

# Multi-branch CNN for Multi-scale Age Estimation

Marco Del Coco<sup>(✉)</sup>, Pierluigi Carcagnì, Marco Leo, Paolo Spagnolo,  
Pier Luigi Mazzeo, and Cosimo Distanto

National Research Council of Italy, Institute of Applied Science and Intelligent  
System, Ecotekne Campus, via Montebello, 73100 Lecce, Italy  
[marco.delcoco@isasi.cnr.it](mailto:marco.delcoco@isasi.cnr.it)

**Abstract.** Convolutional Neural Networks (CNNs) attracted growing interest in recent years thanks to their high generalization capabilities that are highly recommended especially for applications working *in the wild* context. However CNNs rely on a huge number of parameters that must be set during training sessions based on very large datasets in order to avoid over-fitting issues. As a consequence the lack in training data is one of the greatest limits for the applicability of deep networks. Another problem is represented by the fixed scale of the filter in the first convolutional layer that limits the analysis performed through the subsequent layers of the network.

This paper proposes a way to overcome these problems by the use of a multi-branch convolutional neural network with a reduced deep. In particular its effectiveness for age group classification has been proved demonstrating how, on the one hand, the reduced deep avoids the over-fitting issues, whereas, on the other hand, the multi-branch structure introduces a parallel multi-scale analysis capable to catch multiple size patterns.

**Keywords:** CNN · Deep learning · Age estimation

## 1 Introduction

Human beings are intrinsically used to estimate soft biometric of other people and, consequentially, they modulate their behavior to the age and gender of the ones involved in the interaction. It turns out that a more natural human-machine interaction could be achieved by introducing these biometric biases also in the human-machine interaction processes. This could improve acceptability and usability of assistive technologies [17] that are part of our everyday life [2], and in particular of social robots already largely used in clinical contexts [4]. In the light of the above the reliability of the soft biometric estimation algorithms becomes a key point in the human-machine scenario. Unfortunately this is still an open challenge since traditional approaches focused on the selection of specific biometric descriptors particularly suitable for a specific application

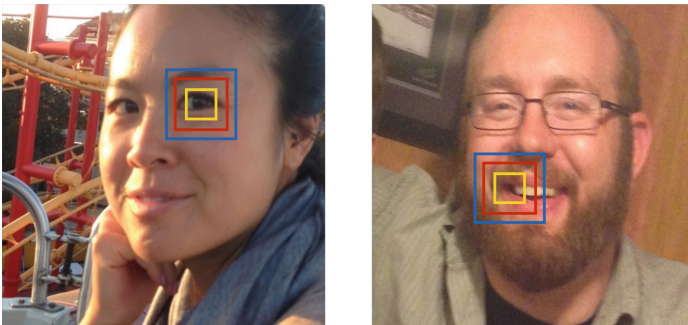
context but unsuitable for the use in unconstrained environments [3]. Where pose, illumination and occlusion changes have to be effectively addressed.

The recent explosion of the Convolutional Neural Network (CNN), jointly with the exponentially hardware advancements, has allowed to overcome the above limitations bringing to generalization capabilities. This is mainly due to the huge number of parameters that characterizes the CNN structure, as well as to the complete adaptability of the convolutional filtering side of the network that makes possible to discover a larger number of hidden structures. It is straightforward that the big amount of parameters requires a large amount of data for carrying out a learning phase free of over-fitting issues but, unfortunately, well annotated data are, nowadays, a limited resource due to the long and tedious phase for collecting and labeling them.

Another limit is that a CNN can adapt the inner parameters of each filter, but the filter size is chosen by the operator making the initial scale of analysis constrained to a fixed value.

A possible solution to avoid the overfitting issues, when available data are undersized, is to contain the network depth (i.e. the number of inner parameters), as suggested in [18]. This approach makes sense especially when the problem is quite simple (few output classes) with respect to classifications problems requiring many output classes (people or object recognition).

In this work the small sized network proposed in [18] is exploited for age group classification on the Adience face dataset [7]. In addition an innovative multi-branch structure (with different staring filter sizes) is proposed in order to achieve a better multi-scale analysis with respect to the one naturally accounted by classical CNN. In other words, the main contribution of this paper is that the proposed approach takes advantage of the parallel multi-scale analysis that is directly performed on the raw signal (see Fig. 1). In the proposed architecture, that of course has a larger number of parameters than the single branch one, overfitting issues are limited since the backpropagated errors are in parallel distributed among the different branches.



