Age and Gender Classification using Convolutional Neural Networks

Gil Levi and Tal Hassner
Department of Mathematics and Computer Science
The Open University of Israel
gil.levi100@gmail.com hassner@openu.ac.il

Abstract

Automatic age and gender classification has become relevant to an increasing amount of applications, particularly since the rise of social platforms and social media. Nevertheless, performance of existing methods on real-world images is still significantly lacking, especially when compared to the tremendous leaps in performance recently reported for the related task of face recognition. In this paper we show that by learning representations through the use of deep-convolutional neural networks (CNN), a significant increase in performance can be obtained on these tasks. To this end, we propose a simple convolutional net architecture that can be used even when the amount of learning data is limited. We evaluate our method on the recent Adience benchmark for age and gender estimation and show it to dramatically outperform current state-of-the-art methods.

1. Introduction

Age and gender play fundamental roles in social interactions. Languages reserve different salutations and grammar rules for men or women, and very often different vocabularies are used when addressing elders compared to young people. Despite the basic roles these attributes play in our day-to-day lives, the ability to automatically estimate them accurately and reliably from face images is still far from meeting the needs of commercial applications. This is particularly perplexing when considering recent claims to super-human capabilities in the related task of face recognition (e.g., [48]).

Past approaches to estimating or classifying these attributes from face images have relied on differences in facial feature dimensions [29] or "tailored" face descriptors (e.g., [10, 15, 32]). Most have employed classification schemes designed particularly for age or gender estimation tasks, including [4] and others. Few of these past methods were designed to handle the many challenges of unconstrained imaging conditions [10]. Moreover, the machine learning methods employed by these systems did not fully exploit the massive numbers of image examples and data available through the Internet in order to improve classification capabilities.

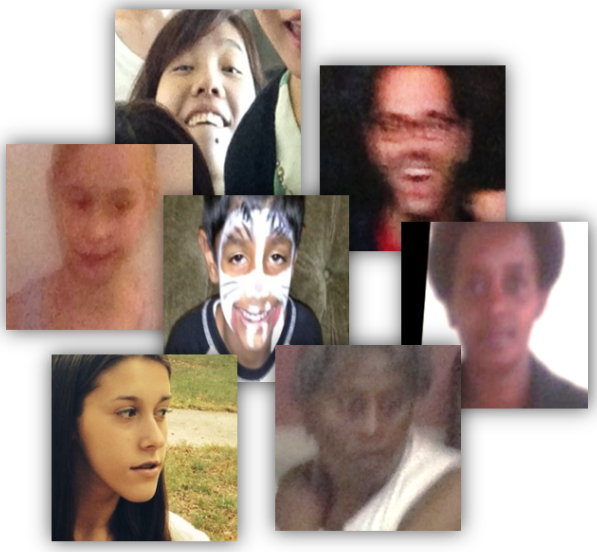

Figure 1. **Faces from the Adience benchmark for age and gender classification [10].** These images represent some of the challenges of age and gender estimation from real-world, unconstrained images. Most notably, extreme blur (low-resolution), occlusions, out-of-plane pose variations, expressions and more.

In this paper we attempt to close the gap between automatic face recognition capabilities and those of age and gender estimation methods. To this end, we follow the successful example laid down by recent face recognition systems: Face recognition techniques described in the last few years have shown that tremendous progress can be made by the use of deep convolutional neural networks (CNN) [31]. We demonstrate similar gains with a simple network architecture, designed by considering the rather limited availability of accurate age and gender labels in existing face data sets.

We test our network on the newly released Adience

benchmark for age and gender classification of unfiltered face images [10]. We show that despite the very challenging nature of the images in the Adience set and the simplicity of our network design, our method outperforms existing state of the art by substantial margins. Although these results provide a remarkable baseline for deep-learning-based approaches, they leave room for improvements by more elaborate system designs, suggesting that the problem of accurately estimating age and gender in the unconstrained settings, as reflected by the Adience images, remains unsolved. In order to provide a foothold for the development of more effective future methods, we make our trained models and classification system publicly available. For more information, please see the project webpage www.openu.ac.il/home/hassner/projects/cnn_agegender.

