DOI: 10.15439/2016F472
ACSIS, Vol. 8. ISSN 2300-5963

A compact deep convolutional neural network architecture for video based age and gender estimation

Bartłomiej Hebda
AGH University of Science and Technology
Krakow, Poland
E-mail: hebda.bartlomiej@gmail.com

Tomasz Kryjak, *Member IEEE*
AGH University of Science and Technology
Krakow, Poland
E-mail: tomasz.kryjak@agh.edu.pl

***Abstract*—In this paper research on a compact deep convolutional neural network (DCNN) architecture for age and gender estimation from facial images has been presented. The proposed solution was tested on the FERET and the Adience Benchmark databases. In the first case a 98.6% accuracy for gender and 86.4% for age estimation was obtained. For the Adience database, which contains images recorded in unconstrained conditions and is much more demanding, a 62.0% for gender and 42.0% for age accuracy was obtained. When compared to the reference results on a much larger network, the performance should be considered as satisfactory. The research shows that a compact DCNN with small input images can provide quite good classification results.**

I. Introduction

A VISION system which allows the estimation of age and gender of a person using a face image can have a number of important practical applications in biometric and statistics systems, as well as advanced human-computer interfaces (HCI) and so-called smart advertising (with personalized content). It typically consists of a face detection [1] and the actual estimation module. The designed vision system must be resistant to changes in face appearance (facial expressions, hairstyle, presence of beard or moustache, glasses and to some extent also make-up), different lightening conditions, inaccurate face localization in the image, face orientation (frontal, from profile and also rotated), size of the input image and it's quality (presence of noise, blur caused by movement of the person, underexposure, overexposure, shadows, etc.).

Deep convolutional neural networks (DCNN) are one of the most interesting tool available for the image processing and machine learning community in recent years. It is worth noting that the approach in not new – the first concepts were already proposed in the 70's of the last century [2]. However, due to limited performance of the available computing platforms, these solution were not used. The breakthrough came with the emergence of programmable graphic processing units (GPU). They proved to be an almost perfect platform for neural networks implementation, mainly because of the massive parallelization possibility and floating point support. At the same time learning methods for such complex structures were developed and refined. In addition, the dynamic growth of social networks services like Facebook, Flickr or Instagram allowed to obtain quite easy access to huge image databases. Currently DCNNs are used for almost all machine learning tasks – from speech recognition through image recognition to information about social networks users categorization. Only in computer vision the following applications should be mentioned: face detection, pedestrian detection, road sign detection and recognition and object tracking. What is more, the DCNN based approaches usually significantly outperform the "classic" (i.e. feature extraction and classification – e.g. HOG + SVM) ones.

In this paper, a compact DCNN for age and gender estimation from facial images is presented. The obtained results indicate, that it is possible to use a quite small, energy and resource efficient architecture, without much loss on classification performance. Moreover, all experiments were performed on a typical PC computer, without a powerful GPU accelerator.

The reminder of this paper is organized as follows. In Section II a brief review of age and gender estimation algorithms is presented. The proposed solution is described in Section III. Then, in Section IV the evaluation results are presented and discussed. The paper ends with a summary and indication of future research directions.

II. Age and gender estimation systems

Age and gender estimation is a quite popular topic in the computer vision community. An in-depth review is far beyond the scope of this article and therefore only selected works directly related to DCNNs are presented.

A. DCNN based solutions

One of the earlier works on gender recognition using CNNs was presented in paper [3]. The solution consisted of a face detection and gender recognition module – both using neural networks. The architecture involved three layers (two hidden and output). Input images had a 32×32 pixels resolution. The reported accuracy on the FERET dataset was 97.2%.

Most of the work related to the DCNNs appeared in 2015 and later. In the article [4] two approaches were compared: "classic" and DCNN based. In the first case the following

978-83-60810-90-3/$25.00©2016, IEEE

features were considered: HOG (Histogram of Oriented Gradients), LBP (Local Binary Patterns) and SURF (Speeded Up Robust Features). As regression the CCA (Canonical Correlation Analysis) was applied. In the second, many different variants of network architectures were examined (the Caffe library was used). The best result were obtained for two convolutional and one fully connected layer. Input images had 50 × 50 pixels size. The authors noted a significant disproportion between the time required for learning and actual operation in both cases. Finally, for the MORPH database, the "classic" solution obtained 4.25 and DCNN 3.88 mean absolute error (MAE) value.

In the work [5] a DCNN for age and gender estimation was proposed. The network had three convolutinal and two fully connected layers. Input images of size 256×256 were cropped to 227×227. The authors did not use a pre-trained network model – in contrast to many other approaches. On the Adience dataset this solution achieved 86.8% ± 1.4 accuracy for gender and 50.5% ± 5.1 for age estimation. In the latter case 8 age categories were used. If an "off by one" error is allowed, the performance increases to 84.7 ± 2.2. In this study the Caffe library and GPU with 1,536 CUDA cores and 4 GB of RAM were used.