**Fig. 1.** Multiscale CNN processing: the multi-scale convolutional filters allow an analysis of different pattern sizes.

The rest of the paper is organized as follows: Sect. 2 aims to give a short overview of the leading approaches for age group classification presented in literature across last years; Sect. 3 presents the whole system giving a detailed description of the network structure; Sect. 4 deals with the presentation and discussion of experimental results; Sect. 5 is finally devoted to the conclusions and future works discussion.

## 2 Related Works

The problem of automatic age estimation from facial images attracted many researchers in the recent years. The solution has been treated in many ways, from the age group classification to the precise age estimation employing regression techniques. A detailed survey concerning both strategies is presented in [11].

Early researches focused the attention on the hand-crafted rule. First of all, facial features (i.e. nose mouth. . . ) were detected in the face image and their size and mutual distances are computed; successively these measures were employed for the age group classification [15]. A similar approach has been employed in order to model age progression in teenager people [19]. It is obvious that the narrow age range requires a precise measurement of facial parts both in terms of size and distance, making the method unsuitable for *in the wild* application. Other approaches suffering the near frontal constraint are the ones based on the representation of age process as a subspace [9] or a manifold [10].

An alternative way is the use of local features for face aging representation. As an instance, Hidden-Markov-Model were used in [23] for representing face patch distributions. Alternatively, Gaussian Mixture Models (GMM) are employed in [22] for the same intent whereas [21] they were used again for representing the distribution of local facial measurements, but robust descriptors were used instead of pixel patches.

An improved version of relevant component analysis [1] (exploited for distance learning) and locally preserving projections [12] (exploited for dimensionality reduction) are combined with Active Appearance Models (as image feature) in the work proposed in [5].

Another research line has been represented by the robust image descriptors. In [8], Gabor image descriptors were used jointly to a Fuzzy-LDA classifier which describes a face image as belonging to more than one age class.

Two of the main exploited approaches are Gabor filters and local binary patterns (LBP) that have been successfully employed in [6] along with a hierarchical age classifier. An interesting comparison of different descriptors, spacial reduction approaches and classification is shown in [3] where a detailed discussion of pose and scale influence is given.

The main limit of the listed methods is their inapplicability to constrained datasets, making their use in real scenarios, unpredictable.

The necessity of highly generalization capability led the research community toward a growing employment of CNN solutions, facilitated by the availability of high performance hardware (GPU equipped with thousand of cores and big amount of memory) and big dataset usually collected through the web.

One of the first solution has been represented by [16] where the LeNet-5 network has been successfully applied to the character recognition system. Another surprising application of CNN has been the classification task on *Imagenet* dataset, a dataset containing hundreds of categories in unconstrained environment [14].

Currently, one of the most interesting challenge is represented by face analysis on face datasets with significant pose variations, occlusions, and poor quality i.e. the Adience dataset. A preliminary solution has been presented in [7] by means of LBP descriptor variations [20] plus SVM. Successively, in [18], the authors suggest the use of a CNN made up by a reduced number of layers respect to the approach proposed in [14]; this is possible considering the limited number of classes in the age group problem (compared to the Image net one occurring in [14]) and allow to avoid the overfitting problem. Due to the notable results obtained this last work as been chosen as the main competitor for the proposed solution.

### 3 Multi-branch CNN

The most exploited CNN architecture is oriented to the modeling of *linear* and *deep* networks characterized by many parameters and great generalization capabilities. Unfortunately this kind of structure presents two main limits: the *overfitting problem* (occurring when the training data are not sufficient for the scope) and the *scale of analysis* that is constrained to the starting convolutional filter size and could be unsuitable to catch key patterns. This last aspect can be avoided by means of a brute force research of the best initial scale. Also in this case any assurance is given concerning the capacity of the network to capture all the important patterns across the image.