2. Related Work

Before describing the proposed method we briefly review related methods for age and gender classification and provide a cursory overview of deep convolutional networks.

2.1. Age and Gender Classification

Age classification. The problem of automatically extracting age related attributes from facial images has received increasing attention in recent years and many methods have been put fourth. A detailed survey of such methods can be found in [11] and, more recently, in [21]. We note that despite our focus here on age group *classification* rather than precise age estimation (i.e., age regression), the survey below includes methods designed for either task.

Early methods for age estimation are based on calculating ratios between different measurements of facial features [29]. Once facial features (e.g. eyes, nose, mouth, chin, etc.) are localized and their sizes and distances measured, ratios between them are calculated and used for classifying the face into different age categories according to hand-crafted rules. More recently, [41] uses a similar approach to model age progression in subjects under 18 years old. As those methods require accurate localization of facial features, a challenging problem by itself, they are unsuitable for in-the-wild images which one may expect to find on social platforms.

On a different line of work are methods that represent the aging process as a subspace [16] or a manifold [19]. A drawback of those methods is that they require input images to be near-frontal and well-aligned. These methods therefore present experimental results only on constrained data-sets of near-frontal images (e.g UIUC-IFP-Y [12, 19] ,FG-NET [30] and MORPH [43]). Again, as a consequence, such methods are ill-suited for unconstrained images.

Different from those described above are methods that use local features for representing face images. In [55] Gaussian Mixture Models (GMM) [13] were used to represent the distribution of facial patches. In [54] GMM were used again for representing the distribution of local facial measurements, but robust descriptors were used instead of pixel patches. Finally, instead of GMM, Hidden-Markov-Model, super-vectors [40] were used in [56] for representing face patch distributions.

An alternative to the local image intensity patches are robust image descriptors: Gabor image descriptors [32] were used in [15] along with a Fuzzy-LDA classifier which considers a face image as belonging to more than one age class. In [20] a combination of Biologically-Inspired Features (BIF) [44] and various manifold-learning methods were used for age estimation. Gabor [32] and local binary patterns (LBP) [1] features were used in [7] along with a hierarchical age classifier composed of Support Vector Machines (SVM) [9] to classify the input image to an age-class followed by a support vector regression [52] to estimate a precise age.

Finally, [4] proposed improved versions of relevant component analysis [3] and locally preserving projections [36]. Those methods are used for distance learning and dimensionality reduction, respectively, with Active Appearance Models [8] as an image feature.

All of these methods have proven effective on small and/or constrained benchmarks for age estimation. To our knowledge, the best performing methods were demonstrated on the Group Photos benchmark [14]. In [10] state-of-the-art performance on this benchmark was presented by employing LBP descriptor variations [53] and a dropout-SVM classifier. We show our proposed method to outperform the results they report on the more challenging Adience benchmark, designed for the same task.

Gender classification. A detailed survey of gender classification methods can be found in [34] and more recently in [42]. Here we quickly survey relevant methods.

One of the early methods for gender classification [17] used a neural network trained on a small set of near-frontal face images. In [37] the combined 3D structure of the head (obtained using a laser scanner) and image intensities were used for classifying gender. SVM classifiers were used by [35], applied directly to image intensities. Rather than using SVM, [2] used AdaBoost for the same purpose, here again, applied to image intensities. Finally, viewpoint-invariant age and gender classification was presented by [49].

More recently, [51] used the Webers Local texture Descriptor [6] for gender recognition, demonstrating near-perfect performance on the FERET benchmark [39]. In [38], intensity, shape and texture features were used with mutual information, again obtaining near-perfect results on the FERET benchmark.

