Deep Label Distribution Learning for Apparent Age Estimation

Xu Yang[1], Bin-Bin Gao[2], Chao Xing[1], Zeng-Wei Huo[1], Xiu-Shen Wei[2], Ying Zhou[1], Jianxin Wu*[2], and Xin Geng†[1]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
Key Laboratory of Computer Network and Information Integration, Ministry of Education, China
{x.yang,xingchao,huozw,zhouying1,xgeng}@seu.edu.cn
[2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{weixs,gaobb,wujx}@lamda.nju.edu.cn

Abstract

In the age estimation competition organized by ChaLearn, apparent ages of images are provided. Uncertainty of each apparent age is induced because each image is labeled by multiple individuals. Such uncertainty makes this age estimation task different from common chronological age estimation tasks. In this paper, we propose a method using deep CNN (Convolutional Neural Network) with distribution-based loss functions. Using distributions as the training tasks can exploit the uncertainty induced by manual labeling to learn a better model than using ages as the target. To the best of our knowledge, this is one of the first attempts to use the distribution as the target of deep learning. In our method, two kinds of deep CNN models are built with different architectures. After pre-training each deep CNN model with different datasets as one corresponding stream, the competition dataset is then used to fine-tune both deep CNN models. Moreover, we fuse the results of two streams as the final predicted ages. In the final testing dataset provided by competition, the age estimation performance of our method is 0.3057, which is significantly better than the human-level performance (0.34) provided by the competition organizers.

1. Introduction

In recent years, facial age estimation has attracted more and more researchers' attentions because of its abundant real-world applications, *e.g.*, security control [8], human-computer interaction [10] or personal identification [11].

In the past few years, a lot of methods have been proposed to solve the age estimation problem. To name a few, Lanitis *et al*. [11, 10] used a quadratic function called aging function to solve chronological age estimation. Geng *et al*. [8, 7] proposed the AGES algorithm based on the subspace trained on a data structure called aging pattern vector. Fu *et al*. [4, 3] used multiple linear regression on the discriminative aging manifold to solve the age estimation problem. In addition, SVR [9], GMM (Gaussian Mixture Model) [18] and patch-based HMM (Hidden Markov Model) [20] were used to predict ages from facial images.

However, in the competition organized by ChaLearn [2], the age of each image in the dataset is labeled by multiple individuals rather than its chronological age. For each image, its mean age m and the corresponding standard deviation σ are given. The uncertainty induced by standard deviation makes this age estimation different from the common chronological one. Thus, the aforementioned methods can hardly use these uncertainty information directly because they use ages as the targets instead of age distributions. In this apparent age estimation task, thanks to the induced uncertainty (*i.e.*, standard deviation), it is natural to treat age distributions as the ground truth. In this paper, we propose a deep learning model with distribution-based loss function, which is more suitable in this estimation task. This proposed method can also be treated as one kind of LDL (Label Distribution Learning) [6].

We choose the deep CNN model as our basic model because of its satisfactory performance in various areas of face-related tasks, *e.g.*, face recognition [16], face alignment [17], face verification [19, 5], age and gender classification [12], etc. In this paper, we build two deep CNNs of different architectures as two streams to solve the apparent age estimation problem. The first one is based on a pop-

*This work was supported by the National Natural Science Foundation of China under Grant 61422203 and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

†J. Wu and X. Geng are the corresponding authors.

Figure 2: Three steps of the images pre-processing in our method. First, we employ the DPM model [13] to detect the main facial region of each image. Then the detected face is fed into a public available facial point detector software [17] to get the corresponding five facial key points including the left/right eye centers, nose tip and left/right mouth corners. Finally, based on these facial points, we utilize face alignment for these facial images.

ular pre-trained deep network, *i.e.*, VGG-16 [15]. For this deep CNN model, it is fine-tuned three times with different datasets in our method. The second deep CNN model is based on a novel architecture, we utilize different types of inputs, *i.e.*, different augmentation methods used on our own collected data, to train this deep CNN model. After that, we use the competition dataset to fine-tune this model. Finally, we fuse the results of two streams as the final predicted ages. In the final evaluation phase, our method achieved 0.3057 age estimation performance, which is significantly better than the human-level performance (0.34).

