Proceedings of the Federated Conference on Computer Science and Information Systems pp. 817–821 DOI: 10.15439/2018F188 ISSN 2300-5963 ACSIS, Vol. 15

Towards a Framework for Semi-Automated Annotation of Human Order Picking Activities Using Motion Capturing

Christopher Reining*, Fernando Moya Rueda†, Michael ten Hompel*, Gernot A. Fink†
*Chair of Materials Handling and Warehousing, †Pattern Recognition in Embedded Systems Group
TU Dortmund, Dortmund, Germany
{christopher.reining, michael.ten.hompel, fernando.moya, gernot.fink}@tu-dortmund.de

***Abstract*—Data creation for Human Activity Recognition (HAR) requires an immense human effort and contextual knowledge for manual annotation. This paper proposes a framework for semi-automated annotation of sequential data in the order picking process using a motion capturing system. Additionally, it introduces proper annotation labels by defining process steps, human activities and simple human movements in order picking scenarios. An attribute representation based on simple human movements meets the challenges set by the versatility of activities in warehousing.**

I. Introduction

ORDER picking is the process of pulling items from a warehouse to satisfy specific customer orders. This basic warehousing process makes up more than half of the total operating expenses [1, p.1-30]. Sub-processes may be partially automatized in high-wage countries. Nevertheless, manual order picking systems remain dominant in practice [2]. To evaluate order picking systems, manual processes need to be quantitatively determinable [3]. Manual assessment of the order picking efficiency is unfeasible as trained specialists would be required to manually gather the necessary information in a highly versatile environment. Due to advancements in sensor technology and data processing, IT-supported approaches of Human Activity Recognition (HAR) gain significance.

HAR is a classification task where time-series segments are assigned to a specific activity class [4], [5], [6]. The authors in [5] provided the first approach of HAR in the order picking process. They recorded multichannel time-series from Inertial Measurement Units (IMUs). IMUs were attached to both arms and the torso of three workers in two warehouses. IMUs provide measurements of three different sensors: accelerometers, gyroscopes and magnetometers for three axes (x, y, z). The authors followed a standard pipeline in pattern recognition; that is, segmenting sequences, extracting handcrafted features, and training a classifier. They used a sliding window approach for segmenting time-series segments. For each of these segments, statistical features were computed and processed by three classifiers. Recently, deep convolutional neural networks (CNN) and recurrent neural networks (RNNs) were successfully used for recognizing human activities [6], [7], [8], [9]. A combination of convolutional layers and recurrent units is proposed in [7] for recognizing activities of daily life. In [8], different deep architectures were deployed to recognize human locomotion activities. In particular, they used a CNN, a long-short term memory (LSTM) network, and a bi-directional LSTM network. The authors in [6] proposed a CNN for solving HAR in the order picking process. In contrast to previous architectures, this CNN contains parallel branches. Each of these branches is composed of two or three convolutional layers and max-pooling operations processing segments per IMU. This architecture, called IMU-CNN, showed the state-of-the-art performance in HAR.

The success of deep architectures in different tasks heavily depends on the amount of data. Nowadays, large collections of data are available for tasks such as image classification, image segmentation and face recognition. However, this is not the case for HAR, which datasets are rather small and scarce. Providing data collections involves recording high quality raw data along with their respective class annotations. Data should be large, variate and correctly labeled. This process in HAR is more challenging in comparison with other tasks. For image classification datasets, label annotations can be carried out using a combination of unsupervised clustering and manual work [10]. However, HAR is diverse involving different type of data sources, e.g. from videos, or multichannel time-series from on-body sensors. HAR faces challenges with regards to environment settings, number of participants and number of sensors [4]. Furthermore, due to the large intra- and inter-class variability of the human movements, a large number of experiments must be carried out, which draw motion repetitions from the same or different persons [7]. These circumstances increase the data collection and annotation efforts. Obtaining and annotating data from videos is computational expensive, and, in the case of multichannel time-series signals, signals are visually hard to interpret. In both cases, annotations are carried out manually, involving the synchronization of the time-series with videos, observing the actions and labeling the sequences. This procedure takes enormous time. For example, annotations of time-series in the order picking dataset in [11] demanded $26min$ in average per minute of annotated data. In addition, annotations are inconsistent among different

The work on this publication has been supported by Deutsche Forschungsgemeinschaft (DFG) in the context of the research project "Adaptive, ortsabhängige Aktivitätserkennung und Bewegungsklassifikation zur Analyse des manuellen Kommissionierprozesses"

