Age-Related Factor Guided Joint Task Modeling Convolutional Neural Network for Cross-Age Face Recognition

Haoxi Li, *Student Member, IEEE*, Haifeng Hu, *Member, IEEE*, and Chitung Yip

***Abstract*— Cross-age face recognition has remained a popular research topic as most regular facial recognition systems have failed in dealing with facial changes through age. In order to enhance the system's capability of discriminating facial identity features in spite of age changes, this paper proposes a novel deep convolutional network method for cross-age face recognition called age-related factor guided joint task modeling convolutional neural networks, which combines an identity discrimination network with an age discrimination network that shares the same feature layers. By alternatively training the fusion networks and the combined factor model, the cross-age identity features and cross-identity age features can be effectively separated with high inter-class distension and intra-class compactness. Extensive experiments have been performed on the benchmark aging data sets, including MORPH, CACD-VS, and Cross Age LFW. The results have demonstrated the superiority and effectiveness of our model.**

***Index Terms*— Age-related factor guided joint task modeling convolutional neural network, cross-age face recognition, joint task factor analysis.**

I. Introduction

FACE recognition has been a popular topic in pattern recognition with extended tasks under different variations such as pose [1]–[4], illumination [5], [6], expression [7], [8] and age [9], [10], etc. Among them, cross-age face recognition has attracted research interests recently and yet still lacks sufficiently reliable solutions [11]. It remains a challenge due to the sophisticated aging process which poses different nonlinear effect on different individuals. Different methods including generative and discriminative approaches have been proposed to tackle the issue [12].

Recently, deep convolution neural networks (CNN) have been successfully applied to classification problem and achieved the-state-of-art recognition performance. These methods, however, fail to achieve a more satisfactory result due to the inherent defects of the traditional approaches, or that they have ignored the latent correlations between the aging process and the identity features.

In order to find a more discriminative identity features from the face images, we propose a novel deep convolutional network method called age-related factor guided joint task CNN (AFJT-CNN) for cross-age face recognition. The proposed network combines a target network, with an auxiliary network that shared the same feature layers.The purpose of the target network is to obtain a discriminative feature that is only related with identity, while the auxiliary network separate the age-related features from the identity feature in the target network. The two subnetworks are trained to be adversarial so that the auxiliary subnetwork can improve the performance of the identity recognition networks. Then the center loss is used in the training scheme to make the deep features more discriminative. In this way, our model is more adaptive to cross age face recognition problem, as supported by our experimental results in Section 4.

The major contributions of this paper are summarized as follows:

- we propose a new deep learning network architecture called AFJT-CNN to specifically address the cross age face recognition task. By adversarial learning the parameters in two CNNs for face and age, the age-invariant deep face features can be extracted, which are more robust to the variations caused by the aging process.
- we suggest that non-linear age features that cannot be entirely separated from identity feature is also of great significance to the performance of the scheme. Based on such assumption, we propose a novel joint task factor analysis model (JTFA) for obtaining the parameters needed for the identity discrimination network and an auxiliary subnetwork for extracting a feature related with both age and identity.
- Extensive experiments have shown that the proposed approach outperforms the state-of-the-art on three large face aging datasets (MORPH Album2 [13], CACD-VS [14] and Cross-Age LFW [15]).

II. Related Works

Early cross-age face recognition methods [16]–[19] usually generate faces of different ages through certain simulation

Manuscript received June 2, 2017; revised November 25, 2017, February 9, 2018, and March 6, 2018; accepted March 21, 2018. Date of publication March 23, 2018; date of current version May 1, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61673402, Grant 61273270, and Grant 60802069, in part by the Natural Science Foundation of Guangdong under Grant 2017A030311029, Grant 2016B010109002, Grant 2015B090912001, Grant 2016B010123005, and Grant 2017B090909005, in part by the Science and Technology Program of Guangzhou under Grant 201704020180 and Grant 201604020024, and in part by the Fundamental Research Funds for the Central Universities of China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christoph Busch. *(Corresponding author: Haifeng Hu.)*

The authors are with the School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: huhaif@mail.sysu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2018.2819124

schemes for recognition. However, they are computationally expensive and unstable due to the non-linear facial appearance change across age. Subsequently, many traditional discriminative methods [12], [14], [20]–[24] are proposed to tackle recognition tasks without explicit age modeling. They usually focus on adopting robust age-invariant image descriptors and classifiers with superior discriminative power. Reference [23] proposes an alternating greedy coordinate descent (AGCD) boosting method to alternatively train an identity classifier and an age classifier. The two classifiers use a shared patch-based feature pool, such that the auxiliary age classifier can help separate the age-sensitive features from the identity classifier. CARC [14] proposes a reference encoding method to extract a feature representation that the encoding of images of the same person in adjacent years are similar. Reference [12] proposes an HFA method that extracts the patch-based histogram of oriented gradient features from face images and decomposes it into face component and age component using factor analysis model. Although these patch-based discriminative methods have sought to obtain an age-invariant representation of face feature, they do not explicitly remove the undesired age-sensitive information using an end-to-end training scheme. Relying on manually extracted features without age priori, these methods have unstable performances in large age gap recognition tasks.

Due to the great potential of CNN approaches in classification problem, there is an increasing number of works that attempt to combine CNN with age-invariant face recognition tasks. For instance, DeepFace [25] reports that a deeply-learned face representation achieves the accuracy close to human-level performance on LFW dataset. Reference [26] learns a patch-based deep CNN with the identification-verification supervisory signal and further adds supervision to early convolutional layers, greatly boosting the face recognition accuracy. FaceNet [27] achieves 99.63% verification accuracy on LFW with a deep CNN trained to directly optimize the embedding itself, rather than an intermediate bottleneck layer as in previous deep learning approaches. More recently, [28] achieves a new record in verification accuracy: 99.77% on LFW with a two-stage approach that combines a multi-patch deep CNN and deep metric learning. However, these general recognition networks do not consider the aging effect of face images. They are not robust on cross age face recognition tasks, due to the limited cross-age datasets for training. Reference [29] devises an aging-aware denoising auto-encoder model(a^2-DAE) to complete the traditional tasks of facial synthesis and recognition. a^2-DAE generates the images of four different age groups of large gaps of the same person, and feeds them into 4 parallel CNNs. However, Generative methods like [29] are often unstable in real world scenarios, and are computationally costly. Reference [30] proposes a latent factor guided convolutional neural networks (LFCNN) method that separates the identity-related component from the age-related component by alternately training a deep framework on the web-scale face training set and a factor analysis model on a smaller cross-age face training set. LFCNN decomposes facial features into identity features and aging features. However, it has assumed that age and identity features are combined linearly, and can be decomposed completely, which ignores the probability of a more complicated correlation existing between the identity and the aging.