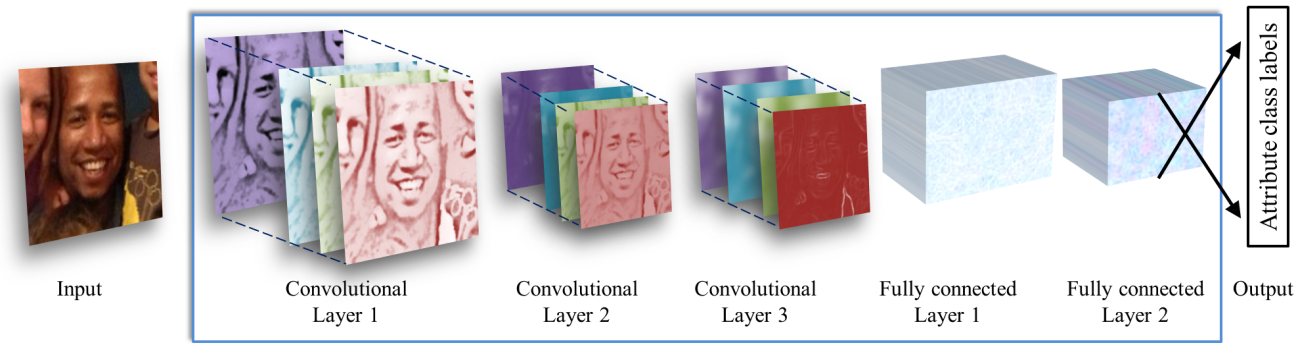

Figure 2. **Illustration of our CNN architecture.** The network contains three convolutional layers, each followed by a rectified linear operation and pooling layer. The first two layers also follow normalization using local response normalization [28]. The first Convolutional Layer contains 96 filters of 7×7 pixels, the second Convolutional Layer contains 256 filters of 5×5 pixels, The third and final Convolutional Layer contains 384 filters of 3×3 pixels. Finally, two fully-connected layers are added, each containing 512 neurons. See Figure 3 for a detailed schematic view and the text for more information.

Most of the methods discussed above used the FERET benchmark [39] both to develop the proposed systems and to evaluate performances. FERET images were taken under highly controlled condition and are therefore much less challenging than in-the-wild face images. Moreover, the results obtained on this benchmark suggest that it is saturated and not challenging for modern methods. It is therefore difficult to estimate the actual relative benefit of these techniques. As a consequence, [46] experimented on the popular Labeled Faces in the Wild (LFW) [25] benchmark, primarily used for face recognition. Their method is a combination of LBP features with an AdaBoost classifier.

As with age estimation, here too, we focus on the Adience set which contains images more challenging than those provided by LFW, reporting performance using a more robust system, designed to better exploit information from massive example training sets.

2.2. Deep convolutional neural networks

One of the first applications of convolutional neural networks (CNN) is perhaps the LeNet-5 network described by [31] for optical character recognition. Compared to modern deep CNN, their network was relatively modest due to the limited computational resources of the time and the algorithmic challenges of training bigger networks.

Though much potential laid in deeper CNN architectures (networks with more neuron layers), only recently have they became prevalent, following the dramatic increase in both the computational power (due to Graphical Processing Units), the amount of training data readily available on the Internet, and the development of more effective methods for training such complex models. One recent and notable examples is the use of deep CNN for image classification on the challenging Imagenet benchmark [28]. Deep CNN have additionally been successfully applied to applications including human pose estimation [50], face parsing [33], facial keypoint detection [47], speech recognition [18] and action classification [27]. To our knowledge, this is the first report of their application to the tasks of age and gender classification from unconstrained photos.