The contributions of this paper are listed as follows:

1 By using distribution-based loss functions as training targets in deep CNN models, the uncertainty information induced by standard deviation is exploited to solve the apparent age estimation problem. In addition, by using distribution as the target can make a face image contribute to not only the learning of its own age, but also the learning of its neighboring ages. Therefore, we can use the data more sufficiently.

2 We totally downloaded 119,539 face images from Internet and labeled these images manually. By using these additionally collected face images, the deep CNNs used in our method have a stronger prediction ability compared with the model only trained on the images provided by this competition.

This paper is organized as follows. In Section 2, we will introduce the proposed Deep Label Distribution Learning method. Implementation details and experimental results will be presented in Section 3. Finally, we conclude our method and present the future issues in Section 4.

2. Deep Label Distribution Learning for apparent age estimation

The whole process of our Deep Label Distribution Learning method is shown in Figure 1. There are two different streams to get their own predicted ages, and the predicted ages from two streams will be fused as the final results. The details of two streams will be discussed in Section 2.2 and 2.3, respectively.

2.1. Facial images pre-processing

Figure 2 shows three steps of the pre-processing. We use the Coc-DPM algorithm [13] for face detection. Then the detected faces are fed into a public available facial point detector software [17] to detect five facial key points, including the left/right eye centers, nose tip and left/right mouth corners. At last, we align all the faces based on the detected five points.

2.2. The first stream

In the first stream of our method, we use the popular pre-trained CNN model, *i.e.*, VGG-16 [15], as the basic model for age estimation. Figure 3 shows the architecture of this deep CNN model. The whole process of training and predicting is listed as follows:

1) Fine-tuning on the MORPH dataset [14].

In this step, we fine-tune VGG-16 on the pre-processed MORPH [14] dataset which contains 55,134 images.

2) Fine-tuning again on two datasets collected from search engines and public facial datasets.

In this step, we fine-tune the deep model obtained in the first step on two different datasets separately, then two different deep CNN models are obtained. The first dataset includes 27,197 images downloaded from *Google*. The second dataset includes 37,606 images downloaded from *Baidu*, *Bing*, FG-Net [1] and Adience [12].

3) The last fine-tuning on the competition dataset with two different loss functions.

In this step, we use 3,615 competition images (training and validation datasets) to fine-tune two deep CNN models obtained in the second step. Two different loss functions are used in these two deep CNN models separately. The first one is the KL divergence loss function, which is a distribution-based loss function. By using the KL divergence loss function, instead of considering each face image as an example with one label/age, each face image is treated as an example associated with a label distribution to exploit the uncertainty information. Given the mean age m_n and standard deviation σ_n of the n^{th} face from the competition dataset, a Gaussian distribution with mean m_n and standard deviation σ_n is generated for the n^{th} face. The second loss function is the softmax loss function.