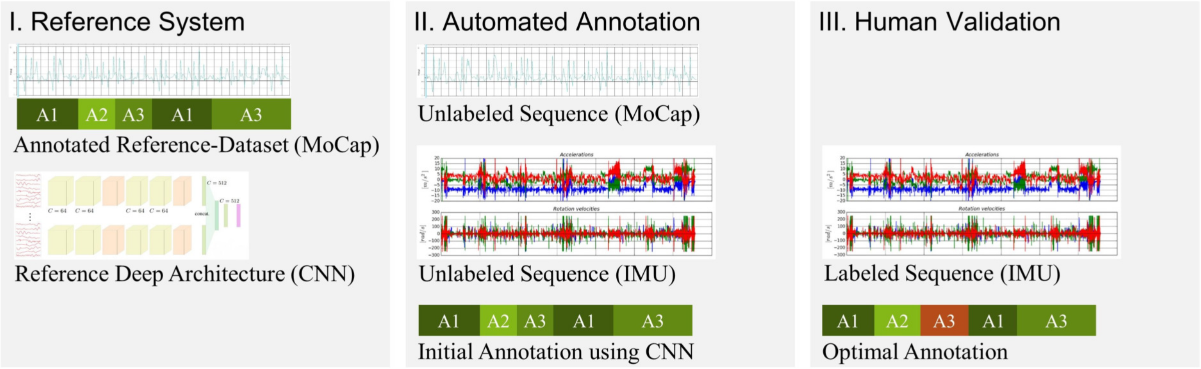

Fig. 1: Framework for semi-automated Annotation

annotators. Repetitions in the annotation process enhance the data quality [11], but escalate the data collection effort.

Apart from the annotation effort, the definition of activities is of high interest. In order picking scenarios, coarse activities like walking, picking and searching are often used [6], [11]. However, these scenarios are highly versatile involving a variety of activities. A possible way out is the definition of more finely subdivided activities. This implies more effort for creation and annotation of datasets. Following [9], activities could be represented by a set of attributes. Attributes are high level semantic descriptions of activities [12], [13]. These attributes are shared among all of the activities. For example, attributes like moving or not moving a foot and the velocity could define walking, running, and standing. Using an attribute based representation, problems like imbalanced data and overfitting are reduced. Sequential data from the most frequent activities could be used for learning attributes that are shared with less frequent activities, as simple human movements are shared among activities. In general, attribute annotations in the context of multichannel time-series HAR are not available. The annotations are related with specific coarse activities, for example standing or walking. However, there are no annotations of attributes describing those coarse activities. In [9], attribute representations for HAR are learned using an evolutionary algorithm, starting from a random combination of attributes. The learned attribute representations are suitable for solving HAR as classification task. However, their semantic interpretation is missing and therefore not understandable by humans.

II. Method

Datasets consisting of multichannel time-series from on-body sensors are of special interest in order picking. Usually, multiple sensors, e.g. IMUs, are worn by a worker gathering recordings in a simple and non-invasive manner. Besides, these sensors are impersonal, i.e. recordings do not portray the identity of the person. In comparison with HAR using videos, they do not suffer from occlusion, as the person's visibility changes along videos. In addition, IMUs are rather economic. Nevertheless, datasets from these devices are hard to annotate manually. As they are difficult to interpret by a human, additional video material is necessary to visualize the respective activity. This paper presents a framework, see Figure 1, to annotate multichannel time-series from on-body sensors using a deep learning model that is trained on highly accurate data. This framework is divided in three parts. First, sequential high quality data are created and annotated from a controlled environment as a reference dataset. Humans are recorded following activities that are commonly seen in order picking scenarios. Proper annotation labels are defined and, in addition, an attribute representation for human activities is introduced. This attribute representation is based on basic human activities and warehousing components. A deep model for solving HAR is learned on the reference dataset. Second, using this model, sequential data from an uncontrolled environment are initially labeled. This initial label includes the computation of uncertainty for the initial predictions. Third, uncertain predictions are revised by human work for final labeling.

A. *Controlled Environment*

On the one hand, naturalistic, real-life data are desired. On the other hand, data is prune to be disturbed in uncontrolled environments [14]. The primary reason to use a controlled environment set-up is the high accuracy of the available sensors. Interfering signals can be averted, and recording sessions can be conducted and repeated with different settings. The Motion Capturing (MoCap) that has been used for this paper is based on photogrammetry methods for measuring object positions on $2D$ and $3D$ spaces using a string of cameras. As an installation of the motion capturing system in a real warehouse is not practicable, it is located at the "InnovationLab Hybrid Services in Logistics" of the chair of materials handling and warehousing at the TU Dortmund University. The MoCap system consists of 38 cameras that cover a space of approximately $22m \times 10m \times 6m$. It uses passive markers to track rigid and flexible objects, such as drones, robots or humans in real time [15]. The passive markers reflect incoming infrared signals to the cameras, and their $3D$ positions are determined via triangulation.