3. A CNN for age and gender estimation

Gathering a large, *labeled* image training set for age and gender estimation from social image repositories requires either access to personal information on the subjects appearing in the images (their birth date and gender), which is often private, or is tedious and time-consuming to manually label. Data-sets for age and gender estimation from real-world social images are therefore relatively limited in size and presently no match in size with the much larger image classification data-sets (e.g. the Imagenet dataset [45]). Overfitting is common problem when machine learning based methods are used on such small image collections. This problem is exacerbated when considering deep convolutional neural networks due to their huge numbers of model parameters. Care must therefore be taken in order to avoid overfitting under such circumstances.

3.1. Network architecture

Our proposed network architecture is used throughout our experiments for both age and gender classification. It is illustrated in Figure 2. A more detailed, schematic diagram of the entire network design is additionally provided in Figure 3. The network comprises of only three convolutional layers and two fully-connected layers with a small number of neurons. This, by comparison to the much larger architectures applied, for example, in [28] and [5]. Our choice of a smaller network design is motivated both from our desire to reduce the risk of overfitting as well as the nature

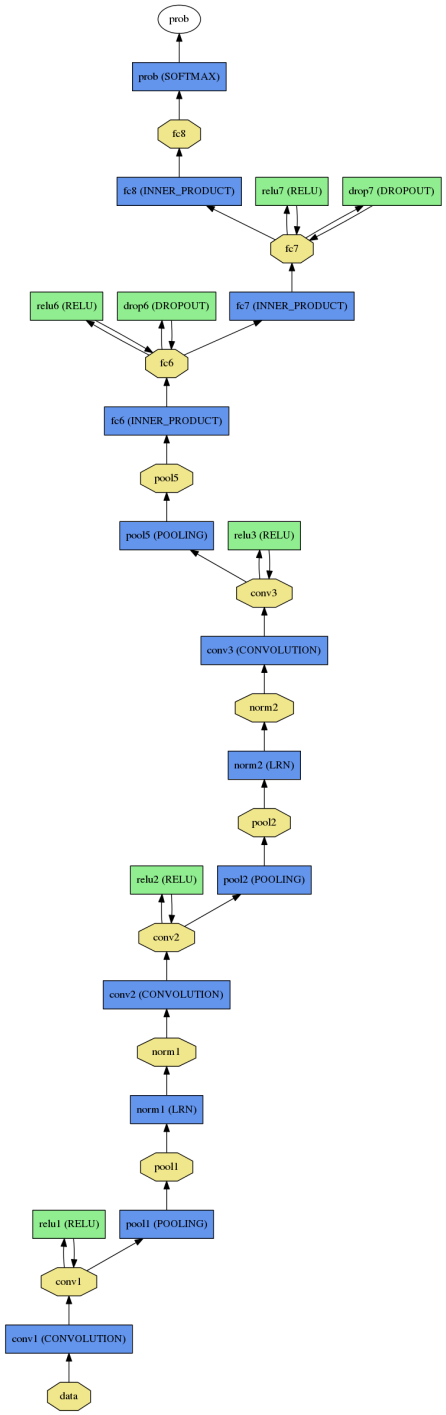

Figure 3. **Full schematic diagram of our network architecture.** Please see text for more details.

of the problems we are attempting to solve: age classification on the Adience set requires distinguishing between eight classes; gender only two. This, compared to, e.g., the ten thousand identity classes used to train the network used for face recognition in [48].

All three color channels are processed directly by the network. Images are first rescaled to 256×256 and a crop of 227×227 is fed to the network. The three subsequent convolutional layers are then defined as follows.

1. 96 filters of size $3 \times 7 \times 7$ pixels are applied to the input in the first convolutional layer, followed by a rectified linear operator (ReLU), a max pooling layer taking the maximal value of 3×3 regions with two-pixel strides and a local response normalization layer [28].

2. The $96 \times 28 \times 28$ output of the previous layer is then processed by the second convolutional layer, containing 256 filters of size $96 \times 5 \times 5$ pixels. Again, this is followed by ReLU, a max pooling layer and a local response normalization layer with the same hyper parameters as before.