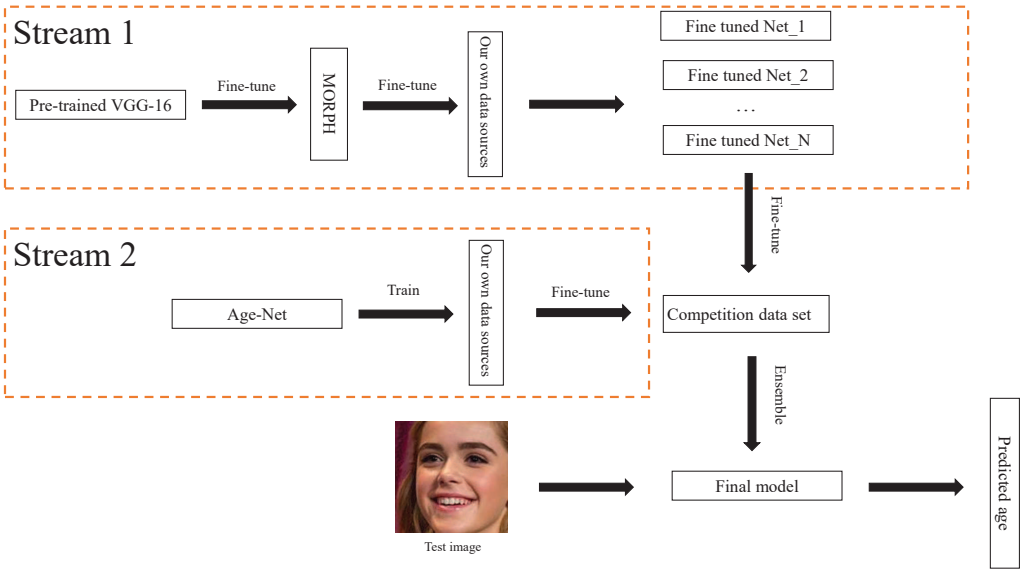


Figure 1: The framework of the proposed Deep Label Distribution Learning method. In our method, two different deep CNN models are built as two streams. The first one is fine-tuned three times on three different datasets, *i.e.*, MORPH [14], some other collected datasets and the competition dataset. The second one is trained by different types of our own collected data and then fine-tuned on the competition dataset. For the test images, two predicted ages from two streams are fused as the final predicted age.

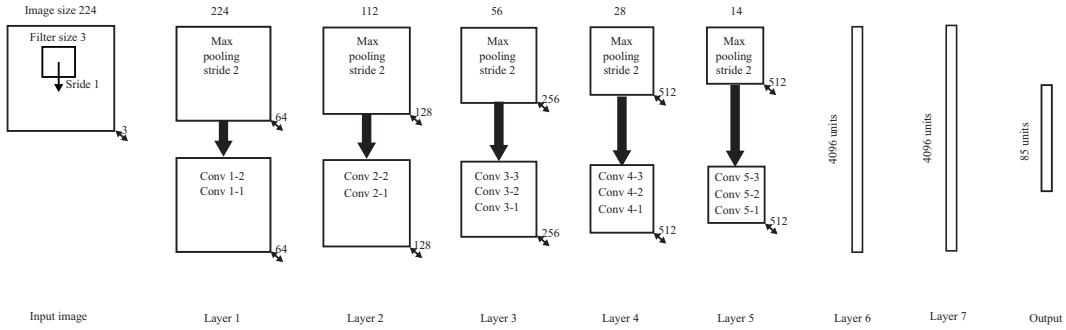


Figure 3: Architecture of the deep CNN model used in the first stream. In this deep CNN, a 224×224 alignment face image is presented as input. The convolutional filter is 3×3 with stride and pad 1, followed by a parameter ReLU (not shown) layer and a 2×2 max-pooling layers with stride 2. This deep CNN has 5 convolutional layers with the same parameters. Last 2 layers are fully connected layers, and a ReLU (not shown) layer follows each fully connected layers. The final layer is an 85-dimension output layer with the KL divergence loss function.

4) Ensemble different results of four different deep CNN models.

Because different datasets (in the second step) and different loss functions (in the third step) are used in this stream, totally four deep CNN models are obtained. Therefore, different deep CNN models will capture different feature representations from facial images. Thus four deep CNN can provide different useful information for predicting ages. We concatenate the features extracted from these four deep models and use a distance-based voting ensemble method to predict age from this concatenated feature. As shown in Figure 3, the dimension of the output layer is 85, by concatenating, we can get a 340 dimension feature which is denoted as $\boldsymbol{x}$. For the n^{th} face image with its age which is denoted as t_n in the training set, we can get a corresponding concatenated feature $\boldsymbol{x}_n$. When a new face image comes, we first compute its concatenated feature $\boldsymbol{x}^*$ from four deep CNNs and then use the following approach to predict its age