The proposed network is based on two main points:

- *A reduced network depth*: as suggested in [18] a reduction of the number of layers in the network helps to reduce the overfitting issue, especially in case such as gender or age group classification characterized by a small number of output classes.
- *A multi-branch architecture*: it is devoted to capture multiple scale appearance pattern keeping, at the same time, a sufficient robustness to the overfitting problem thanks to the independent back-propagation of the error.

The proposed network architecture, illustrated in Fig. 2, is made up by 2 main sides: the multi-branch one, aimed to decompose and filter the input image on multiple scales, and the single-branch one, aimed to collect the extracted features and to feed a series of full connected neural layers. RGB registered facial images of  $256 \times 256$  pixels are centrally cropped to  $227 \times 227$  pixels that are the inputs of the network. The  $i$ -th branch is referred as  $B(i)$ , the  $j$ -th convolutional block (involving different layers) is referred as  $CL(j)$  and, finally, the  $k$ -th fully-connected layer is referred as  $FC(k)$ .

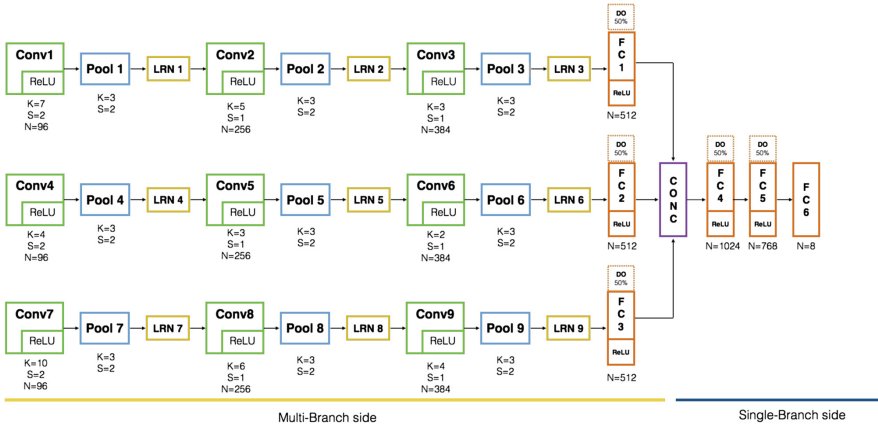


Fig. 2. Proposed network

B(1)-CL(1): it uses 96 filters of size  $4 \times 4$  working directly on the input image. It is followed by a rectified linear operator (ReLU) and successively by a max pooling layer (aimed to take the maximal value of  $3 \times 3$  regions with two-pixel strides) and a final local response normalization (LRN) layer.

B(1)-CL(2): this layer block processes the output of the previous one by means of 256 filter of size  $96 \times 3 \times 3$ . As for the previous one, it is followed by a ReLU, a max pooling layer and a LRN layer working with the same parameters used before.

B(1)-CL (3): This last convolutional layer operates on the received input from the bottom one by applying a set of 384 filters of size  $256 \times 2 \times 2$  pixels followed by only a ReLU and a max pooling layer.

B(1)-FC (1): this first fully connected layer works on the received output of the last convolutional layer and contains 512 neurons. It is followed by a ReLU and a dropout layer.

B(2)-CL(4): it uses 96 filters of size  $7 \times 7$  working directly on the input image. It is followed by a rectified linear operator (ReLU) and successively by a max pooling layer (aimed to take the maximal value of  $3 \times 3$  regions with two-pixel strides) and a final local response normalization (LRN) layer.

B(2)-CL(5): this layer block processes the output of the previous one by means of 256 filter of size  $96 \times 5 \times 5$ . As for the previous one, it is followed by a ReLU, a max pooling layer and a LRN layer working with the same parameters used before.

B(2)-CL (6): This last convolutional layer operates on the received input from the bottom one by applying a set of 384 filters of size  $256 \times 3 \times 3$  pixels followed by only a ReLU and a max pooling layer.

B(2)-FC (2): this first fully connected layer works on the received output of the last convolutional layer and contains 512 neurons. It is followed by a ReLU and a dropout layer.

