Apparent Age Estimation Using Ensemble of Deep Learning Models

Refik Can Mallı[*] Mehmet Aygün[*] Hazım Kemal Ekenel

Istanbul Technical University
Istanbul, Turkey
{mallir,aygunme,ekenel}@itu.edu.tr

Abstract

In this paper, we address the problem of apparent age estimation. Different from estimating the real age of individuals, in which each face image has a single age label, in this problem, face images have multiple age labels, corresponding to the ages perceived by the annotators, when they look at these images. This provides an intriguing computer vision problem, since in generic image or object classification tasks, it is typical to have a single ground truth label per class. To account for multiple labels per image, instead of using average age of the annotated face image as the class label, we have grouped the face images that are within a specified age range. Using these age groups and their age-shifted groupings, we have trained an ensemble of deep learning models. Before feeding an input face image to a deep learning model, five facial landmark points are detected and used for 2-D alignment. We have employed and fine tuned convolutional neural networks (CNNs) that are based on VGG-16 [24] architecture and pretrained on the IMDB-WIKI dataset [22]. The outputs of these deep learning models are then combined to produce the final estimation. Proposed method achieves 0.3668 error in the final ChaLearn LAP 2016 challenge test set [5].

1. Introduction

Age estimation from face images is an interesting computer vision problem. It has various applications, ranging from customer relations to biometrics and entertainment. Although, there are various studies on real age estimation, [7, 6, 26] to name a few, problem of apparent age estimation from face images is a recently introduced topic, which has received significant attention [4, 20, 14, 22, 15].

Besides the common difficulties in facial image processing and analysis, such as pose and illumination, the main challenges associated with age estimation have been the im-
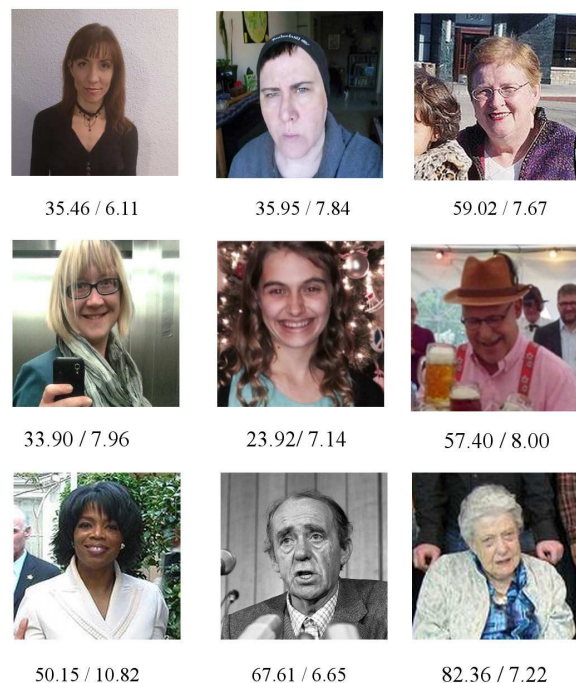

Figure 1. Sample images from the ChaLearn LAP 2016 apparent age estimation challenge dataset. Numbers below the images represent their average apparent age and standard deviations. The higher the standard deviation is, the harder it gets to predict apparent age of a person. Especially, in old age group, standard deviations are generally large, which makes it difficult to perform accurate age prediction from the images of old people. For instance, in LAP 2016 training set, there are 1095 images with standard deviations less than three, and only 31 of them belong to someone older than 40.

pact of ethnicity and subjective factors. The task of apparent age estimation alleviates the challenges posed by subjective factors. Since every person ages differently, real age may not be easy to deduce from face images. However, apparent age estimation is based on annotaters' perception of sub-

---

[*]Both authors contributed equally to this work.

jects' ages, who are displayed in the images. Therefore, the judgments for the age labes are expected to be based on visual appearance cues rather than personal characteristics. On the other hand, these perceived ages are subjective and they depend on the annotators, leading to result in multiple age labels for the same face image. This is a very intriguing problem, considering that for object classification problems, we normally have a single label per image. Sample images from the ChaLearn LAP 2016 dataset can be seen in Figure 1.

In this study, we have proposed a novel approach for apparent age estimation, which addresses the problem of imprecise, uncertain multiple labels by grouping the face images that are within a specified age range, and by training an ensemble of deep learning models using these age groups and their age-shifted groupings.