The purpose of the MoCap system is to construct and record skeleton data from workers performing activities in an

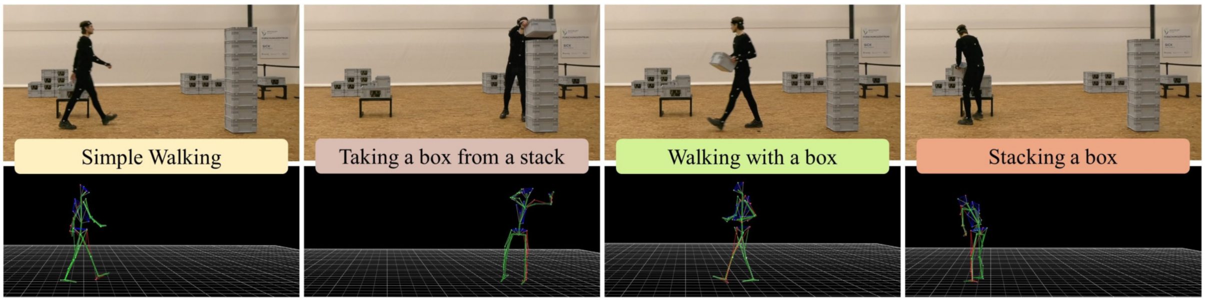

Fig. 2: Exemplary Activities and Motion Capture Data Skeleton

order picking scenario, as shown in Figure 2. Workers wear a specific suit with a set of passive markers. The MoCap computes the global $3D$ positions and it constructs a human skeleton. The MoCap System provides global poses from different parts of the human body, e.g. head, torso, arms and feet. A pose is a combination of position and angular values in $[X, Y, Z]$ of a certain reference system.

B. Annotation of Order Picking Activities

From a macro-level perspective, the human activity in order picking systems can be segregated into basic activities such as locomotion, retrieving and confirming [1, p.1-30]. An obvious approach of HAR would be to interpret each activity as a class. However, this approach is incapable to deal with the versatility of actions in real-world systems. Members of the same class differ significantly in terms of motions and tasks that are executed by the pickers [3]. For example, a warehouse employee can simply walk or walk while carrying a box. A single class cannot account for such distinctions. There is a wide variety of components that influence the human activity, ranging from the type of storage and collecting unit to the information technology [1, p.1-30], [16], [17]. These components and their combinations define coarse order picking process steps, e.g. *putting a box from a shelf onto a cart*. However, process steps can be composed of fine human activities such as *taking a box from a shelf* and *putting a box onto a cart*. Thus, each relevant process step needs to be defined with regards to human activities. This approach offers a high degree of flexibility. On the one hand, the definition of each human activity is fixed so that patterns in the sensor can be recognized and the obtained data is reusable. This is feasible as the definition of human activity is supposed to hold global validity irrespective of a specific context and environment. On the other hand, the definition of process steps is not fixed. Depending on the user's requirements, process steps can be defined very specifically or in more general terms. In addition, following [9], human activities are represented by a set of attributes that describe them semantically. These attributes are simple human movements, for example moving an arm or a foot. As shown in [9], attribute representations boost HAR tasks using deep architectures.

The proposal is to annotate time-series with a respective activity and a set of attributes, see Figure 3. The definition of both the activities and the attributes must be created a priori by a warehousing specialist to ensure that they are semantically understandable. The attributes are the output of the CNN that operates on the sensor data. The combination of attributes implies a specific activity. The activity sequence is then comprehended as a process step of order picking.

C. Creation of Reference Dataset

A reference dataset for order picking scenarios using the MoCap system, see subsection II-A, is created. The closeness to reality within the controlled laboratory environment was ensured by using the same kind of equipment, such as boxes or racks, that are used in real warehouses.