In 2015, at the IEEE International Conference on Computer Vision Workshop (ICCVW) a competition on age estimation (ChaLearn) was conducted. The Looking at People 2015 (LAP) dataset with 4,961 images was used [6]. The DEX system from ETH Zurich, Switzerland obtained the best results [7]. In the first step the face was detected. Then, it was rescaled to 256×256 pixel size. The authors used a pre-trained DCNN on the ImageNet dataset. In addition, it was fine-trained on the IMDB-WIKI database (260,282 images) and finally on the LAP set. In total 20 separate networks were used. This allowed to achieve 3.21 MAE value. The Caffe framework and NVIDIA Tesla K40C GPU were used. Learning a single network lasted 5 days, fine-learning 3 hours and processing a single image about 200 ms.

This short (compared to the available material) overview can be concluded with the following statement: "age and gender estimation with DCNNs is very accurate (sometimes even better than human), however the used network architectures are very complex". Therefore, their training and usage requires a lot of computing resources – most authors used specialized and expensive GPU accelerators like NVIDIA K40.

B. Our previous solution

During our earlier work [8], a "classical" age, gender and race estimation system was created. In the pre-processing stage the face was detected (Viola-Jones method), eyes and mouth were detected (also by the Viola-Jones approach) and image alignment was performed – an affine transform was used to assure that eyes and mouth are at predefined locations for every considered image. The local binary patterns (LBP) were used as features. After their computation, the image was divided into non-overlapping blocks in which histograms were computed – their concatenation formed the feature vector. The support vector machine (SVM) with radial basis functions (RBF) was used as classifier. In case of non-binary problems (race, age) multiple SVMs were used – one for each class.

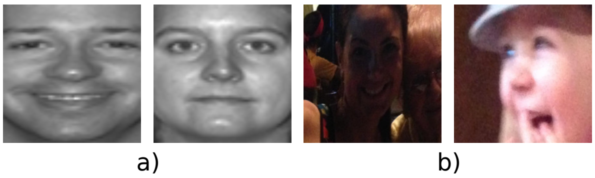

Figure 1: Samples from the FERET (a) and Adience (b) databases

In the experiments on the FERET database quite good performance (94.0% for gender, 86,7% for race and 69.5 % for age in 10 years interval) was obtained. Unfortunately, real-life tests in unconstrained conditions on images acquired with a standard USB webcam revealed the weakness of this solution. Actually, even for gender determination it was quit difficult to get a proper result. This was a direct motivation to improve the vision system and use the DCNN approach.

III. The proposed solution

In this work the impact of simplifying the network structure, especially the size of the input image, on the overall accuracy of age and gender recognition was evaluated. Also the computing resources and single image processing time were considered. The ultimate goal of the ongoing research is to implement DCNN solutions in embedded devices and support real-time video stream processing.

A. The used datasets

In the experiments two databases were used: FERET [9] and Adience [10]. The first was chosen to compare the obtained results with our previous work [8]. It includes 2,413 images of 856 individuals. Examples are shown in Figure 1a. The biggest drawback is that these pictures were taken under controlled conditions. This results in poor generalization of the trained classifier, especially for cases registered under real-life conditions. In addition modern methods obtain almost 100% accuracy for this set, which indicates that it is not longer an appropriate challenge.

The Adience set was used because it is considered as challenging and the images were registered in unconstrained environment. It contains 26,580 annotated images of 2,284 people. Samples are shown in Figure 1b.

B. The used hardware and software

All experiments were performed on mid-range laptop – Intel Core i5-2410M (2.3 GHz up to 2.9 GHz), 4 GB DDR3 RAM, NVIDIA GeForce GT 540M with 2 GB DDR3 RAM, 96 CUDA cores and 672 MHz core frequency. There are several computing libraries for DCNNs – eg. Caffe, Torch and Google Tensor Flow. After a preliminary analysis, Torch (`http://torch.ch/`) was selected, mainly due to easier installation and configuration (in comparison to Caffe), as