3. Finally, the third and last convolutional layer operates on the $256 \times 14 \times 14$ blob by applying a set of 384 filters of size $256 \times 3 \times 3$ pixels, followed by ReLU and a max pooling layer.

The following fully connected layers are then defined by:

4. A first fully connected layer that receives the output of the third convolutional layer and contains 512 neurons, followed by a ReLU and a dropout layer.

5. A second fully connected layer that receives the 512-dimensional output of the first fully connected layer and again contains 512 neurons, followed by a ReLU and a dropout layer.

6. A third, fully connected layer which maps to the final classes for age or gender.

Finally, the output of the last fully connected layer is fed to a soft-max layer that assigns a probability for each class. The prediction itself is made by taking the class with the maximal probability for the given test image.

3.2. Testing and training

Initialization. The weights in all layers are initialized with random values from a zero mean Gaussian with standard deviation of 0.01. To stress this, we do not use pre-trained models for initializing the network; the network is trained, from scratch, without using any data outside of the images and the labels available by the benchmark. This, again, should be compared with CNN implementations used for face recognition, where hundreds of thousands of images are used for training [48].

Target values for training are represented as sparse, binary vectors corresponding to the ground truth classes. For each training image, the target, label vector is in the length

of the number of classes (two for gender, eight for the eight age classes of the age classification task), containing 1 in the index of the ground truth and 0 elsewhere.

Network training. Aside from our use of a lean network architecture, we apply two additional methods to further limit the risk of overfitting. First we apply dropout learning [24] (i.e. randomly setting the output value of network neurons to zero). The network includes two dropout layers with a dropout ratio of 0.5 (50% chance of setting a neuron's output value to zero). Second, we use data-augmentation by taking a random crop of 227×227 pixels from the 256×256 input image and randomly mirror it in each forward-backward training pass. This, similarly to the multiple crop and mirror variations used by [48].

Training itself is performed using stochastic gradient decent with image batch size of fifty images. The initial learning rate is $e^{-}3$, reduced to $e^{-}4$ after 10K iterations.

Prediction. We experimented with two methods of using the network in order to produce age and gender predictions for novel faces:

- **Center Crop:** Feeding the network with the face image, cropped to 227×227 around the face center.
- **Over-sampling:** We extract five 227×227 pixel crop regions, four from the corners of the 256×256 face image, and an additional crop region from the center of the face. The network is presented with all five images, along with their horizontal reflections. Its final prediction is taken to be the average prediction value across all these variations.

We have found that small misalignments in the Adience images, caused by the many challenges of these images (occlusions, motion blur, etc.) can have a noticeable impact on the quality of our results. This second, over-sampling method, is designed to compensate for these small misalignments, bypassing the need for improving alignment quality, but rather directly feeding the network with multiple translated versions of the same face.

4. Experiments

Our method is implemented using the Caffe open-source framework [26]. Training was performed on an Amazon GPU machine with 1,536 CUDA cores and 4GB of video memory. Training each network required about four hours, predicting age or gender on a single image using our network requires about 200ms. Prediction running times can conceivably be substantially improved by running the network on image batches.

4.1. The Adience benchmark

We test the accuracy of our CNN design using the recently released Adience benchmark [10], designed for age and gender classification. The Adience set consists of images automatically uploaded to Flickr from smart-phone devices. Because these images were uploaded without prior manual filtering, as is typically the case on media webpages (e.g., images from the LFW collection [25]) or social websites (the Group Photos set [14]), viewing conditions in these images are highly unconstrained, reflecting many of the real-world challenges of faces appearing in Internet images. Adience images therefore capture extreme variations in head pose, lightning conditions quality, and more.

The entire Adience collection includes roughly 26K images of 2,284 subjects. Table 1 lists the breakdown of the collection into the different age categories. Testing for both age or gender classification is performed using a standard five-fold, subject-exclusive cross-validation protocol, defined in [10]. We use the in-plane aligned version of the faces, originally used in [10]. These images are used rater than newer alignment techniques in order to highlight the performance gain attributed to the network architecture, rather than better preprocessing.