Convolutional Neural Networks (CNNs) have shown significant performance improvement in several computer vision problems, such as image classification [23], object detection [23], image segmentation [30], and face recognition [19]. Moreover, all the best performing systems proposed in ChaLearn LAP 2015 [4] were based on CNNs [22, 15, 31]. Due to these reasons, we have opted for CNNs for the proposed system and have conducted a thorough study to efficiently transfer already existing models for the problem at hand. In addition, ensemble models are known to increase performance further. To benefit from this, we have generated different age groupings and train multiple CNN models. Finally, we have combined the outputs of these CNN models to produce the final estimation result.

The contributions of this study can be summarized as follows: (i) We proposed an apparent age estimation system that takes into account the imprecise, uncertain multiple labels available for the task. Instead of using avarage age labels as class labels, we have grouped the face images that are within a specified age range. An ensemble of CNNs have been trained using these age groups and their age-shifted groupings. (ii) We have conducted an extensive assessment about transferability of existing CNN models for apparent age estimation. (iii) We have analyzed the apparent age estimation problem in detail and pointed the challenges associated to it.

The rest of the paper is organized as follows: In Section 2, a brief overview of related work is provided. In Section 3, the problem is stated and the challenges associated to it are pointed. The proposed method is explained in detail in Section 4. Experimental results are presented and discussed in Section 5. Finally, in Section 6, the paper is concluded with a brief summary and discussion.

2. Related Work

Deep convolutional neural networks have performed impressively on very large-scale image recognition prob-

lems [24, 9, 13]. Well known deep network architectures [24, 27] have been recently utilized in apparent age estimation studies.

Deep EXpectation of Apparent Age From a Single Image (DEX) [22], which uses the CNN architecture presented in [24] was the winner of ChaLearn LAP 2015 apparent age estimation challenge [5]. In this method, the problem is approached as a classification problem with 101 age classes. They fine tuned VGG-16 model, which was pretrained on ImageNet [23], using IMDB -WIKI [22] age dataset, that was created with the images crawled from the web. They splitted Chalearn LAP 2015 [4] data to 20 different groups and trained 20 different models using these groups. Finally, they combined the output of each of the network using the weighted sum. The faces were not aligned in this study. They only performed face detection [16] with different rotations and picked the one that had the best detection score. Also, they added 40% margin to the faces before cropping. In case they cannot detect the face in the image, they used the entire image. They outperformed human performance and won the ChaLearn LAP 2015 apparent age estimation challenge [5].

In contrast to the DEX method, AgeNet [15] approached to the problem as both classification and regression problem. They used large-scale deep convolutional neural networks [27] for both tasks. Firstly, they normalized ages by dividing them with 100 and added a sigmoid layer as the output of regression model, so that the output is also normalized to [0,1]. In their second model, they trained their networks for classification task using Gaussian label distribution. While Euclidean loss was used for training phase of the regression model, cross-entropy loss was used for the classification model. All of the models were pretrained on face identification dataset using the CASIA-WebFace [29] and on age datasets [2, 18, 21]. Face detection and facial landmark localization were performed before face normalization. Two different face normalization methods, namely Exterior and Interior, were applied to the images. In total, eight different models were trained. Four of them were used for regression and four of them were used for classification. Two regression and two classification networks used the Exterior face template and the remaining four regression/classification networks used the Interior representation. Also, two different face crop sizes were used for training different networks. Finally, the output of each network were fused to produce the final estimation.

In contrast to DEX and AgeNet, in which the CNNs were used both for feature extraction and classification, Zhu *et al*. [31] just used the CNNs as a feature extractor and employed additional classifiers, such as Support Vector Machines (SVM), Support Vector Regressors (SVR) [25], and Random Forests (RF) [1] for apparent age estimation. Similar to the AgeNet method, they also trained their network

first on the face identification dataset [29] and then on the age datasets [3, 21, 17, 2, 18]. They classified ages first into ten groups using the deep features that were extracted from CNN and were classified with SVM. After that, they used local age estimators, which used SVR and RF for the final predictions within each group.