According to our common sense, the aging process varies between different person, and therefore the age features and the identity features might be combined in a more complicated way [31]. Based on such a new understanding of the correlations between aging features and identity features, and utilizing the leaning ability of CNN, we devise a new scheme to extract facial identity features robust to aging. Different from LFCNN [30] which uses the web-scale datasets to train the convolution unit and the aging dataset to train a latent factor, we build a multiloss network that is directly trained on the aging face datasets. In this way we include the age information in the training process to improve the robustness on aging. Our method is also different with a^2-DAE [29] that tries to reconstruct the face images in different large age groups for face verification. Without the aging simulation, our method will focus on finding a discriminative subspace that is robust on aging, which can significantly improve the face recognition performance.

III. The Proposed Approach

A. Problem Modeling

Verifying whether two faces of different ages belong to the same person can be a challenging task, as facial appearances can change dramatically through time. Some works in the literatures [12], [30], [32] have suggested that an original set of features extracted from a face can be represented by a linear combination of average face features, identity features and age features. This hypothesis, however, is based on the presumption that all age features can be separated from identity features in a linear fashion. And therefore, all the possible person-specific aging features, i.e. age features that are not independent of identity will be ignored according to this model. This is a clear contradiction to the common sense, as the aging process apparently varies between different persons [31]. Furthermore, person-specific aging information will be mixed with the extracted identity features, increasing intra-class distances among persons, hence depressing the efficiency of the entire system.

We have developed our method to rectify this negligence, by proposing a new type of auxiliary feature containing person-specific aging information to help extracting a more competent identity feature for verification. In the following we first introduce the structure of our deep network in subsection B. In subsection C we propose a method based on factor analysis to obtain the crucial parameters for the network. In subsection D we introduce the training process of the network.

B. The Structure of the Joint Task Modeling Network

In this section, we introduce the proposed deep learning network, i.e., AFJT-CNN, which is composed of two major parts: a convolutional unit for feature extraction, and two parallel discriminative networks for identity and auxiliary feature learning.

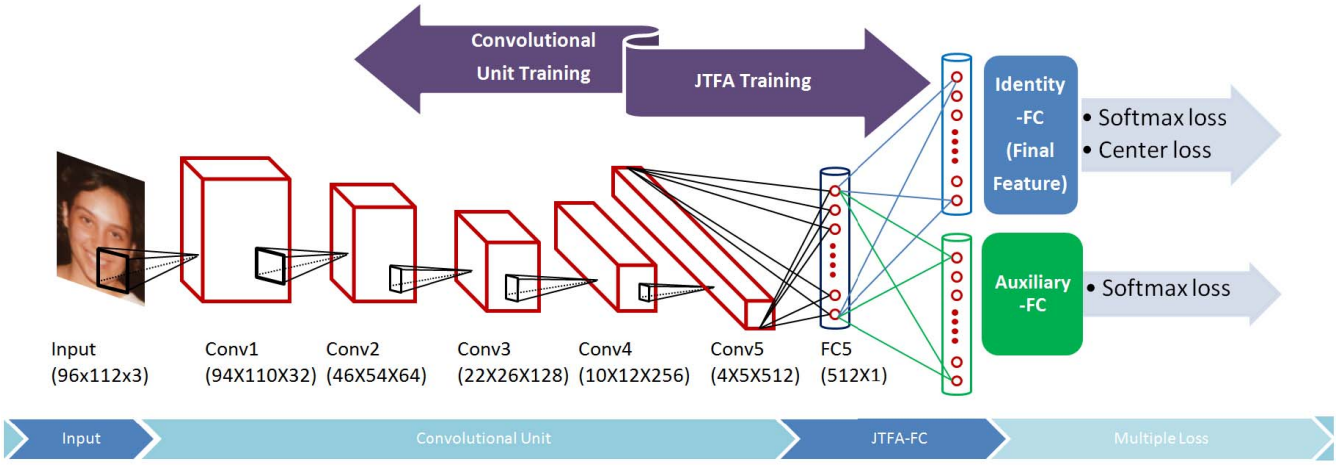


Fig. 1. The architecture of AFJT-CNN. We iteratively train the convolutional unit by multiple loss and train the identity-FC layer and auxiliary-FC layer by JTFA. The identity-FC layer features are used as the finally age-invariant representation for face recognition.

The structure of the convolution unit follows typical CNNs, alternatively stacking convolution layer, activation layer and pooling layer. We construct the convolution unit with 5 convolution layers as shown in Figure 1. The filter sizes in both convolution and local convolution layers are 3×3 with stride 1, followed by PReLU nonlinear units [33]. Weights in three local convolution layers are locally shared in the regions of 4×4, 2×2 and 1×1 respectively. The number of the feature maps are 128 for the convolution layers and 256 for the local convolution layers. The max pooling is used for enhancing the robustness to potential translation and subsampling.