- B(3)-CL(7):** it uses 96 filters of size  $10 \times 10$  working directly on the input image. It is followed by a rectified linear operator (ReLU) and successively by a max pooling layer (aimed to take the maximal value of  $3 \times 3$  regions with two-pixel strides) and a final local response normalization (LRN) layer.
- B(3)-CL(8):** this layer block processes the output of the previous one by means of 256 filter of size  $96 \times 6 \times 6$ . As for the previous one, it is followed by a ReLU, a max pooling layer and a LRN layer working with the same parameters used before.
- B(3)-CL (9):** This last convolutional layer operates on the received input from the bottom one by applying a set of 384 filters of size  $256 \times 4 \times 4$  pixels followed by only a ReLU and a max pooling layer.
- B(3)-FC (3):** this first fully connected layer works on the received output of the last convolutional layer and contains 512 neurons. It is followed by a ReLU and a dropout layer.
- CONCAT:** a concatenation layer aimed to concatenate the output of the 3 fully-connected above mentioned.
- FC (6):** this fully connected layer, containing again 1024 neurons, receives the 1536-dimensional output of the previous fully connected. It is followed by a ReLU and a dropout layer.
- FC (7):** this fully connected layer, containing again 768 neurons, receives the 1024-dimensional output of the previous fully connected. It is followed by a ReLU and a dropout layer.
- FC (8):** this is the last step toward the mapping of the input image to the age categories (8 output layer).

As a final step, the output of the last fully connected layer is given as input to a soft max layer that produces an output in terms of probability for each class. The final prediction is the class corresponding to the maximum value of a soft-max layer.

## 4 Experimental Results

In order to evaluate the performance of the proposed architecture and its computational load, k-fold validation and tests on multiple hardware platforms have been carried out on the Adience dataset [7] (details in Table 1). The dataset consists of images automatically uploaded to Flickr from smart-phone devices.

**Table 1.** Adience face dataset: face distribution among different age classes.

Age group	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60+	Total
(Label)	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Male	745	928	934	734	2308	1294	392	442	8192
Female	682	1234	1360	919	2589	1056	442	427	9411
Both	1427	2162	2294	1653	4897	2350	825	869	19487

The entire Adience collection includes roughly 26K images of 2,284 subjects. However it represents an highly challenging dataset (due to the presence of make-up, accessories and extremely rotated poses). In order to stress the performance of the proposed solution, the in-plane aligned version of the faces, originally used in [7], has been employed. In this way it is possible to highlight the generalization capabilities of the proposed solution, enhancing the advantage of a multiscale approach in unconstrained contexts. The training/testing procedure was based on a cross subject-exclusive validation that means that the same subject was not present, even with different images, in both the test and training sets used in a training/test round. More specifically the dataset under investigation was randomly split into  $k$  folds (the folding provided by [18]) with  $k = 5$  has been employed. For each of the  $k$  validation steps,  $k - 1$  folds were used for training, whereas the remaining fold was used for evaluating the estimation/validation capabilities.

The training was performed with a backward optimization process aided by a stochastic gradient descend (SGD). More precisely a batch size of 50 elements and a momentum  $\mu = 0.9$  were used.

Both training and testing procedures were performed using the Caffe open source framework [13] exploiting a Nvidia GTX 960 equipped with 1024 cores and 4 GB of video memory.

The outcomes were expressed in form of confusion matrix (computed as mean on 5 folds) both for the propose solution (Table 2) and the solution in [7] (Table 3). Moreover, the accuracies referred to the precise class and to the 1-class off were reported at the and of the tables and resumed in Fig. 3.

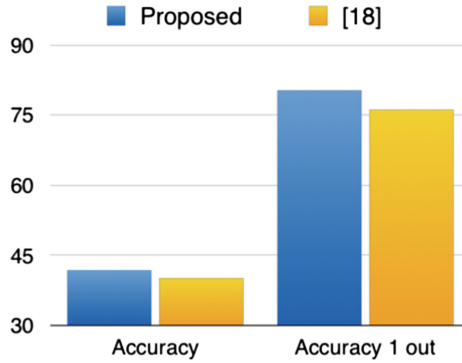
In order to be as fair as possible, the average accuracies have been computed as means of the confusion matrix diagonals, avoiding the unreliable results eventually due to the unbalancing of the classes.

**Table 2.** Accuracy performance on multi-branch network (proposed) approach: confusion matrix of *predicted* vs *real* classes; the last row reports the exact accuracy and the 1 class-off accuracy.