3. Problem Definition and Challenges

The ChaLearn LAP dataset [5] is based on apparent ages, which are annotated by more than 10 people for every image. In contrast to real age labels, apparent age labels are sensitive to environmental conditions, where the picture has taken and also changes with respect to annotators' perception. These reasons make the ground truth age labels uncertain. Standard deviation of the average age scores of the images are shown in Figure 2. It can be observed that mean value of standard deviation per age label increases, when the age increases. This indicates that estimating apparent age of the subjects, who are older than 30 is more challenging.
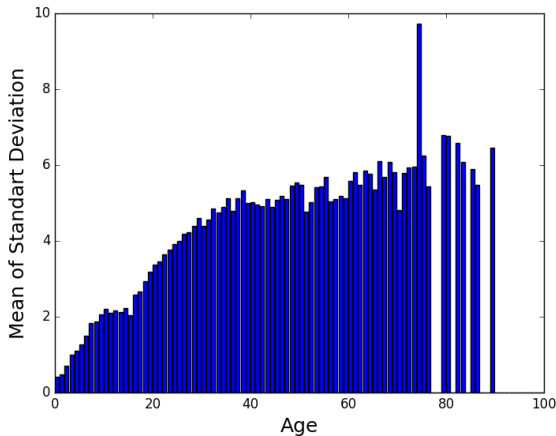
Figure 2. Mean value of standard deviation per age label increases when age labels increase.

The competition dataset consists of 7591 images and it is divided into training, test, and validation sets. The number of images in each set is given in Table 1. Training and validation sets are provided with corresponding average age estimates by the annotators and their standard variations. Test set is spared for the final submission and the systems' performances are assessed on it. One of the observation that can clearly be made from Figure 3 is that the number of images provided for each age is not equal. Imbalanced data distribution is an important problem that should be handled by extracting representative features from the data without overfitting [8].

Part of Dataset	Number of Images
Training	4113
Validation	1500
Test	1978
Total	7591

Table 1. Number of images provided in the ChaLearn LAP 2016 dataset.

Variation contained in the images, both in terms of image quality and facial appearance, is another challenge that prevents correct estimation of the apparent ages. Even though there are images that labeled with the same age, they could look significanly different from each other. Figure 4 displays face images, which are labeled with an identical age, from different genders and ethnicities. Moreover, pictures are taken in different lighting conditions and some of them has photo effects on them. There are images that only a part of the face could be seen and sometimes they are covered with hair. Some of the people also have make-up on their faces or they wear accessories such as hats and glasses.
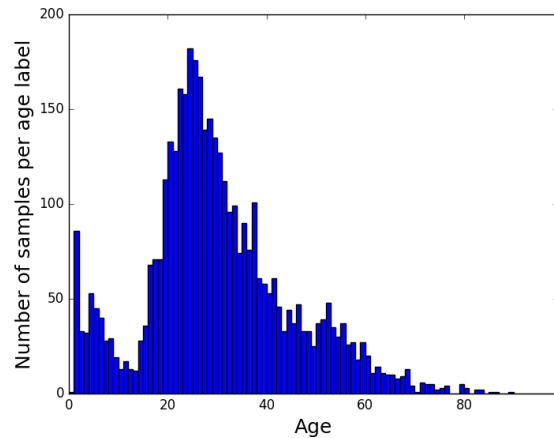
Figure 3. Age distribution of training set shows that dataset includes a lot of images between age of 20 and 40.

4. Proposed Method

Our proposed method follows the pipeline illustrated in Figure 6. For system training, we have used the provided face images in the ChaLearn LAP 2016 apparent age estimation challenge dataset [5]. The building blocks of the system are explained in the following subsections.

4.1. Face Detection

We have employed the Mathias *et al*. [16] face detector to obtain bounding boxes of faces in the images. Some images in the dataset contain more than one face. In this case,

Figure 4. Sample face images from the training set that are labeled as 25 years old. Although the age labels are the same, facial attributes, image resolution, genders and ethnicities of people can be very different.
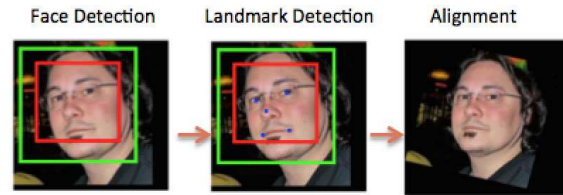

Figure 5. Overview of the face detection and alignment pipeline. Firstly, face is detected (red box) and 30% margin is added (green box). In the second step, five facial landmarks are localized and lastly the faces are aligned with respect to these landmarks.

we have picked the faces with greater detection scores. To handle faces that appear at the corner or at the border of the images, we have added zero padding and created square sized images. Since, in 2% of the images, the Mathias *et al*. face detector did not return any faces, we have utilized one more face detector, namely the one available in the dlib library [12], and run it on these images. In the end, only in 1.5% of the images, no faces have been detected. The entire image has been fed to the CNN model in this case. In case of detection, the bounding boxes of faces have been enlarged by 30% to benefit from more facial attributes for age estimation, i.e. hair and shape.