For this reference dataset, eight activities have been recorded: *Standing (none)*, *Walking (none)*, *Standing (box)*, *Walking (box)*, *Reaching forward (none)*, *Lifting (box)*, *Putting down (box)*, *Straighten up (none)*. Here, the words *box* and *none* express whether a worker walks with or without a box. Thus, the sequence of *reaching forward (none)* and *lifting (box)* implies the process step *picking up a box*. The box was a standard small load carrier with the dimensions L 600 mm x W 400 mm x H 220 mm and a gross weight of 4 kg.

The sample recording for this paper was conducted with eight participants of which four have been female and four male. Their height ranged from 161 to 192 cm and the average age was 25. Five participants have been right-handed and three participants have been left-handed. Previous research suggests that the handedness and gender have an impact on the motion [18]. The amount eight participants is equivalent to state-of-the-art approaches [19].

The activities were not recorded in a sequence and subsequently segregated into activities. Rather, they were recorded successively as modular units to ensure the creation of a balanced dataset; that means, all activities have a similar number of recordings regardless of their occurrence in a given scenario. All standing and walking activities have been recorded for five minutes per participant in 5 individual recordings of 60sec each. Both the activities *Reaching forward (none)* and *lifting (box)* were recorded in a single run to reduce the recording effort. The box was picked 10 times from 9 different heights, from the ground level up to a stock of 8 boxes. The participants approached the stack from different starting positions to ensure a natural motion. The boxes had to be lifted with both hands.

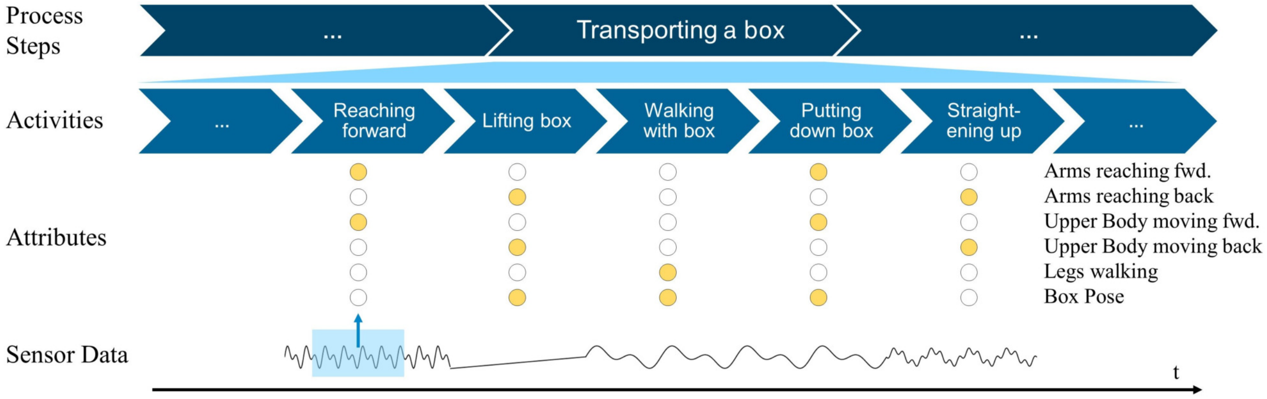

Fig. 3: Attribute based representation of a process step composed of activities that are semantically described

Apart from that, no instructions were given. The total amount of 90 recordings was likewise recorded for the activities *put down (box)* and *straighten up (none)*. A testing data set of 60 sec were recorded for each participant. In the testing data set, the participant conducts a sequence of the previously classified and annotated activities in an arbitrary order and duration. This data set is manually annotated. 202 recordings were conducted with eight participants each, resulting in a total of 1616 recordings. As the data set is based on skeleton poses, one can visualize them easily for annotation purposes. The annotation of walking and standing data sets is simple, as there is no alteration of neither the attributes nor the activity over time. The activities that included the stacked boxes contained not only the two activities *Reaching forward (none)* and *lifting (box)*, as well as *put down (box)* and *straighten up (none)*. The stack was approached and departed by the participants by feet. Therefore, the two *walking* activities and the two *standing* activities were annotated as well. Having this modular recording from activities, the annotation took approximately $2.5min$ per recorded minute.

D. Convolutional Neural Networks for HAR based on Skeleton Datasets

This paper uses the deep architectures, proposed in [6], [9]. These architectures are suitable for multichannel time-series. They are composed of temporal-convolutions and pooling layers, which perform convolution and downsampling operations along the time axis. These architectures extract hierarchical-temporal relations of human movements creating abstract representations of an input sequence. Fully-connected layers connect these representations creating a global one of the input data. The network will compute an attribute representation of an input sequence. This representation is a vector $a \in \mathbb{B}$ containing $1s$ and $0s$ in which 1 for having or not an attribute, the sigmoid activation function is applied to each element of the output layer. Its output corresponds to pseudo-probabilities for each attribute a_i being present in the representation.