The convolutional unit is then followed by two parallel subnetwork classifiers which shares the same feature t from the convolutional unit as their input. One of them is the identity subnetwork which extracts the identity feature t_{id}, while the other is the auxiliary subnetwork that extracts the auxiliary feature t_{aux}. Note that both subnetworks consist of FC layers and their corresponding loss functions, i.e. the two subnetworks are equivalent to the following matrix operations:

$$\begin{aligned} t_{id} &= Ut + \alpha \\ t_{aux} &= Vt + \beta \end{aligned} \tag{1}$$

In our implementation, we use softmax loss and center loss as our cost function to train the networks. We adopt these losses for the following reasons: First, the softmax loss is adopted by both the identity network and the auxiliary network to extract age-invariant and age-related information respectively. Second, the recently proposed center loss [28] is proven to be very effective in boosting the discriminative ability of CNNs, and therefore we adopt it in our identity network. The center loss is not adopted in the auxiliary network, since it does not need to extract discriminative features. Apart from that, the goal of center loss is to decrease the intra-class distance between different samples of the same class. The auxiliary task tries to classify sample images of different identities and ages. Since the number of samples for each group is small, the center loss is no longer effective for such task. Hence, the loss function of each of the two subnetworks for identity classification and auxiliary task can be represented as:

$$\begin{aligned} L_{id}(D;U,\alpha) &= -\sum_{k=1}^{L} log\frac{e^{W_{id}^{\mathrm{T}}t_{id}^{k}+b_{id}^{k}}}{\sum_{i=1}^{n} e^{W_{id}^{i\ \mathrm{T}}t_{id}^{k}+b_{id}^{i}}} \\ &\quad + \frac{\gamma}{2}\sum_{k=1}^{L}\|t_{id}^{k} - c_{y_{id}}\|_2^2 \\ L_{aux}(D;V,\beta) &= -\sum_{k=1}^{L} log\frac{e^{W_{aux}^{\mathrm{T}}t_{aux}^{k}+b_{aux}^{k}}}{\sum_{i=1}^{n} e^{W_{aux}^{i\ \mathrm{T}}t_{aux}^{k}+b_{id}^{i}}} \end{aligned} \tag{2}$$

where $W_{id}, W_{aux}, b_{id}, b_{aux}$ are the parameters of the softmax classifier. When we perform face recognition tasks, we take out the identity feature vectors t_{id} of the images from the probe and gallery set and calculate their cosine similarity as the comparison score. As we will elaborate in the next section, a factor analysis model is applied to learn the parameters $\{U, V, \alpha, \beta\}$ for the subnetworks instead of the fc-layer. After that, these parameters will be fixed to update the parameters in the convolutional unit through stochastic gradient descent(SGD).

C. Joint Task Factor Analysis

In our scheme, the subnetwork classifiers H_{id} and H_{aux}, as we mentioned above, have parameters$\{U, V, \alpha, \beta\}$. Since some aging information is person-specific and cannot be entirely separated from identity feature in a linear way, we propose an auxiliary feature to represent all of the age-identity correlated information and separate it from the identity features. To extract the auxiliary features, the auxiliary classifier H_{aux} is set as a softmax classifier, in which each class represents a unique identity and age group. We develop a JTFA model that separates the identity-related features from the auxiliary features denoted as:

$$\begin{aligned} &\mathbf{t} = \boldsymbol{\mu} + F\mathbf{x}_i + G\mathbf{y}_{ij} + \epsilon, \\ &\forall i = 1\ldots N \quad and \quad j = 1\ldots g_i, M = \sum_i g_i \end{aligned} \tag{3}$$

The JTFA model decomposes a facial image into the linear combination of four components: (i) the mean vector $\boldsymbol{\mu}$, (ii) identity component $F\mathbf{x}$, (iii) age-related component $G\mathbf{y}$, and (iv) a noise term ϵ. An intuitive idea is to collect all the face images of different persons for each age group. However, different persons may have various face aging effect. Hence we use the identity-age groups instead of the age groups to classify our training set, so that the age-related component in the face features can be extracted more discriminatively. $\mathbf{x}_i$ and $\mathbf{y}_{ij}$ means the latent factor variables with identity information and age-related information. $\mathbf{x}_i$ changes only with different identities i while $\mathbf{y}_{ij}$ differs with different identities i and age-group j. $p(\mathbf{x}_i)$ and $p(\mathbf{y}_{ij})$ follow Gaussian probabilistic model. We denote I_n by n order unit matrix, then $p(\mathbf{x}_i)$ and $p(\mathbf{y}_{ij})$ are denoted as follow:

$$p(\mathbf{x}_i) = N(0, I_p), \;\; p(\mathbf{y}_{ij}) = N(0, I_q). \tag{4}$$

As the latent factors can not be observed directly, a critical step is to solve the model parameters $\theta = \{F, G, \epsilon\}$ through the learning process of the joint task factor analysis. We can learn these parameters by maximizing the following objective function:

$$Lc = \sum_i \sum_j \ln p_\theta(\mathbf{t}_k, \mathbf{x}_i, \mathbf{y}_{ij}). \tag{5}$$

We adopt an Expectation Maximization (EM) strategy to solve the above function, which alternately updates these model parameters and the statistics characters of latent factors. The priori distributions of latent factors are shown in (4). In particular, with the fixed θ, we can estimate the posterior distribution of the latent variables. The target of learning stage is to estimate the parameters θ based on training feature set $T_t = \{\mathbf{t}_k^s \in \mathbb{R}^d | k = 1 \ldots L, s = 1 \ldots L_k\}$, where d is the dimension of the data $\mathbf{t}_k^s$ and L_k is the number of images in the kth group. Given initial estimation θ_0, we can maximize the expectation of the objective function in (5) by updating θ:

$$E[Lc] = \sum_i \sum_j \iint p_{\theta_0}(\mathbf{x}_i, \mathbf{y}_{ij} | T) \times \ln p_\theta(\mathbf{t}_{ij}^s, \mathbf{x}_i, \mathbf{y}_{ij}) d\mathbf{x}_i d\mathbf{y}_{ij}. \tag{6}$$

Similarly, we denote that the ith person have N_i images and the jth identity-age group has M_{ij} images. The detailed EM steps are described below:

Expectation Step (E): According to (3), $\boldsymbol{\mu}$ is the mean of the training samples:

$$\boldsymbol{\mu} = \frac{1}{\sum_i N_i} \sum_i \sum_j \sum_{M_{ij}} \mathbf{t}_{ij}^s. \tag{7}$$

To optimize the objective function in equation (5), an estimation of joint distribution of the latent variables given model parameters $\{\mathbf{x}_i, \mathbf{y}_j\}$ are calculated. The estimation of the first and second conditional moments are given by:

$$\begin{aligned} E[\mathbf{x}_i] &= \frac{F^{\mathrm{T}}\Psi^{-1}}{N_i} \sum\nolimits_{n=1}^{N_i} (\mathbf{t}_k^s - \boldsymbol{\mu}), \\ E[\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}] &= \frac{I - F^{\mathrm{T}}\Psi^{-1}F}{N_i} + E[\mathbf{x}_i]E[\mathbf{x}_i]^{\mathrm{T}}. \\ E[\mathbf{x}_i \mathbf{y}_{ij}^{\mathrm{T}}] &= \frac{F^{\mathrm{T}}\Psi^{-1}G}{(N_i M_{ij})^{\frac{1}{2}}} + E[\mathbf{x}_i]E[\mathbf{y}_{ij}]^{\mathrm{T}}. \end{aligned} \tag{8}$$

where $\Psi = FF^{\mathrm{T}} + GG^{\mathrm{T}} + \epsilon$ and $E[\cdot]$ denotes expectation. The estimation of moments $E[\mathbf{y}_{ij}]$, $E[\mathbf{y}_{ij}\mathbf{y}_{ij}^{\mathrm{T}}]$, $E[\mathbf{y}_{ij}\mathbf{x}_i^{\mathrm{T}}]$ are calculated in a similar way. As we obtain the expectations of the latent factor variations, we can calculate the model parameters using maximum likelihood estimation in the next step.

Maximization Step(M): According to (6), we optimize parameters θ that maximize the expectation of the log-likelihood function. The likelihood is represented as:

$$\begin{aligned} p_\theta(\mathbf{t}_{ij}, \mathbf{x}_i, \mathbf{y}_{ij}) = (2\pi)^{-(d+p+q)/2} |\epsilon|^{-1/2} exp((\mathbf{t}_{ij} - \boldsymbol{\mu} \\ - F\mathbf{x}_i - G\mathbf{y}_{ij})^{\mathrm{T}} \epsilon^{-1} (\mathbf{t}_{ij} - \boldsymbol{\mu} - F\mathbf{x}_i - G\mathbf{y}_{ij})) \\ \times exp(-\frac{1}{2}(\|\mathbf{x}_i\|_2 + \|\mathbf{y}_{ij}\|_2)). \end{aligned} \tag{9}$$

Thus, the expectation of F derivative of the log-likelihood has the form of:

$$\begin{aligned} E[\frac{\partial Lc}{\partial F}] = \sum_{i=1}^{N} \sum_{j=1}^{g_i} \sum_{s=1}^{M_{ij}} \iint p(\mathbf{x}_i, \mathbf{y}_{ij} | T, \theta_0) \epsilon^{-1} \\ \times \{(\mathbf{t}_{ij} - \boldsymbol{\mu} - G\mathbf{y}_{ij})\mathbf{x}_i^{\mathrm{T}} - F\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\} d\mathbf{x}_i d\mathbf{y}_{ij}. \end{aligned} \tag{10}$$

Then the optimal F is obtained at $E[\frac{\partial Lc}{\partial F}] = 0$, as well as optimal G, which can be determined by:

$$\begin{aligned} [F \vdots G] \sum_{i=1}^{N} \sum_{j=1}^{g_i} M_{ij} \begin{bmatrix} E[\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}] & E[\mathbf{x}_i \mathbf{y}_{ij}^{\mathrm{T}}] \\ E[\mathbf{y}_{ij} \mathbf{x}_i^{\mathrm{T}}] & E[\mathbf{y}_{ij} \mathbf{y}_{ij}^{\mathrm{T}}] \end{bmatrix} \\ = \sum_{i=1}^{N} \sum_{j=1}^{g_i} \sum_{s=1}^{M_{ij}} [(\mathbf{t}_{ij}^s - \boldsymbol{\mu}) E[\mathbf{x}_i]^{\mathrm{T}} \vdots (\mathbf{t}_{ij}^s - \boldsymbol{\mu}) E[\mathbf{y}_{ij}]^{\mathrm{T}}]. \end{aligned} \tag{11}$$

With equation (11), we can solve $\{F, G\}$ by a simple linear transform. To calculate ϵ, a common constraint that $\epsilon = \sigma^2 I_d$ is used. And then the coefficient σ can be optimized by:

$$\sigma^2 = \sum_{i=1}^{N} \sum_{j=1}^{g_i} \sum_{s=1}^{M_{ij}} \frac{(\mathbf{t}_{ij}^s - \boldsymbol{\mu})^{\mathrm{T}} (\mathbf{t}_{ij}^s - \boldsymbol{\mu} - FE[\mathbf{x}_i] - GE[\mathbf{y}_{ij}])}{d \sum_{i=1}^{N} N_i}. \tag{12}$$

(11) and (12) indicate that we can obtain the estimation of all the parameters in θ only using the first-order and second-order statistics calculated in the E-step. In the M-step, the model parameters are updated to maximize the expectation of the log-likelihood probability that training samples are generated by these latent factors. After the model parameters converges, the identity factor for the ith class and the age-related factor for the (i, j) group can be inferred by the reconstruction of first moment of x. Denote the deep features F_{conv} extracted in the convolution unit as the input t_{ij}^s,

JTFA model can be formed by an fc-layer with the following parameters:

$$\begin{aligned} t_{id} &= Ut + \alpha, t_{aux} = Vt + \beta \\ U &= FF^{\mathrm{T}}\Psi^{-1}, \alpha = -FF^{\mathrm{T}}\Psi^{-1}\boldsymbol{\mu} \\ V &= GG^{\mathrm{T}}\Psi^{-1}, \beta = -GG^{\mathrm{T}}\Psi^{-1}\boldsymbol{\mu} \end{aligned} \tag{13}$$