4.2. Landmark Detection and Alignment

For face alignment, we have first detected 68 facial landmarks using the approach presented in [11], which is available in the dlib [12] library. After localizing 68 points, we have calculated five points (nose tip, eye centers, lip corners) and aligned the faces with respect to these points, as described in [28]. In Figure 5, the red box shows the original bounding box from face detection and the green one represents the enlarged box.

4.3. System Training

In this subsection, we present the processes that have been performed for system training.

4.3.1 Data Augmentation

The competition dataset includes in total 4113 images for training and 1500 images for validation. Each age label is averaged over more than 10 different human voters. In Figure 3, age distribution of training set is shown. It can be clearly seen that the distribution is not even. Moreover, the number of training samples is too low to train a deep CNN architecture. To tackle this problem, we have employed adaptive augmentation on the dataset, that increases the total number of images, while trying to make the age distribution of the set even. Our augmentation method includes random rotation, zooming, and color channel shifting. All images have been resized to 256×256 pixels resolution. We have also used five different cropped 224×224 pixels area of each image and combined them, while training the deep CNN models.

4.3.2 Age - Shifted Grouping

ChaLearn LAP 2016 dataset includes standard deviation of each ground truth age label. Average of standard deviations in the annotations of the dataset is 4.012. This causes a problem, since each face image has multiple age labels and the provided average face is not an exact label. To account for this problem, instead of using average age of the annotated face image as the class label, we have grouped the face images that are within a specified age range. Using these age groups and their age-shifted groupings, we have trained an ensemble of deep learning models.

In contrast to grouping ages without changing boundaries of each group, which was proposed by Zhu *et al*. [31], we have implemented a grouping technique with a different perspective. Our proposed method consists of three different CNNs that group human ages into 34 classes. In addition to that, we have also shifted the age group indices by one for each deep model. In Figure 7, three different age groupings are illustrated.
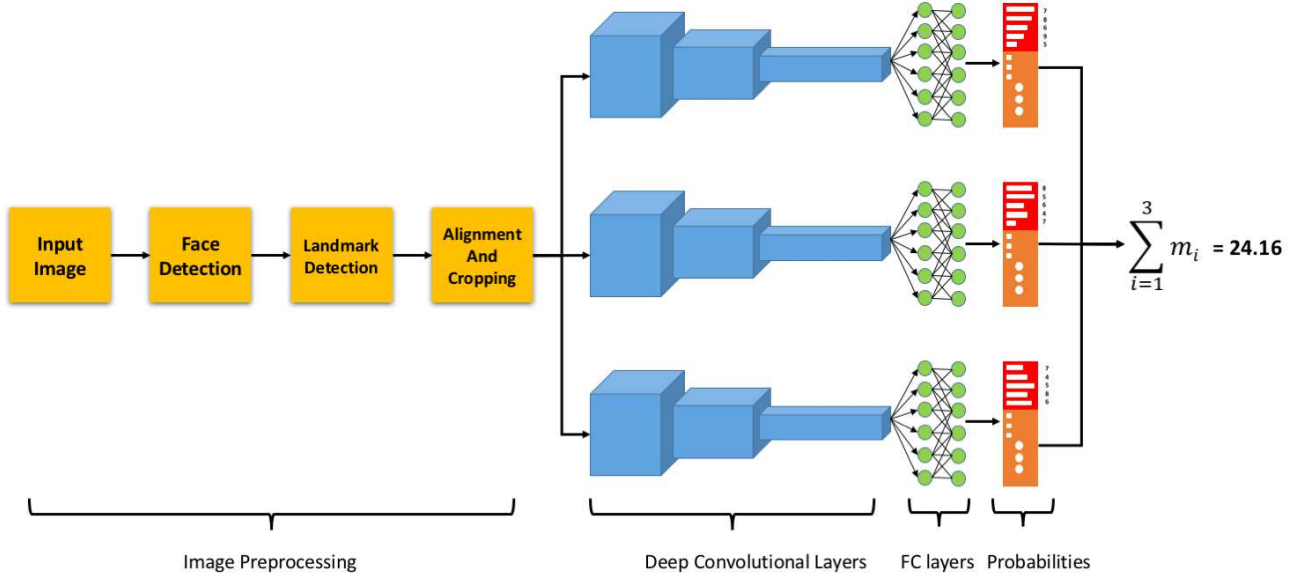
Figure 6. Overview of the proposed system. It consists of three main steps: In image preprocessing step, using Mathias *et al*. face detector, faces are found. Then five landmarks are located and faces are aligned with respect to these landmarks. Finally, aligned faces are cropped with 60% additional margin. After preprocessing, three different VGG-16 model-based trained networks are fed with cropped faces. Using their softmax probabilities that are generated by the last fully connected layers (FC) of deep learning models ($m_i, i = 1, 2, 3$), final score is obtained.
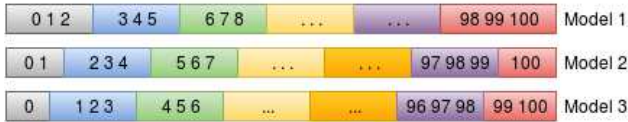