We emphasize that the same network architecture is used for all test folds of the benchmark and in fact, for both gender and age classification tasks. This is performed in order to ensure the validity of our results across folds, but also to demonstrate the generality of the network design proposed here; the same architecture performs well across different, related problems.

We compare previously reported results to the results computed by our network. Our results include both methods for testing: center-crop and over-sampling (Section 3).

4.2. Results

Table 2 and Table 3 presents our results for gender and age classification respectively. Table 4 further provides a confusion matrix for our multi-class age classification results. For age classification, we measure and compare both the accuracy when the algorithm gives the exact age-group classification and when the algorithm is off by one adjacent age-group (i.e., the subject belongs to the group immediately older or immediately younger than the predicted group). This follows others who have done so in the past, and reflects the uncertainty inherent to the task – facial features often change very little between oldest faces in one age class and the youngest faces of the subsequent class.

Both tables compare performance with the methods described in [10]. Table 2 also provides a comparison with [23] which used the same gender classification pipeline of [10] applied to more effective alignment of the faces; faces in their tests were synthetically modified to appear facing forward.

Figure 4. **Gender misclassifications.** Top row: Female subjects mistakenly classified as males. Bottom row: Male subjects mistakenly classified as females

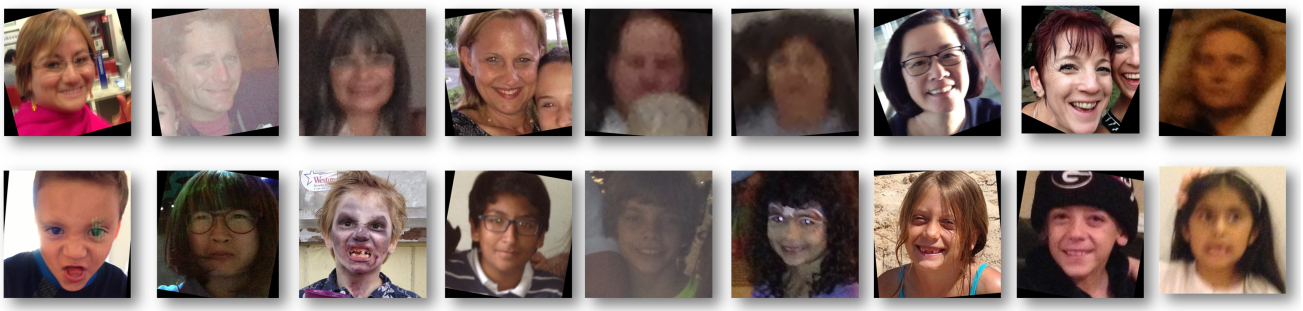

Figure 5. **Age misclassifications.** Top row: Older subjects mistakenly classified as younger. Bottom row: Younger subjects mistakenly classified as older.

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-	Total
Male	745	928	934	734	2308	1294	392	442	8192
Female	682	1234	1360	919	2589	1056	433	427	9411
Both	1427	2162	2294	1653	4897	2350	825	869	19487

Table 1. **The AdienceFaces benchmark.** Breakdown of the AdienceFaces benchmark into the different Age and Gender classes.

Method	Accuracy
Best from [10]	77.8 ± 1.3
Best from [23]	79.3 ± 0.0
Proposed using single crop	85.9 ± 1.4
Proposed using over-sample	**86.8 ± 1.4**

Table 2. **Gender estimation results on the Adience benchmark.** Listed are the mean accuracy ± standard error over all age categories. Best results are marked in bold.

Method	Exact	1-off
Best from [10]	45.1 ± 2.6	79.5 ±1.4
Proposed using single crop	49.5 ± 4.4	84.6 ± 1.7
Proposed using over-sample	**50.7 ± 5.1**	**84.7 ± 2.2**

Table 3. **Age estimation results on the Adience benchmark.** Listed are the mean accuracy ± standard error over all age categories. Best results are marked in bold.