The architecture was designed for handling sequences from multichannel time-series, which are measured from m individual portable-devices. These devices are located on different parts of the human body. Convolutional and pooling layers are configured in parallel branches for processing these sequences. Specifically, a single branch processes sequences from a single device increasing the descriptiveness. This architecture is called CNN-IMU. Besides, this configuration allows for more robustness against different and asynchronous devices. This architecture contains m convolutional branches, one per device. Each branch is composed of four temporal-convolution, two max-pooling layers and a fully-connected layer.

Different from [6], [9], the input sequences are not measurements from any portable sensor located on human body parts. Sequences are provided from the MoCap System, see subsection II-A, which provide global poses of human segments. Then, for each of these segments, one has six different measurements. There are in total 22 human segments, e.g. the head, torso, feet, knees and arms. In total, 134 channels have been taken into account. The global pose sequences are normalized with respect to the lower back human-segment. This is necessary to avoid a dependency of the human activity recognition to a global position of warehousing equipment in the laboratory. Each of this measurements is taken as a channel, similar to sequences from portable devices. One considers in total 132 channels and $m = 22$ branches. In the CNN-IMU, convolutions are computed along the time axis, and their filters are shared among the channels.

For training, the following configurations are employed. Sequences from persons $1 - 6$, person 7 and person 8 are used as training, validation and testing sets respectively. The parameters of the networks are updated by minimizing the binary-cross entropy loss using the stochastic gradient descent with the RMSProp update rule as in [7], [9]. Sequence segments, extracted using a sliding window approach, are fed to the networks. These segments are assigned the most frequent ground truth. In general, learning rates are decreased by $\gamma = 0.1$ at a certain epoch or iteration during training. Additionally, we use dropout with probability of 50% on the inputs of the first and second fully-connected layer, and orthogonal initialization [7]. As suggested in [6], [7], input sequences were normalized per channel to a range $[0, 1]$. Moreover, a Gaussian noise of $\mu = 0$ and $\sigma = 0.01$ is added, simulating inaccuracies on the MoCap System. For a given attribute representation A describing the aforementioned activities in the reference dataset, a nearest neighbour approach is used for predicting a specific activity by measuring the cosine distance from the CNN's output for a certain input sequence $\tilde{a}$ to the set $a \in A$. Different sets A of attribute representations, provided by experts, will be evaluated.

E. Human Validation

Following a sliding window approach with a window size of T and step of s, an unlabeled sequence from the reference dataset and an unlabeled sequence from IMU's measurements are segmented. A set of D sequences of size T are then obtained. These sequences are fed to the CNN-IMU computing their attribute representations. By means of a nearest neighbor, these sequences are assigned to the activity where the distance between their representations is minimal. Following [20], an uncertainty measure can be computed for each of the predictions. This measure give a value of how certain a CNN is with respect to a prediction. Uncertain predictions are then revised by experts for generating the final annotation of the sequence.

III. Discussion and Conclusion

This contribution proposed a framework to reduce the annotation effort for multichannel time-series. An attribute based representation creates a high-level semantic description of activities. This is beneficial to make full use of imbalanced data, avoid overfitting and to recognize unseen activities. The logical connection of activities and process steps has been explained and an exemplary attribute representation has been provided. Motion Capture datasets of eight activities including a training and validation data set have been recorded with eight participants each, resulting in a total 1616 recordings. The recordings have been annotated, normalized and used to train a state-of-the-art CNN. Recording further participants and the manual annotation of the MoCap data requires few manual effort. The attributes used by the CNN can be understood by a human and thus transferred to new activities.

Based on the proposed framework, a large multichannel time series can be annotated with respect to semantics in a semi-automated manner. It is not restricted to IMU data but can be used for other sources, such as video data, as well.

References

[1] R. Manzini, Ed., *Warehousing in the global supply chain: advanced models, tools and applications for storage systems*. Springer, 2012.

[2] E. H. Grosse, C. H. Glock, and W. P. Neumann, “Human factors in order picking: a content analysis of the literature,” *International Journal of Production Research*, vol. 55, no. 5, pp. 1260–1276, Mar. 2017.

[3] K. Weisner and J. Deuse, “Assessment Methodology to Design an Ergonomic and Sustainable Order Picking System Using Motion Capturing Systems,” *Procedia CIRP*, vol. 17, pp. 422–427, Jan. 2014.