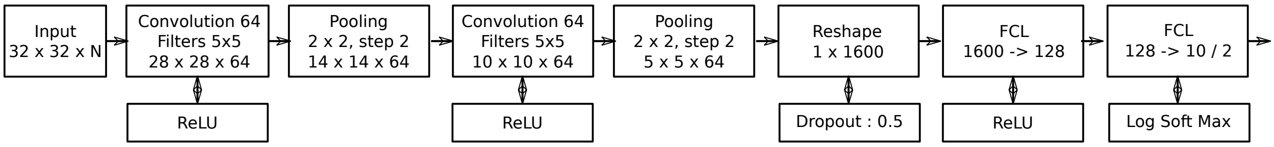

Figure 2: The used DCNN architecture

Table I: Results for the FERET (F) and Adience (A) dataset

	Conv. 1	Conv. 2	# Training	# Test	# Epoch	Acc. gender	Acc. age	Computing time
F1	5×5	5×5	100	100	100	81.3%	41.2%	48s
F2	5×5	5×5	250	100	200	96.0%	70.1%	181s
F3	5×5	5×5	500	200	300	88.5%	74.5%	534s
F4	5×5	5×5	1000	500	500	93.7%	58.9%	1795s
F5	5×5	5×5	2000	500	500	**98.6%**	85.4%	3324s
F6	3×3	5×5	2000	500	200	97.6%	**86.4%**	1572s
F7	7×7	5×5	2000	500	200	97.6%	85.6%	1683s
F8	3×3	3×3	2000	500	200	98.0%	**86.4%**	1354s
F9	7×7	7×7	2000	500	200	97.6%	85.8%	1709s
A1	5×5	5×5	2000	500	300	50.2%	42%	2317 s
A2	5×5	5×5	12000	500	200	51.4%	32.1%	3h
A3	5×5	5×5	12000	500	200	59.5%	33.1%	3h
A4	5×5	5×5	12000	500	200	**62.0%**	**42.0%**	3h

well as a more convenient to manage network model and the whole project (LUA language). Additionally, in our opinion the available documentation is very extensive and there are many resources on the Internet (ready scripts and educational materials). It is also possible to import models from the Caffe library.

C. Image preprocessing

In the first stage, the images from the FERET and Adience databases were rescaled to 32×32 pixels. Unfortunately, on the used hardware, experiments for higher resolution lasted too long (several hours) and for larger models were even impossible. Then, the original RGB colour space was transformed to YUV, which allowed to separate luminance and chrominance information and slightly improve the performance.

In the next step the input data was normalized using the well known scheme with mean and standard deviation. First, this two values were computed for each colour component separately. Then, from all pixel in the input image the mean value was subtracted and the result was divided by the standard deviation.

D. Network architecture

As stated earlier, the main assumption of the presented work was to design a simple and computationally efficient network architecture. Therefore, it was decided to use 32×32 pixels input images. It should be emphasized that usually larger images are considered (cf. Section II-A).

The used network architecture was essentially based on the one described in [5]. However, many simplifications were introduced. The scheme is presented in Figure 2. It consists of the following components: two convolutional layers (Convolution 64), three ReLU transfer layers (in-place operation – $f(x) = max(0, x)$), two subsampling layers with a *max* operator (Pooling), data reshape, two fully connected layers (FCL), in-place dropout with 0.5 probability and LogSoftMax operation $(f_i(x) = log(\frac{1}{\sum_j e^{x_j}} \cdot e^{x_i}))$.

E. Network training

In the experiments the following labels were used. For gender: 1 – man, 2 – woman. For age: 0 – 9, as the range $[0, 100]$ was divided into 10 intervals (we used the value 10 due to compatibility with our previous research). Some solutions allow to determine the age with one year accuracy – for example [7]. However, this increases the complexity of the network structure and for a lot of applications this precision is not required.

After creation of the network model, the weights were initialized (in this work random values were used). Then, during an iterative learning process (using stochastic gradient descent (SDG) method) the weights were modified to reduce the classification error.

IV. Evaluation

For the FERET database 9 experiments were carried out. Their main aim was to examine how the size of used convolutional filters and sizes of the training and test sets, as well as the number of iterations (epochs) affected the classification performance. The obtained results are summarized in Table I (upper part). As the training times for both networks (age, gender) were similar, only the mean value is presented.

Like expected, increasing the number of training samples had a positive effect on the recognition efficiency – using 2,000 images allowed to obtain a 98.6% gender recognition accuracy. It is worth noting, that this is 4% better than the "classic" solution descried in Section II-B and work [8].

Age estimation is a more complex issue. In addition, it was not easy to provide enough training and test samples for each of the 10 classes. Again, the best results were obtained for

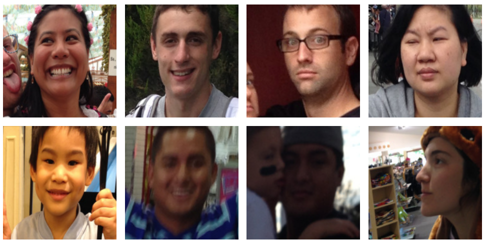

Figure 3: Sample positive (top row) and negative (bottom row) gender classification results from the Adience dataset

the largest network – 86.4%. It is over 15% better than the "classical" solution and slightly better than reported in [3].