Algorithm 1 Model Training for JTFA

Input: deep feature set of training images $T = \{\mathbf{t}_k \in \mathbb{R}^d | k = 1 \dots L\}$, with identity labels and age group information
Output: model parameters $\{U, V, \alpha, \beta\}$
1) Initialization: $F, G, H \leftarrow rand(-0.1, 0.1)$, $\sigma^2 \leftarrow 0.1$.
2) Compute the mean of the training samples μ with equation (7).
3) E-step: Compute the estimation of joint distribution of the latent variables with equation (8).
4) M-step: Update the model parameters $\theta = \{F, G, \sigma^2\}$ with equations (11) and (12).
5) Go to step 3) until convergence.
6) Compute the model parameters with equation (13)

D. AFJT-CNN Learning Framework

In this section we introduce the learning strategy of our AFJT-CNN model. One advantage of CNN is that the entire network is capable of learning features from raw image data without prior knowledge. The main disadvantage of this model, however, is that it requires a large number of training samples in order to achieve a competitive performance. In our implementation, the face dataset without age information are used to initialize the network, while the cross-age face dataset are used to alternately fine-tune the convolutional unit and JTFA FC-layer.

By combining the discrimination of CNN and aging robustness of JTFA, we alternately optimize the parameters of the multiloss network and the parameters for JTFA. In the multiloss network training, the convolutional unit tends to extract features that include both the identity and the person-specific aging information. More specifically, after initializing the face identity subnetwork using the web-scale face training sets, the convolutional unit is shared by the age-related subnetwork. Then we calculate the JTFA model parameters using the cross-age face datasets for the FC layer of two subnetworks. Furthermore, the cross-age face datasets are used to alternately train the shared convolutional unit W and the FC layer U, V, α, β. During the training process, the parameters W of the convolutional unit W_{conv} are updated based on the parameters of the identity factor guided FC layer, while in the next cycle, $\{U, V, \alpha, \beta\}$ are also updated using the new W. The procedure iteratively goes on as is shown in Algorithm 2.

IV. Experiments

In this section we perform experiments on some of the most commonly used databases to demonstrate the effectiveness of our scheme. The databases we used include Morph Album 2 [13], CACD [14] and Cross Age LFW [15], and the

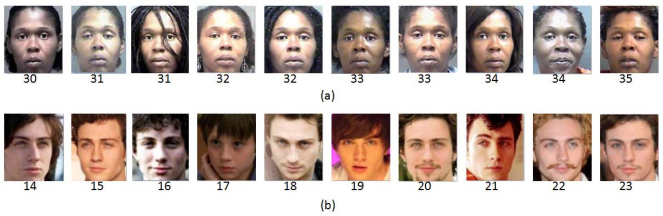


Fig. 2. Cropped images in dataset (a) MORPH Album 2 and (b) CACD.

Algorithm 2 Model Training for AFJT-CNN

Input: Cross age face image training set $T_a = \{\mathbf{I}_{ij}^s | i = 1 \dots N, j = 1 \dots g_i, s = 1 \dots M_j\}$, with identity labels and age group information and general face image training set with just identity labels $T_i = \{\mathbf{I}_i^s | i = 1 \dots L, s = 1 \dots L_i\}$,
Output: convolution unit parameters W and JTFA parameters $\{U, V, \alpha, \beta\}$
1) $p \leftarrow 0$.
2) Initialize the parameters W^0 by Xaiver filter.
3) Initialization: $F, G \leftarrow rand(-0.1, 0.1)$, $\sigma^2 \leftarrow 0.1$.
4) Pretraining the parameters W^0 on the identity subnetwork using the general face imagesT_i.
5) while not converge do
6) $p \leftarrow p + 1$
7) Fix W^p and train the JTFA model parameters $\theta = \{F, G, \sigma^2\}$ with training set T_a.
8) Calculate $\{U^p, V^p, \alpha^p, \beta^p\}$ using equations (13) and replace the two FC layer.
9) Update the convolution unit parameters W^{p+1} with training set T_a using the joint task multiple loss, the parameters are shared by two subnetworks.
10) Go to step 5) until convergence.

testing result has demonstrated the superiority of our scheme over most of the previous approaches.

For each face image in the datasets, the recently proposed MTCNN algorithm [34], which proposes a cascaded CNN based framework for real time joint face detection and alignment, is used to detect the facial landmarks. Then the faces are globally cropped to 112×96 according to the 5 facial landmarks (two eyes, nose and two mouth corners) by similarity transformation. The images are horizontally flipped for data augmentation. Example images are shown in Fig. 2. In our experiments, we compare the proposed approach with state-of-the-art cross-age face recognition methods including: (i) Gong's HFA method [12]. (ii) Chen's CARC method [14] (iii) Li's MEFD method [32]. (iv) Gong's LF-CNN method [30]. In above mentioned methods, they perform face identification experiments on MORPH Album 2 dataset, and face verification experiments on CACD-VS dataset and Cross Age LFW dataset. For a fair comparison between our method and other state of the art methods, we follow the same training and testing splits as these papers in our experiments. We use the same training strategies and testing procedures as [30] on MORPH and CACD-VS dataset, and we test our method on CALFW dataset following the

unrestricted with labeled outside data protocol [15]. To separate face identification from face verification, identification tasks and verification tasks are introduced in subsection B and C respectively.