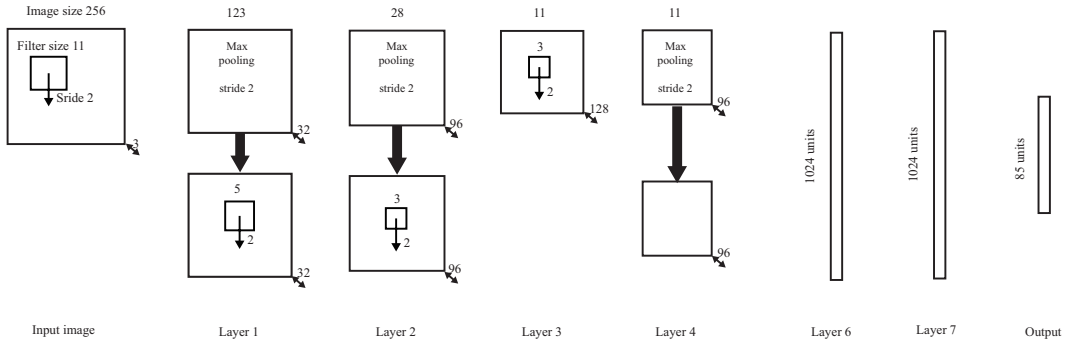


Figure 4: Architecture of the deep CNN model used in the second stream. In this deep CNN, a 256×256 alignment facial image is presented as input. The convolutional filter is $11 \times 11 \times 32$ with stride 2, followed by a 3×3 max-pooling layers with stride 2. In addition, batch normalization is applied to the resulting feature maps. Then, they are passed through a parameter ReLU (not shown) and pooled. Similar operations are repeated in layers 2, 3 and 4. The last two layers of this model are fully connected layers. The final layers is an 85-dimension output layer with the KL divergence loss function.

t^*:

$$t^* = \sum_{n=1}^{N} t_n K(\boldsymbol{x}^*, \boldsymbol{x}_n)\,. \tag{1}$$

The distance measure $K(\cdot,\cdot)$ is defined as follows:

$$K(\boldsymbol{x}^*, \boldsymbol{x}_n) = \exp(-\alpha||\boldsymbol{x}^* - \boldsymbol{x}_n||_2^2)\,, \tag{2}$$

where $\boldsymbol{x}_n^{th}$ is the concatenated feature of the n^{th} image in the training dataset, t_n is the age of the n^{th} image and N is the number of images in the training dataset.

2.3. The second stream

For the second stream, we propose a new CNN architecture which is shown in Figure 4.

In this stream, we first use the collected facial images to pre-train this deep CNN model. After that, we fine-tune the model by utilizing 2,479 training and 1,136 validation images from the competition, and then obtain the age predictions of the final testing images. In order to boost the estimation performance of this deep CNN model, we combine the power of multiple networks trained on different types of inputs: 1) The aligned face images in the RGB colorspace; 2) The gray-level images with images gradient magnitudes and orientations; 3) The single channel gray-level images; 4) The HSV colorspace images; 5) The three channels images obtained by the sobel filter and 6) The sobel-level images with the ones in the RGB colorspace. By using such different types of input, this deep CNN model can learn different aspects of useful information about ages. During predicting, an alignment face image is operated by a random horizontal scale variants and flipped with 50 times. Thus 300 results are obtained from the above six types deep CNNs. At last, the final age estimations can be obtained by averaging these 300 predicted results.

2.4. Fusing in the age layer

Finally, we fuse two streams by a simple strategy. The strategy is that if the predictions of the two streams are within 11 years difference, we average their predictions as the finally predicted results; if not, then we take the predictions of the first stream as the final results.

3. Experiments

In this section, we will first describe the details about the datasets used in our method, *i.e.*, our own collected datasets and the competition dataset. And then, we present the experimental results of the proposed method on the competition dataset.

3.1. Datesets

3.1.1 Competition dataset

In the competition [2], totally 3,615 face images (2,479 in the training dataset and 1,136 in the validation dataset) with their apparent ages and standard deviations are provided. nine sampled facial images of the competition dataset are shown in Figure 6.

3.1.2 Collected data

Training an effective deep CNN model heavily relies on a large number of training images. However, the organizer only provides 3,615 training images. In this case, we collect 119,539 additional images from different sources, *e.g.*, Internet and public facial datasets. Table 1 shows the details of the additional datasets, including the names of their resources, numbers and usages. The corresponding age distributions of these data sets are shown in Figure 5.