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60+
0-2	68.39%	28.32%	1.98%	0%	0.95%	0.08%	0.2%	0.08%
4-6	14.01%	64.4%	18.91%	0.32%	1.85%	0.32%	0.05%	0.14%
8-13	1.17%	24.48%	58.44%	4.6%	10.32%	0.75%	0.09%	0.14%
15-20	0.3%	5.79%	18.4%	15.74%	56.21%	3.26%	0.18%	0.12%
25-32	0.2%	1.49%	9.19%	8.26%	73.33%	7.03%	0.34%	0.14%
38-43	0.43%	2%	8.43%	4.26%	62.48%	18.3%	2.96%	1.13%
48-53	0.6%	1.81%	9.98%	1.57%	30.48%	40.24%	9.88%	5.54%
60+	0.23%	2.74%	7.77%	2.06%	13.83%	30.63%	16.46%	26.29%
Acc: 41.85%					Acc (1 out): 80.33%			

**Table 3.** Accuracy performance on single-branch network [18] approach: confusion matrix of *predicted* vs *real* classes; the last row reports the exact accuracy and the 1 class-off accuracy.

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60+
0-2	73.5%	23.01%	1.9%	0%	0.87%	0.44%	0.16%	0.12%
4-6	16.5%	61.49%	18.12%	0.55%	2.96%	0.18%	0%	0.18%
8-13	1.92%	21.95%	50.7%	7.46%	16.42%	1.31%	0.19%	0.05%
15-20	0.48%	5.91%	15.68%	15.98%	57.48%	3.74%	0.6%	0.12%
25-32	0.4%	1.66%	9.17%	7.76%	74.26%	5.84%	0.73%	0.18%
38-43	0.78%	2.65%	7.91%	2.7%	65.96%	14.91%	3.57%	1.52%
48-53	0.24%	2.53%	6.99%	3.37%	45.06%	28.43%	8.55%	4.82%
60+	0.8%	3.31%	6.63%	2.97%	23.2%	29.6%	12%	21.49%
Acc: 40.11%					Acc (1 out): 76.18%			



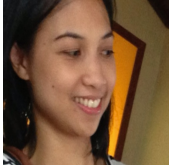



**Fig. 3.** Exact accuracy and the 1 class-off accuracy

A quick analysis of the results immediately highlights the superiority of the proposed approach respect to the single branch approach in [18]. Both the accuracies show that the use of multiple scales allows to improve the classification performance (+1.74% in exact class and +4.15% in 1 class-off class). Moreover, the greater increment on the 1 class-off accuracy proves as the proposed method allows a general polarization towards the right class.

Going deeper in the observation of confusion matrices it is possible to observe that, concerning the majority of the age classes, the proposed network guarantees an increment in accuracy. The most relevant improvement is obtained on the 60+ class (exact class = 4.8%, 1 class-off = 9.26) where the notable spread in age is well managed by the analysis on multiple scales. On the contrary, the 0-2 class is characterized by a narrow age gap and it is quite close to the successive one 4-6. This makes it suffering of the proposed approach that, catching more information, tends to spread the results leading to a sensible polarization toward



**Table 4.** Some example of sample wrongly classified by the single branch approach but correctly recognized by the proposed multi-branch structure. GT is the ground truth and the numbers are the labels referred in Table 1

				
GT	4	7	2	7
[18]	3	5	4	6
Proposed	4	7	2	7

the 4–6 class. Anyway, looking to the 1 class-off results for the first class, we can see as the proposed approach gets an accuracy of 96.71% against the 96.51%.

Table 4 shows some examples of images wrongly classified by the single branch approach that the multi-scale analysis has been capable to classify on the correct class. It is clear as orientation, illumination, quality and scale are some of the most evident issues on these samples. The proposed network is probably able to exploit its multi-branch structure to make the trained model more suitable to the aforementioned variations.

As an additional investigation, the execution time on different hardware configurations, has been computed for both the proposed network and the solution in [7]. Resulting computational times are reported in Table 5. More precisely the test have been performed on an i7 desktop pc equipped with 8 GB of RAM and, alternatively, on an NVIDIA Titan X and a NVIDIA GTX 970. The execution times have been moreover measured on the NVIDIA Jaton TX1 embedded system. It is clear as the proposed architecture introduces an increment of the computational time, that is, anyway, largely into the magnitude of the *real time processing*.