A. Detail Settings in our Experiments

We implement the AFJT-CNN model using the Tensorflow library [35] and MATLAB R2015b software. Unless otherwise specified, the batch size is 128. The experiments are performed on a computer with an Intel E5-2680v3 CPU and a NVIDIA GTX1080Ti graphic card. Our model takes up 8.6 GB video memory. We first pretrain the convolutional unit with identity subnetwork for about 12 epoches using the CACD and CASIA dataset, which have 680K face images. When testing our model on CACD-VS, we remove all the identities in CACD-VS from the training data. The learning rate of CNN pretraining is set to 1e-3. When the training error plateaus, it decays to 1e-4 and 1e-5. The total time cost of pretraining is about 20 hours. Then the alternative training procedure costs 12 hours. When evaluating our model on the test set, the time cost is 1.08ms of each image. The weight of the center loss is set as $\lambda = 0.008$. To better evaluate the performance of the proposed age-invariant deep face features, our model uses the simple Cosine Distance and the Nearest Neighbor rule as the classifier. The cross-age face datasets are divided into 8 different aging groups such that faces in different groups have different auxiliary label in the auxiliary classifier. The code of our method is published at https://github.com/lz00309/AFJTCNN.

B. Experiments for Face Identification

1) Experiments on Morph Dataset: The MORPH Album 2 dataset is one of the largest publicly available face aging dataset, composed of about 55,000 face images of 13367 persons captured at different ages. Following the configurations of training and testing split in [20] and [30]we use all the persons and partition them into the training set and the test set. The multiloss network is pretrained on the CASIA and CACD dataset. Then we finetune the network and perform JTFA on all the face images from the training set of half of the persons. The test set comprises a gallery and a probe set selected from the rest persons. The gallery consists of the face images corresponding to the youngest age of these persons, and the probe set consists of images of the oldest ages for each person. In order to verify the contribution of the JTFA, we conduct an ablation experiment by finetuning the multiloss network on our training set without JTFA training and with the fc-layer unfixed. With the same experiment settings as [30], we compare the AFJT-CNN model with several state-of-the-art methods, and the rank-1 identification rates are reported in Table I.

From these results in Table I, we can see that our method remains the best performance. We have the following observations. First, the identification rate of the CNN-baseline is only 91.67%, which is inferior to most of the other results in Table 1. This shows that directly applicating the deep CNN model to cross age face datasets is not robust for identification task. Second, the performance of CNN-baseline can be improved to 95.90% by fine-tuning with the additional MORPH training data. However, this result (95.90%) is still inferior to the top-performing result of the LF-CNN method in the literature (97.51%). This is because the CNN baseline model has taken no consideration regarding the aging influence. The result of the multiloss network without JTFA (96.04%) shows that without JTFA training, the improvement with the auxiliary network is limited, because the network has no ability to separate the collected age-related auxiliary information from the face identity features. Therefore it is natural to expect that the significant improvement of our AFJT-CNN method results from both multiloss training and JTFA.

TABLE I

RANK-1 IDENTIFICATION RATE ON MORPH DATASET

Method	Identification rate(%)
HFA(2013) [12]	91.14
CARC(2014) [14]	92.80
MEFA(2015) [32]	93.80
MEFA+SIFT+MLBP(2015) [32]	94.59
LF-CNNs(2016) [30]	97.51
CNN-baseline	91.67
CNN-baseline(finetuned by MORPH)	95.90
CNN(finetuned with joint task multiloss)	96.04
AFJT-CNN(proposed)	**97.85**

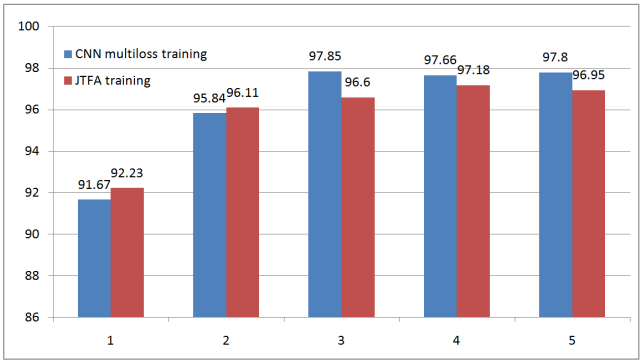


Fig. 3. Rank-1 Identification rate changes with different iterations of training on MORPH.

2) Effectiveness of Multiloss Training: To evaluate the multiloss architecture of our AFJT-CNN model, we design an experiment to test the recognition performance of our model in each iteration of the learning process in Algorithm 2, as illustrated in Fig. 3. The two parameters W_{conv} and $\{U, V, \alpha, \beta\}$ are updated in a step-wise manner, i.e. to update one parameter while fixing the other, and vice versa. Fig. 3 clearly shows that the iterative learning process consistently contributes to the performance improvement of cross age face recognition, converging to a good result quickly. During the first two iterations the recognition performance increases continually. When the performance reaches a peak, there are gaps under the same number of iteration between the convolutional unit training and factor analysis training. It is due to the combination of center loss and JTFA. The main goal of center loss is to reduce the intra-class distance between different samples of

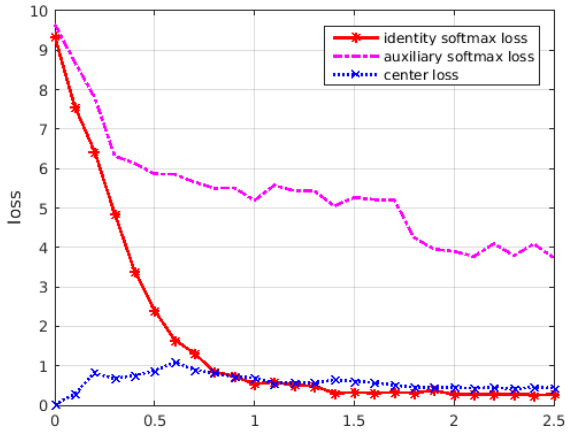


Fig. 4. The multiple loss of AFJT-CNN during fine-tuning on MORPH.