For the public facial datasets, we download the database

Table 1: The details of the datasets used in our work.

No.	Source	Number	The first stream	The second stream
1	MORPH	54,736	The 1^{st} Fine-tuning	–
2	Google image search	27,197	The 2^{nd} Fine-tuning	Pre-training
3	Bing image search	11,220	The 2^{nd} Fine-tuning	–
4	Baidu image search	6,212	The 2^{nd} Fine-tuning	–
5	FG-NET	959	The 2^{nd} Fine-tuning	Pre-training
6	Adience	19,215	The 2^{nd} Fine-tuning	–
7	Competition	3,615	The 3^{rd} Fine-tuning	Fine-tuning
	Total	123,154	–	–

Table 2: Performance of our method in the validation dataset. The 0.3377 performance is obtained by fusing the results of the first stream and the second stream. The fusing strategy is described in Section 2.4.

Model	The 1^{st} stream	2^{nd} stream	After fusing
Performance	0.3534	0.3610	0.3377

which can provide the face images and their related age labels such as MORPH [14], FG-NET [1] and Adience [12].

From the Internet, we collect 44,629 face images by using popular image search engines, *e.g.*, *Google*, *Bing* and *Baidu*. We search these face images by using age-related tags, *e.g.*, "80 years old" tag or the "baby" tag. Then we label these face images manually by crowdsourcing.

3.2. Results

In Figure 6, it shows nine examples of face images provided by the competition, with the corresponding generated age distributions and the predicted distributions from our method. The generated age distribution here is the Gaussian distribution with the provided mean and standard deviation, and this distribution can be treated as a kind of "ground truth" distribution.

The performance of age estimation is evaluated by the formula provided by the competition, which is:

$$\epsilon = 1 - \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right) \tag{3}$$

where t is the predicted age, m is the mean apparent age and σ is the standard deviation. Table 2 and Table 3 show the age estimation performance of our method in the validation dataset and testing dataset, respectively. The comparison results of the top 5 teams are also shown in Table 3.

4. Conclusion

In this paper, we propose a method to solve the apparent age estimation problem. Two deep CNN models in different architectures are separately trained as two corresponding streams on different datasets. Moreover, the distribution-based loss function, *i.e.*, the KL divergence loss function, is used in our model as the training target to exploit the information of the standard deviations. In order to improve the estimation performance of both two streams, we also utilize the additional 119,539 face images collected from Internet and public facial datasets. Finally, out method achieved 0.3057 performance and ranked the 4^{th} place in Age Estimation at the ChaLearn LAP challenge [2] in association with ICCV 2015. In the future, we will try different types of loss functions and incorporating more information from label distributions.

Table 3: Comparison results of the top 5 teams in the final evaluation phase of this competition. In addition, the human-level estimation performance of this task is 0.34 provided by the competition organizers [2].

Rank	Team	Performance
1	CVL-ETHZ	0.264975
2	ICT-VIPL	0.270685
3	WVU-CVL	0.294835
4	**SEU-NJU (Ours)**	0.305763
5	UMD	0.373352

References

[1] The FG-NET Aging database.

[2] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzàlez, H. J. Escalante, and I. Guyon. ChaLearn 2015 apparent age and cultural event recognition: Datasets and results. In *ICCV, ChaLearn Looking at People workshop*, 2015.

[3] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. In *ICME*, pages 578–584, 2008.

[4] Y. Fu, Y. Xu, and T. S. Huang. Estimating human age by manifold analysis of face pictures and regression on aging features. In *ICME*, pages 1383–1386, 2007.