Figure 7. Shifted Age Groups

4.3.3 Deep Learning with CNNs

In this work, VGG-16 deep CNN models pretrained on the IMDB-WIKI dataset from [22] have been used for fine tuning on the challenge dataset. We have changed the last neural network layer of our networks to output 34 classes corresponding to different age groups. To prevent overfitting, we have changed the learning rate policy of models. They have been fine tuned for each layer separately. Learning rates of feature extraction layers are lowered, while fully connected layers' learning rates remain high. All models are trained with softmax loss function.

4.3.4 Determination of Score Value

To perform transition from age classification to age regression, i.e. to provide an age estimation output based on the CNN model classification outputs, classification results have been weighted summed with multiplication of their age indexes and probabilities. Unlike to the DEX

method [22], we have sorted probability vector output of the CNN models and selected predefined top k elements and calculated weighted sum of them. This variable has also been tuned and we have selected the one that has produced the best score. At the last stage of age estimation, all three scores from three different models have summed up to produce the final output.

5. Experiments

Experiments have been conducted on the ChaLearn LAP 2016 validation set, which consists of 1500 images in total. As the error measure, we have used the same ϵ-error function that was defined for the ChaLearn LAP challenge, which is given in Equation 1. Since LAP dataset has been annotated by multiple people, ground truth value has been calculated as the average of votes, μ, with standard deviation, σ. In error function x represents the predicted age and the value of the error function can be between 1 (worst) and 0 (best). In the experiments, face detection and alignment procedures described in Sections 4.1 and 4.2 have been applied.

$$\epsilon = 1 - e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

In the experiments, firstly, we have extracted deep features from FC-7 layer of VGG-FACE [19] model and

13

trained a 3-layer neural network that has two hidden layer using deep features. We have approached the problem as a multi-label classification problem and we have assigned age labels that are also in the boundaries of standard deviation, as true. This model performance has been found to be low with an error of 0.41, without aligning any detected faces. Then, we applied alignment on this setup and the alignment of the face images decreased our error rate to 0.39. Significant improvement is obtained after we used fine tuned VGG-16 [24] models for testing our age-shifted grouping strategy. We have produced three different scores for each CNN, m_i, using sorted softmax probability vector, p_j, and corresponding ages, ω_j. In the first experiment, we have used all of the probabilities and for the second one we have used only maximum probability for calculating the final score, which are formulated in Equations 2 and 3, respectively.

$$m_i = \sum_{j=1}^{34} p_j \omega_j \qquad (2)$$

$$m_i = p_1 \omega_1 \qquad (3)$$

Lastly, we have used the five maximum probabilities to produce the final score of the CNN, as formulated in Equation 4, which has attained the best accuracy in the validation set.

$$m_i = \sum_{j=1}^{5} p_j \omega_j \qquad (4)$$

Finally, we have combined these m_i values that represent the outputs of CNNs to estimate the final score of the system using Equation 5. Since all m_i values are in range of 0-34, they are summed up to produce the final prediction. All experiments are implemented with Caffe Deep Learning Framework [10] on NVIDIA TITAN X GPU. Training of each deep model took approximately 5 hours. The experimental results of these different cases are listed in Table 2.