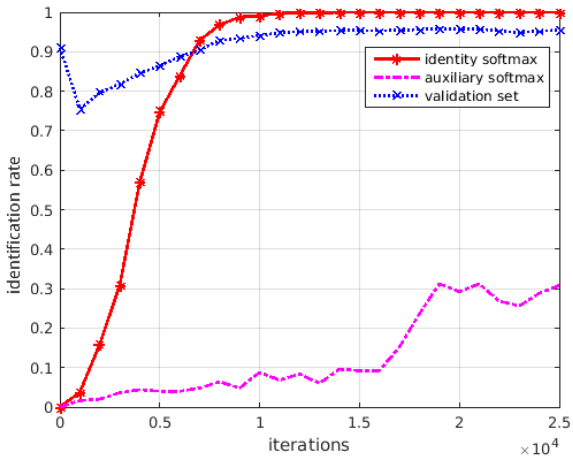


Fig. 5. Rank-1 Identification rate changes of the sub-classifiers and validation dataset during fine-tuning on MORPH.

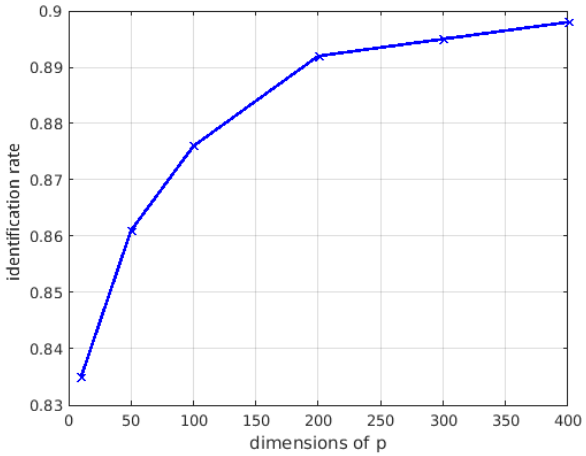


Fig. 6. Identification rate changes with different p of JTFA training on MORPH. q is fixed to 0.

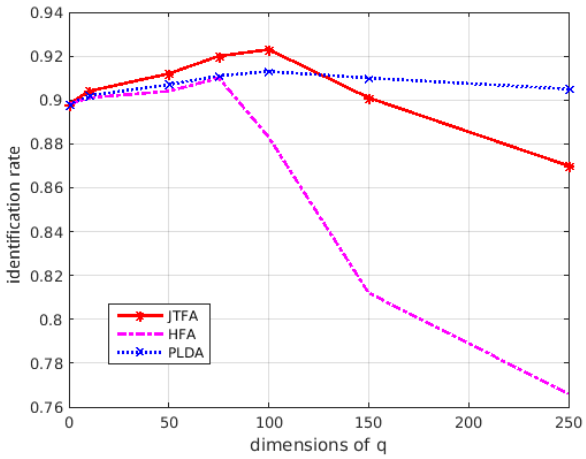


Fig. 7. Identification rate changes with different q of JTFA training on MORPH. p is fixed to 400.

the class. The factor analysis, however, maps all classes into a different feature space, which can have an adverse impact on discrimination ability and recognition rate.

To show how the auxiliary network impacts the identification task during multiloss finetuning, we report the change of loss function and recognition performance as illustrated in Fig.4 and Fig. 5. During the first 11000 iterations, the performances of identity and auxiliary softmax classifiers both increase, while the recognition performance of the validation dataset continues to improve. From iteration 11000 to 19000, the performance of identity classifier on the training set edges up to 100%, while the loss of auxiliary classifier was still decreasing. Meanwhile, the recognition performance of the validation set is still promoting, which shows that the newly added auxiliary classifier helps improve the recognition performance during training of convolutional unit. After iteration 19000, while the loss of auxiliary classifier keeps declining, the identification rate of the validation set decreases, implied that the network has been over-fitting. This phenomenon shows that our proposed multiple loss network can profoundly improve the robustness of CNNs for cross age face recognition.

3) Effectiveness of JTFA Model: We have conducted experiments to verify the advantages of our proposed JTFA algorithm. We compare our JTFA model with two original factor analysis methods: (i)HFA model and (ii)PLDA model. Note that the parameter q is the number of dimensions of the auxiliary latent factor. In HFA the auxiliary latent factor represents the age information, while in PLDA each training face image has a respective latent factor with q dimensions. When $q = 0$ the proposed method degenerates to a latent factor gaussian mixed model. Also, a larger parameter p means that the identity factor has higher dimensions, hence represents more identity information of the face image. A larger parameter q means higher dimensions and more extracted information of the auxiliary classifier as well. We output identification rate curves for two parameters p, q as shown in Fig. 6 and 7. The curves of parameter p shows that the dimension of identity-specific feature, extracted by the identity latent factor x, seizes to contribute to the overall performance when p reaches 400. To show the difference between HFA, PLDA and JTFA, the curves of parameter q is plotted in Fig. 7 with parameter p fixed by 400. When the auxiliary factor y is

TABLE II
FNMR@FMR = 0.01 AND EER ON CACD DATASET

Method	FNMR@FMR=0.01(%)	EER(%)
HD-LBP(2013) [36]	58.2	18.4
HFA(2013) [12]	52.1	15.6
CARC(2014) [14]	43.3	12.7
Human,Average(2015) [14]	46.9	13.5
Human,Voting(2015) [14]	15.6	5.5
LF-CNNs(2016) [30]	2.9	1.5
CNN-baseline	17.8	5.5
CNN(finetuned)	3.7	2.2
CNN(joint task multiloss)	3.8	2.2
AFJT-CNN	**1.0**	**1.0**

deployed(the parameter q becomes larger), the performance is obviously improved in all the factor analysis methods. It is shown in Fig. 7 that our method performs best when $q = 100$, reaching the highest peak in the figure, which demonstrates its superiority compared with the other factor analysis model.

C. Experiments for Face Verification

1) Experiments on CACD-VS Dataset: The CACD dataset is a recently released dataset for cross-age face recognition, containing 163,446 images from 2,000 celebrities with labeled ages. It includes varying illumination, pose variation and makeup and better simulates practical scenario. However, the entire CACD dataset contains some incorrectly labeled samples, and some duplicate images. With same experimental settings as [30], we test AFJT-CNN on CACD-VS, i.e., a subset of CACD, which consists of 4000 image pairs (2000 positive pairs and 2000 negative pairs) and have been carefully annotated. Similar to the above section, we first pretrain the network without the auxiliary network on CASIA and CACD dataset. Note that the corresponding samples in the training set of the persons included in the CACD-VS are removed in this experiment. Then we alternatively train the whole network and perform JTFA on the CACD training set. The identities in CACD-VS are excluded from the training data. After the algorithm converges, we use the fully connected-layer of identity subnetwork as the feature of each face image. Cosine distance is used as our evaluation metric.