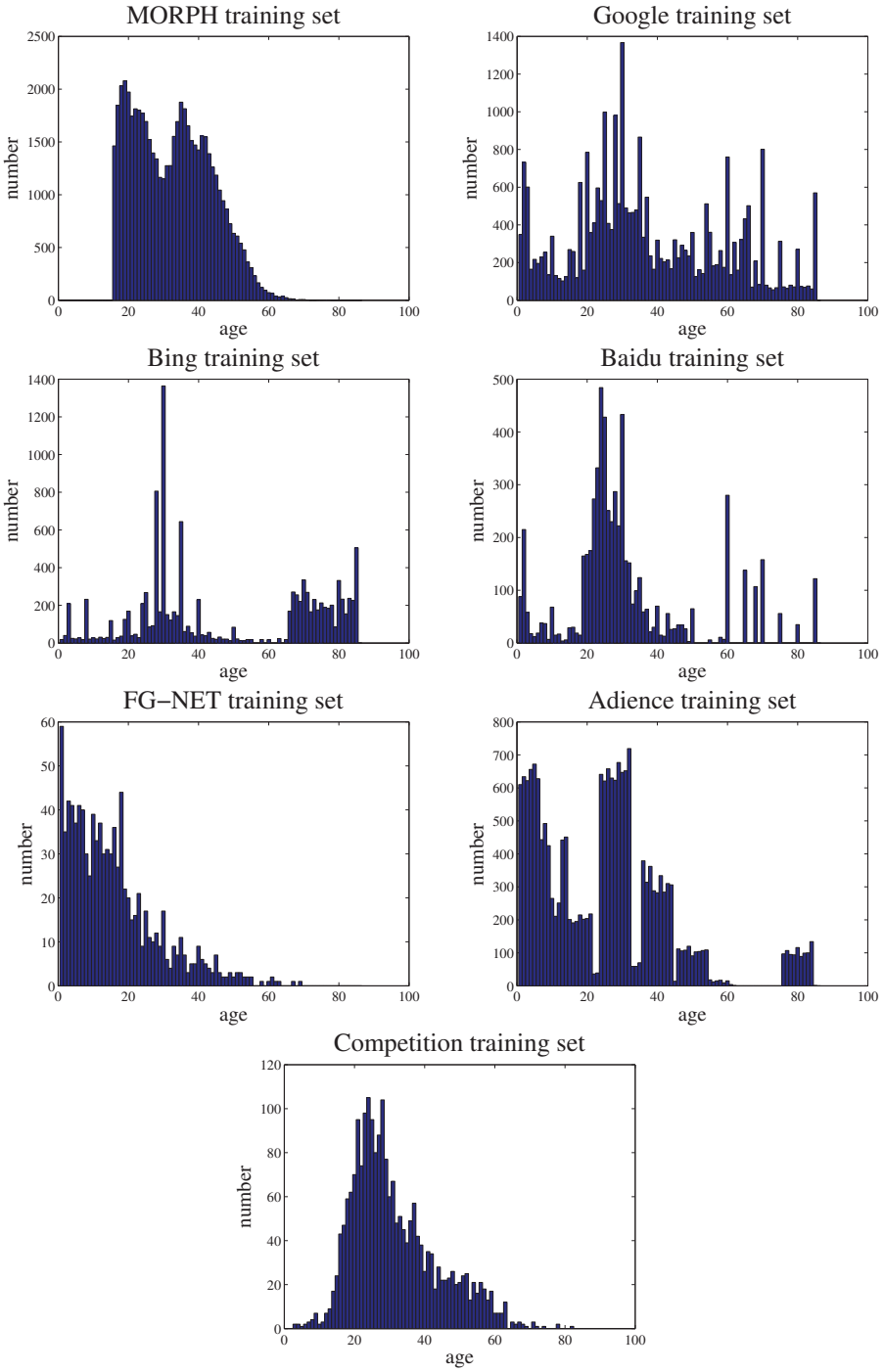
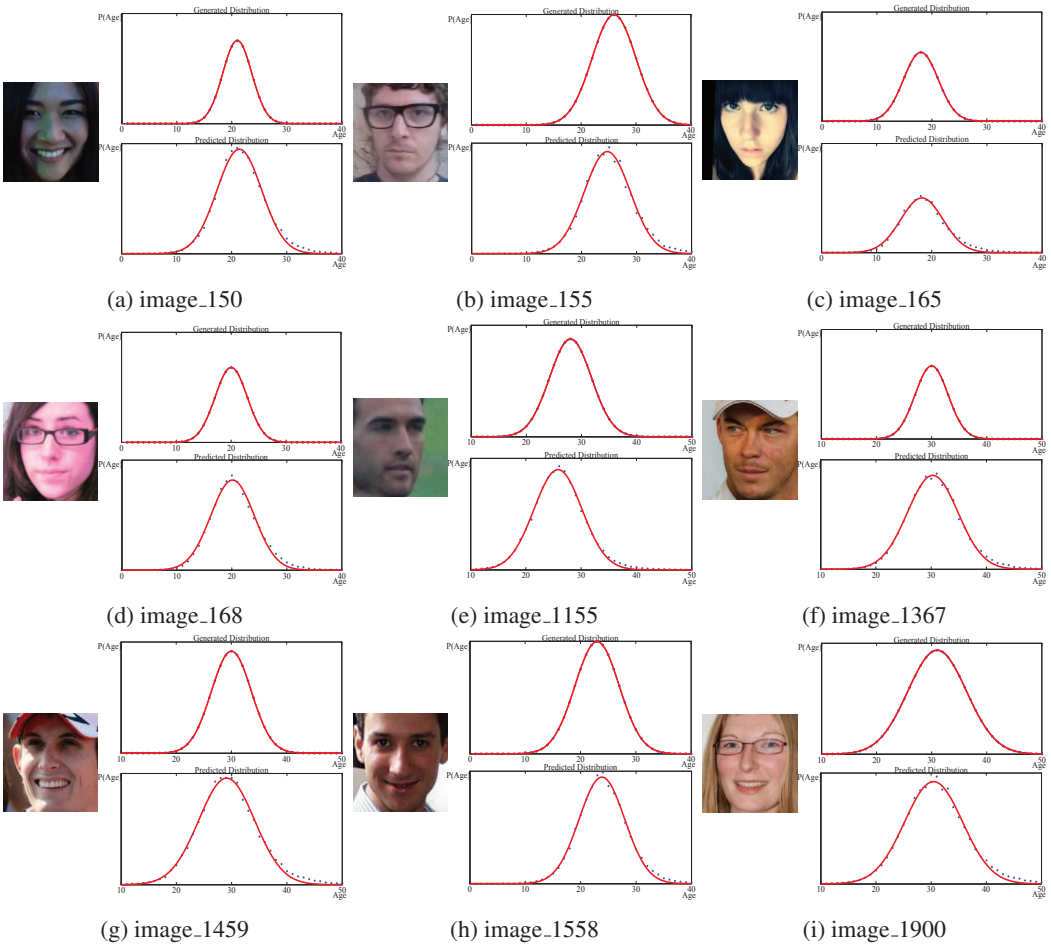