Evidently, the proposed method outperforms the reported state-of-the-art on both tasks with considerable gaps. Also evident is the contribution of the over-sampling approach, which provides an additional performance boost over the original network. This implies that better alignment (e.g., frontalization [22, 23]) may provide an additional boost in performance.

We provide a few examples of both gender and age misclassifications in Figures 4 and 5, respectively. These show that many of the mistakes made by our system are due to extremely challenging viewing conditions of some of the Adience benchmark images. Most notable are mistakes caused by blur or low resolution and occlusions (particularly from heavy makeup). Gender estimation mistakes also frequently occur for images of babies or very young children where obvious gender attributes are not yet visible.

5. Conclusions

Though many previous methods have addressed the problems of age and gender classification, until recently, much of this work has focused on constrained images taken in lab settings. Such settings do not adequately reflect appearance variations common to the real-world images in social websites and online repositories. Internet images, however, are not simply more challenging: they are also abun-

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-
0-2	**0.699**	0.147	0.028	0.006	0.005	0.008	0.007	0.009
4-6	0.256	**0.573**	0.166	0.023	0.010	0.011	0.010	0.005
8-13	0.027	0.223	**0.552**	0.150	0.091	0.068	0.055	0.061
15-20	0.003	0.019	0.081	**0.239**	0.106	0.055	0.049	0.028
25-32	0.006	0.029	0.138	0.510	**0.613**	0.461	0.260	0.108
38-43	0.004	0.007	0.023	0.058	0.149	**0.293**	0.339	0.268
48-53	0.002	0.001	0.004	0.007	0.017	0.055	**0.146**	0.165
60-	0.001	0.001	0.008	0.007	0.009	0.050	0.134	**0.357**

Table 4. **Age estimation confusion matrix on the Adience benchmark.**

dant. The easy availability of huge image collections provides modern machine learning based systems with effectively endless training data, though this data is not always suitably labeled for supervised learning.

Taking example from the related problem of face recognition we explore how well deep CNN perform on these tasks using Internet data. We provide results with a lean deep-learning architecture designed to avoid overfitting due to the limitation of limited labeled data. Our network is "shallow" compared to some of the recent network architectures, thereby reducing the number of its parameters and the chance for overfitting. We further inflate the size of the training data by artificially adding cropped versions of the images in our training set. The resulting system was tested on the Adience benchmark of unfiltered images and shown to significantly outperform recent state of the art.

Two important conclusions can be made from our results. First, CNN can be used to provide improved age and gender classification results, even considering the much smaller size of contemporary unconstrained image sets labeled for age and gender. Second, the simplicity of our model implies that more elaborate systems using more training data may well be capable of substantially improving results beyond those reported here.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600010. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006. 2

[2] S. Baluja and H. A. Rowley. Boosting sex identification performance. *Int. J. Comput. Vision*, 71(1):111–119, 2007. 2

[3] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Int. Conf. Mach. Learning*, volume 3, pages 11–18, 2003. 2

[4] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013. 1, 2

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. 3

[6] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao. Wld: A robust local image descriptor. *Trans. Pattern Anal. Mach. Intell.*, 32(9):1705–1720, 2010. 2

[7] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim. Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition*, 44(6):1262–1281, 2011. 2

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conf. Comput. Vision*, pages 484–498. Springer, 1998. 2

[9] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 2

[10] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Trans. on Inform. Forensics and Security*, 9(12), 2014. 1, 2, 5, 6

[11] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *Trans. Pattern Anal. Mach. Intell.*, 32(11):1955–1976, 2010. 2

[12] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *Int. Conf. Multimedia*, 10(4):578–584, 2008. 2

[13] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1991. 2

[14] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 256–263. IEEE, 2009. 2, 5