Using different filter sizes (3×3, 5×5 and 7×7) had only a minor impact on the final classification performance (not more than 1% difference). A small 3×3 filter allowed to achieve best age estimation – 86.4%.

For the Adience dataset 4 test were conducted. Their results are summarized in Table I (bottom part). Images of size 32×32 were used. In the first test 50% for gender and 42% for age accuracy was obtained. This is much worse than on the FERET base, but the reference DCNN from [5] on this demanding set allowed for approx. 87% for gender and 50.7% for age accuracy. Because the size of the input image could not be increased, two optimizations proposed in the work [5] were evaluated on the compact DCNN. In the third test, the input image was restricted only to the centre part of the original image. In the fourth test, 5 images were passed to network – one centre and 4 from corners. This allowed to compensate the incorrect alignment of faces in the input images. Especially the second modification had a positive impact on the final solution – 62% for gender and 42% for age accuracy. Some sample gender classification results are presented in Figure 3. It should be noted that the unsuccessful cases (bottom row) are quite difficult.

V. Conclusion

In this paper the use of a compact DCNN architecture for age and gender estimation was proposed. The input image size was defined as 32×32 pixels. This allowed to obtain 98.6% accuracy for gender and 85.34% for age in 10 years intervals on the FERET database. When compared to our previous solution based on LBP and SVM, respectively a 4% and 15% improvement was obtained.

For a much more demanding Adience database 62% gender and 42% age accuracy was measured. These are results respectively 25% and 8% worse than for the large DCNN with input image size 227×227. It is worth to emphasize that the measured classification time was 16 ms (vs 200 ms) and this despite of the used computing platform (mid-range notebook).

The obtained results are interesting due to at least three reasons. Firstly, the access to powerful GPU platforms or even clusters in not common. Especially nowadays, where most personal computers are notebooks which are being slowly replaced by advanced smartphones. Secondly, the "didactic" aspect should not be missed. It seems that deep learning and convolutional neural networks should be introduced for graduate students of electrical engineering and computer science faculties. For an effective teaching process, students should be able to carry out simple experiments on their own hardware. In addition, the network training should not take too long (e.g. not several hours) and the classification result should be "decent" (reasonable when compared to large DCNNs). Thirdly, the use of a compact DCNN in an embedded vision system allows to reduce costs (less computing resources required), minimize the energy consumption and perform real-time video stream processing.

In the near future our research will concentrate on analysing the use of reprogrammable FPGA devices, heterogeneous Zynq SoC (which both are proven platforms for embedded real-time vision systems), as well as the recently proposed solutions like NVIDIA Jetson FX-1 for implementing compact, energy efficient but also accurate DCNN based classifiers and vision systems.

Acknowledgment

The work was supported by AGH University of Science and Technology project number 15.11.120.879.

References

[1] M. Szkudlarek and M. Pietruszka, "Fast gpu and cpu computing for head position estimation," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015. doi: 10.15439/2015F410 pp. 231–240.

[2] A. G. Ivakhnenko, "Polynomial theory of complex systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-1, no. 4, pp. 364–378, Oct 1971. doi: 10.1109/TSMC.1971.4308320

[3] F. H. C. Tivive and A. Bouzerdoum, "A gender recognition system using shunting inhibitory convolutional neural networks," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 2006. doi: 10.1109/IJCNN.2006.247311. ISSN 2161-4393 pp. 5336–5341.

[4] I. Huerta, C. Fernández, C. Segura, J. Hernando, and A. Prati, "A deep analysis on age estimation," *Pattern Recognition Letters*, vol. 68, Part 2, pp. 239 – 249, 2015. doi: http://dx.doi.org/10.1016/j.patrec.2015.06.006 Special Issue on "Soft Biometrics".

[5] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2015. doi: 10.1109/CVPRW.2015.7301352. ISSN 2160-7508 pp. 34–42.

[6] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015. doi: 10.1109/ICCVW.2015.40 pp. 243–251.

[7] R. Rothe, R. Timofte, and L. V. Gool, "Dex: Deep expectation of apparent age from a single image," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015. doi: 10.1109/ICCVW.2015.41 pp. 252–257.

[8] B. Hebda and T. Kryjak, "Age, race and gender estimation based on facial images," in *Zeszyty Studenckiego Towarzystwa Naukowego*, 2015, pp. 137—141.

[9] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The {FERET} database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295 – 306, 1998. doi: http://dx.doi.org/10.1016/S0262-8856(97)00070-X

[10] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, Dec 2014. doi: 10.1109/TIFS.2014.2359646