Figure 5: Age distributions of each data resource, the horizontal axis of each histogram represents age and the vertical axis of each histogram means the number of age in the corresponding data resource. The name of resource is shown at the top of each figure.

[5] E. G. L.-M. Gary B. Huang, Honglak Lee. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012.

[6] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *TPAMI*, 35(10):2401–2412, 2013.

[7] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *TPAMI*, 29(12):2234–2240, 2007.



(a) image_150 (b) image_155 (c) image_165

(d) image_168 (e) image_1155 (f) image_1367

(g) image_1459 (h) image_1558 (i) image_1900

Figure 6: Sampled facial images from the competition dataset. In each subfigure, the left is the original face image, the top right is the corresponding generated age distribution with provided mean and standard deviation, and the bottom right is the predicted distribution of the first stream of our method. This figures are best viewed in color.

[8] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *ACM MM*, pages 307–316, 2006.

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

[10] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(1):621–628, 2004.

[11] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *TPAMI*, 24(4):442–455, 2002.

[12] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. *CVPR*, 2015.

[13] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, pages 720–735, 2014.

[14] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FGR*, pages 341–345, 2006.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[16] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014.

[17] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483, 2013.

[18] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang. Regression from patch-kernel. In *CVPR*, pages 1–8, 2008.

[19] M. A. R. Yaniv Taigman, Ming Yang. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014.

[20] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. Huang. Face age estimation using patch-based hidden markov model supervectors. In *ICPR*, pages 1–4, 2008.