[4] A. Bulling, U. Blanke, and B. Schiele, “A tutorial on human activity recognition using body-worn inertial sensors,” *ACM Computing Surveys*, vol. 46, no. 3, pp. 1–33, Jan. 2014.

[5] S. Feldhorst, M. Masoudenijad, M. ten Hompel, and G. A. Fink, “Motion Classification for Analyzing the Order Picking Process Using Mobile Sensors,” in *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*, ser. ICPRAM 2016. Portugal: SCITEPRESS - Science and Technology Publications, Lda, 2016, pp. 706–713.

[6] R. Grzeszick, J. M. Lenk, F. M. Rueda, G. A. Fink, S. Feldhorst, and M. ten Hompel, “Deep Neural Network based Human Activity Recognition for the Order Picking Process,” in *Proceedings of the 4th international Workshop on Sensor-based Activity Recognition and Interaction*. ACM Press, 2017, pp. 1–6.

[7] D. R. Francisco Javier Ordóñez, “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition,” *Sensors*, vol. 16, no. Advances on Data Transmission and Analysis for Wearable Sensors Systems, p. 115, 2016.

[8] N. Y. Hammerla, S. Halloran, and T. Ploetz, “Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables,” *CoRR*, Apr. 2016.

[9] F. M. Rueda and G. A. Fink, “Learning Attribute Representation for Human Activity Recognition,” *arXiv:1802.00761 [cs]*, Feb. 2018. [Online]. Available: http://arxiv.org/abs/1802.00761

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep Face Recognition,” in *British Machine Vision Conference*, 2015.

[11] S. Feldhorst, S. Aniol, and M. ten Hompel, “Human Activity Recognition in der Kommissionierung – Charakterisierung des Kommissionierprozesses als Ausgangsbasis für die Methodenentwicklung,” *Logistics Journal : Proceedings*, vol. 2016, no. 10, Oct. 2016.

[12] H.-T. Cheng, F.-T. Sun, M. Griss, P. Davis, J. Li, and D. You, “NuActiv: Recognizing Unseen New Activities Using Semantic Attribute-based Learning,” in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '13. New York, NY, USA: ACM, 2013, pp. 361–374.

[13] J. Zheng, Z. Jiang, and R. Chellappa, “Submodular Attribute Selection for Visual Recognition,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 39, Issue: 11,)*, ser. 11, vol. 39. IEEE, Nov. 2017, pp. 2242 – 2255.

[14] A. Vinciarelli, A. Esposito, E. André, F. Bonin, M. Chetouani, J. F. Cohn, M. Cristani, F. Fuhrmann, E. Gilmartin, Z. Hammal, D. Heylen, R. Kaiser, M. Koutsombogera, A. Potamianos, S. Renals, G. Riccardi, and A. A. Salah, “Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions,” *Cognitive Computation*, vol. 7, no. 4, pp. 397–413, Aug. 2015.

[15] A. K. R. Venkatapathy, H. Bayhan, F. Zeidler, and M. t. Hompel, “Human machine synergies in intra-logistics: Creating a hybrid network for research and technologies,” in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2017, pp. 1065–1068.

[16] J. Haase and D. Beimborn, “Acceptance of Warehouse Picking Systems: A Literature Review,” in *Proceedings of the 2017 ACM SIGMIS Conference on Computers and People Research*, ser. SIGMIS-CPR '17. New York, NY, USA: ACM, 2017, pp. 53–60.

[17] D. Battini, M. Calzavara, A. Persona, and F. Sgarbossa, “Additional effort estimation due to ergonomic conditions in order picking systems,” *International Journal of Production Research*, vol. 55, no. 10, pp. 2764–2774, May 2017.

[18] R. Müller-Rath, C. Disselhorst-Klug, S. Williams, C. Braun, and O. Miltner, “Einfluss des Geschlechts und der Seitendominanz auf die Ergebnisse der quantitativen, dreidimensionalen Bewegungsanalyse der oberen Extremitäten,” *Zeitschrift für Orthopädie und Unfallchirurgie*, vol. 147, no. 04, pp. 463–471, Jul. 2009.

[19] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Millàn, “Collecting complex activity datasets in highly rich networked sensor environments,” in *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, Jun. 2010, pp. 233–240.

[20] R. Grzeszick, S. Sudholt, and G. A. Fink, “Optimistic and Pessimistic Neural Networks for Scene and Object Recognition,” *arXiv:1609.07982 [cs]*, Sep. 2016.