$$prediction = \sum_{i=1}^{3} m_i \qquad (5)$$

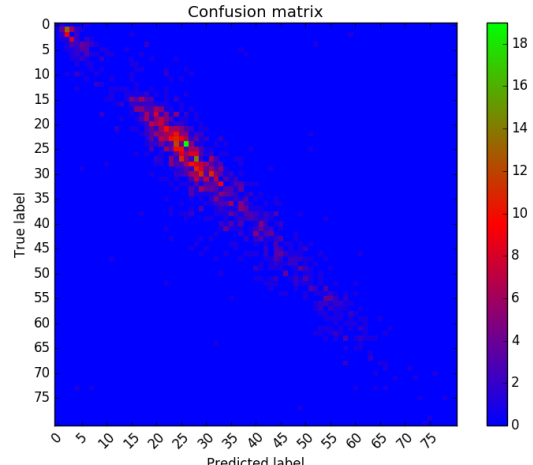| Method	Error
VGG-FACE + w/o Alignment + Neural Network	0.4128
VGG-FACE + Alignment + Neural Network	0.3906
3 CNNs + Alignment + full weighted sum.	0.2907
3 CNNs + Alignment + max. 1 weighted sum.	0.3000
3 CNNs + Alignment + max. 5 weighted sum.	0.2897

Table 2. Experimental Results

Figure 8. Confusion matrix obtained on the validation set (Best displayed in color).

Rank	Team	Test Error
1	OrangeLabs	0.2411
2	palm_seu	0.3214
3	cmp+ETH	0.3361
4	WYU_CVL	0.3405
5	**ITU_SiMiT (Ours)**	0.3668
6	Bogazici	0.3740
7	MIPAL_SNU	0.4569
8	DeepAge	0.4573

Table 3. ChaLearn LAP 2016 final ranking on the test set. Out of 84 registered participants, there were only eight submissions.

6. Conclusions

In this study, we have built an ensemble of deep learning models for apparent age estimation. Together with age groupings, this ensemble model accounts for imprecise, uncertain apparent age labels of the face images. In the proposed system, we have employed five-point face alignment. To equalize age distribution in the dataset, we have performed an adaptive data augmentation. We have utilized different fusion schemes and with these enhancements, we have attained an error rate on the validation set, which is on par with the last year's best performing system. However, the error obtained on the test set, 0.3668, is relatively higher than the one obtained on the validation set. As a future work, we will investigate the reasons for this performance gap. Furthermore, we plan to develop a training scheme that takes into account the standard deviations of the labels.

Acknowledgements

References

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Computer Vision–ECCV 2014*, pages 768–783. Springer, 2014.

[3] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Information Forensics and Security, IEEE Transactions on*, 9(12):2170–2179, 2014.

[4] 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 *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[5] S. Escalera, M. Torres, B. Martnez, X. Bar, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri, and M. Valstar. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *CVPR workshops*, 2016.

[6] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. 2008.

[7] X. Geng, Z. H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging pattern. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[8] H. He and E. A. Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[10] 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.

[11] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[12] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] Z. Kuang, C. Huang, and W. Zhang. Deeply learned rich coding for cross-dataset facial age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 96–101, 2015.

[15] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[16] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.

[17] M. Minear and D. C. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633, 2004.

[18] B. Ni, Z. Song, and S. Yan. Web image and video mining towards universal and robust age estimator. *Multimedia, IEEE Transactions on*, 13(6):1217–1229, 2011.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[20] R. Ranjan, S. Zhou, J. Chen, A. Kumar, A. Alavi, V. Patel, and R. Chellappa. Unconstrained age estimation with deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 109–117, 2015.

[21] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.

[22] R. Rothe, R. Timofte, and L. V. Gool. DEX: Deep EXpectation of apparent age from a single image. In *ICCV, ChaLearn Looking at People workshop*, December 2015.

[23] 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. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[25] A. Smola and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[26] J. Suo, T. Wu, S. Zhu, and S. Shan. Design sparse features for age estimation using hierarchical face model. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[28] X. Wu, R. He, and Z. Sun. A lightened CNN for deep face representation. *arXiv preprint arXiv:1511.02683*, 2015.

[29] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[30] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015.

[31] Y. Zhu, Y. Li, G. Mu, and G. Guo. A study on apparent age estimation. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

Close Estimates

Input						
Aligned						
Apparent	47.07	19.67	5.29	34.11	23.3	37.17
Predicted	47	20	5	34	23	37

Erroneous Estimates

Input						
Aligned						
Apparent	40.95	0.94	54.78	18.42	38.4	10.89
Predicted	35	6	60	24	44	4

Figure 9. Close and erroneous estimates from sample images.