[15] F. Gao and H. Ai. Face age classification on consumer images with gabor feature and fuzzy lda method. In *Advances in biometrics*, pages 132–141. Springer, 2009. 1, 2

[16] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *Trans. Pattern Anal. Mach. Intell.*, 29(12):2234–2240, 2007. 2

[17] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Neural Inform. Process. Syst.*, pages 572–579, 1990. 2

[18] A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE, 2013. 3

[19] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *Trans. Image Processing*, 17(7):1178–1188, 2008. 2

[20] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In *Proc. Int. Conf. Comput. Vision*, pages 1986–1991. IEEE, 2009. 2

[21] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *Biometrics (ICB), 2013 International Conference on*. IEEE, 2013. 2

[22] T. Hassner. Viewing real-world faces in 3d. In *Proc. Int. Conf. Comput. Vision*, pages 3607–3614. IEEE, 2013. 6

[23] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. *Proc. Conf. Comput. Vision Pattern Recognition*, 2015. 5, 6

[24] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 5

[25] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 3, 5

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[27] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 1725–1732. IEEE, 2014. 3

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Inform. Process. Syst.*, pages 1097–1105, 2012. 3, 4

[29] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 762–767. IEEE, 1994. 1, 2

[30] A. Lanitis. The FG-NET aging database, 2002. Available: www-prima.inrialpes.fr/FGnet/html/benchmarks.html. 2

[31] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 1, 3

[32] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Trans. Image Processing*, 11(4):467–476, 2002. 1, 2

[33] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 2480–2487. IEEE, 2012. 3

[34] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *Trans. Pattern Anal. Mach. Intell.*, 30(3):541–547, 2008. 2

[35] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *Trans. Pattern Anal. Mach. Intell.*, 24(5):707–711, 2002. 2

[36] X. Niyogi. Locality preserving projections. In *Neural Inform. Process. Syst.*, volume 16, page 153. MIT, 2004. 2

[37] A. J. O'toole, T. Vetter, N. F. Troje, H. H. Bülthoff, et al. Sex classification is better with three-dimensional head structure than with image intensity information. *Perception*, 26:75–84, 1997. 2

[38] C. Perez, J. Tapia, P. Estévez, and C. Held. Gender classification from face images using mutual information and feature fusion. *International Journal of Optomechatronics*, 6(1):92–119, 2012. 2

[39] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998. 2, 3

[40] L. Rabiner and B.-H. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986. 2

[41] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In *Proc. Conf. Comput. Vision Pattern Recognition*, volume 1, pages 387–394. IEEE, 2006. 2

[42] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross. Soft biometrics for surveillance: an overview. *Machine learning: theory and applications. Elsevier*, pages 327–352, 2013. 2

[43] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 341–345. IEEE, 2006. 2

[44] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999. 2

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014. 3

[46] C. Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4):431–437, 2012. 3

[47] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 3476–3483. IEEE, 2013. 3

[48] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 1891–1898. IEEE, 2014. 1, 4, 5

[49] M. Toews and T. Arbel. Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion. *Trans. Pattern Anal. Mach. Intell.*, 31(9):1567–1581, 2009. 2

[50] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 1653–1660. IEEE, 2014. 3

[51] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, G. Bebis, and A. M. Mirza. Gender recognition from face images with local wld descriptor. In *Systems, Signals and Image Processing*, pages 417–420. IEEE, 2012. 2

[52] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998. 2

[53] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *post-ECCV Faces in Real-Life Images Workshop*, 2008. 2

[54] S. Yan, M. Liu, and T. S. Huang. Extracting age information from local spatially flexible patches. In *Acoustics, Speech and Signal Processing*, pages 737–740. IEEE, 2008. 2

[55] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang. Regression from patch-kernel. In *Proc. Conf. Comput. Vision Pattern Recognition*. IEEE, 2008. 2

[56] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. Huang. Face age estimation using patch-based hidden markov model supervectors. In *Int. Conf. Pattern Recognition*. IEEE, 2008. 2