**Table 5.** Processing time: both the network was tested on 2 desktop configuration and (Nvidia Titan X and Nvidia GTX 970) and an embedded solution Nvidia Jaton TX1

	Execution time [ms]			Model size [MB]
	Jatson TX1	GTX 960	Titan X	
[18]	20.5	3.5	1.4	46
Proposed	70	12	5.3	243

## 5 Conclusions

Deep Learning, especially in computer vision issues, forcefully replaced classical approaches based on local feature extraction, allowing recognition to easily move

toward unconstrained environments. Anyway the lack in data availability sets a serious limit to the development of very deep networks that are constrained to be mildly deep in order to avoid overfitting issues. In this work, the possibility to exploit a multi-branch structure capable to keep the overfitting under control, but meanwhile to take advantage of a multiple scale analysis, has been treated.

Results proved the validity of the suggested solution capable to outperform the leading state of the art solutions (at the best of our knowledge) increasing the average accuracy up to 4.15% for the 1-off class estimation.

## References

1. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions using equivalence relations. In: ICML, vol. 3, pp. 11–18 (2003)
2. Carcagnì, P., Cazzato, D., Del Coco, M., Leo, M., Pioggia, G., Distante, C.: Real-time gender based behavior system for human-robot interaction. In: Beetz, M., Johnston, B., Williams, M.-A. (eds.) ICSR 2014. LNCS, vol. 8755, pp. 74–83. Springer, Cham (2014). doi:[10.1007/978-3-319-11973-1\\_8](https://doi.org/10.1007/978-3-319-11973-1_8)
3. Carcagnì, P., Del Coco, M., Cazzato, D., Leo, M., Distante, C.: A study on different experimental configurations for age, race, and gender estimation problems. *EURASIP J. Image Video Process.* **2015**(1), 37 (2015)
4. Carcagnì, P., Cazzato, D., Del Coco, M., Distante, C., Leo, M.: Visual interaction including biometrics information for a socially assistive robotic platform. In: Second Workshop on Assistive Computer Vision and Robotics (ACVR) (2014)
5. Chao, W.L., Liu, J.Z., Ding, J.J.: Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recogn.* **46**(3), 628–641 (2013)
6. Choi, S.E., Lee, Y.J., Lee, S.J., Park, K.R., Kim, J.: Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recogn.* **44**(6), 1262–1281 (2011)
7. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.* **9**(12), 2170–2179 (2014)
8. Gao, F., Ai, H.: Face age classification on consumer images with gabor feature and fuzzy LDA method. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 132–141. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-01793-3\\_14](https://doi.org/10.1007/978-3-642-01793-3_14)
9. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(12), 2234–2240 (2007)
10. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Process.* **17**(7), 1178–1188 (2008)
11. Han, H., Otto, C., Jain, A.K.: Age estimation from face images: human vs. machine performance. In: 2013 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2013)
12. He, X., Niyogi, P.: Locality preserving projections. In: NIPS, vol. 16 (2003)
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint [arXiv:1408.5093](https://arxiv.org/abs/1408.5093) (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)

15. Kwon, Y.H., et al.: Age classification from facial images. In: 1994 IEEE Computer Vision and Pattern Recognition (CVPR), pp. 762–767 (1994)
16. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**(4), 541–551 (1989)
17. Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.: Computer vision for assistive technologies. *Comput. Vis. Image Underst.* **154**, 1–15 (2017)
18. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–42 (2015)
19. Ramanathan, N., Chellappa, R.: Modeling age progression in young faces. In: Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 387–394. IEEE (2006)
20. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition (2008)
21. Yan, S., Liu, M., Huang, T.S.: Extracting age information from local spatially flexible patches. In: IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP), pp. 737–740 (2008)
22. Yan, S., Zhou, X., Liu, M., Hasegawa-Johnson, M., Huang, T.S.: Regression from patch-kernel. In: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
23. Zhuang, X., Zhou, X., Hasegawa-Johnson, M., Huang, T.: Face age estimation using patch-based hidden Markov model supervectors. In: 19th International Conference on Pattern Recognition (ICPR), pp. 1–4. IEEE (2008)