We report FNMR@FMR=0.01(false non-match rate when false match rate is 1%) and EER(equal error rate) in Table II. From these results, we can find that our AFJT-CNN model beats the CNN baseline one by a significant margin, which further demonstrates the effectiveness of the proposed AFJT-CNN training scheme. On the other hand, the performance of AFJT-CNN outperforms all the state-of-the-art methods, which demonstrates the superiority of our method. The ablation experiment results, that we just finetune with CACD on a baseline CNN and a multiloss network without JTFA training, are also reported. The results of CNN (finetuned with joint task multiloss) is close to the result of CNN finetuned by CACD. The results show that without JTFA training, the added auxiliary network even does not improve the discriminative performance for cross-age face recognition. Fig. 8 shows that the iterative learning process consistently contributes to the EER changes of cross age face recognition, converging to a good result more quickly than MORPH.

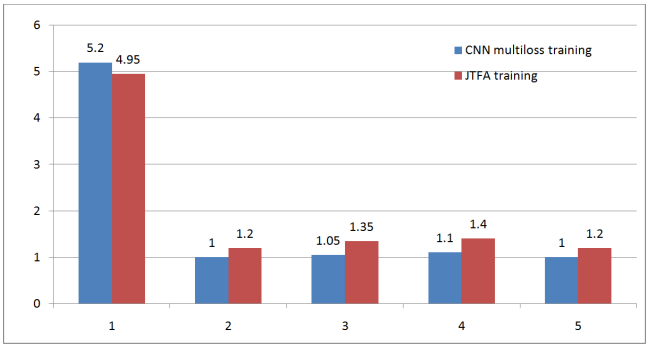


Fig. 8. EER on CACD-VS changes with different iterations of training on CACD.

TABLE III
EER OF DIFFERENT METHODS ON CALFW DATASET

Method	Training Images	EER(%)
VGG-Face [37]	2.6M	**13.5**
Noisy Softmax [38]	0.5M	17.5
CNN-baseline	0.7M	18.1
CNN(joint task multiloss)	0.7M	18
AFJT-CNN	0.7M	**14.8**

TABLE IV
FNMR@FMR = 0.1 ON CALFW DATASET

Method	Training Images	FNMR@FMR=0.1(%)
VGG-Face [37]	2.6M	**17.6**
Noisy Softmax [38]	0.5M	29.2
CNN-baseline	0.7M	34.3
CNN(joint task multiloss)	0.7M	34.1
AFJT-CNN	0.7M	**21.8**

2) Experiments on Cross Age LFW Dataset: Cross-Age LFW (CALFW) is a newly proposed benchmark dataset which contains 4,025 individuals with 2, 3 or 4 images for each persons. It selects 3,000 positive face pairs with age gaps to add aging process intra-class variance. Negative pairs with same gender and race are selected to reduce the influence of attribute difference between positive/negative pairs and achieve face verification instead of attributes classification.

Following the unrestricted with labeled outside data protocol [15], we train on the both CACD and CASIA and test on 6,000 face pairs. The tested face images are aligned by similar transformation using the 5 face landmarks proposed by the dataset and then resize to 112×96. In Table III and Table IV we compare the results of EER and FMR@FNMR = 0.1 among our method and the baselines (including VGG-Face [37], Noisy Softmax [38] and our CNN baseline) on CALFW datasets.

From the results, we have the following observations. Firstly, AFJT-CNN model beats the CNN baseline by a significant margin. The result shows that the proposed method can notably enhance the robustness of cross-age face recognition, demonstrating the effectiveness of AFJT-CNN. Secondly, CNN with joint task multiloss performs just slightly better than the baseline but worse than the AFJT-CNN, which shows that the JTFA training procedure plays an important role in

improving the recognition performance. Finally, we can see that our method can obtain comparable results to the state-of-the-art approaches using relatively small training data from sources other than CALFW, demonstrating the generalization ability of our approach.

V. Conclusion

In this paper, we have proposed a probabilistic Age-related Factor guided Joint Task Convolutional Neural Networks (AFJT-CNN) approach to address the challenging problem of cross-age face recognition. The basic idea of the AFJT-CNN model is to pursue a more robust age-invariant deep face feature descriptor by training an auxiliary network to separate age-related information from identity features. Extensive experiments conducted on three benchmark face aging datasets (MORPH Album2, CACD-VS and CALFW) convincingly demonstrate that our proposed method has a more effective performance than the state-of-the-art approaches.

References

[1] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, p. 37, 2016.

[2] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3539–3545.

[3] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. V. Rohith, "Fully automatic pose-invariant face recognition via 3D pose normalization," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 937–944.

[4] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 980–993, Mar. 2015.

[5] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[6] X. Zou, J. Kittler, and K. Messer, "Illumination invariant face recognition: A survey," in *Proc. IEEE Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2007, pp. 1–8.

[7] J. Anil and L. P. Suresh, "Literature survey on face and face expression recognition," in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Mar. 2016, pp. 1–6.

[8] R. Jameel, A. Singhal, and A. Bansal, "A comprehensive study on facial expressions recognition techniques," in *Proc. 6th Int. Conf. Cloud Syst. Big Data Eng. (Confluence)*, Jan. 2016, pp. 478–483.

[9] V. R. Rai *et al.*, "Recognizing images across age progressions: A comprehensive review," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 572–576.

[10] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the FG-NET ageing database," *IET Biometrics*, vol. 5, no. 2, pp. 37–46, May 2016.

[11] A. Lanitis, "A survey of the effects of aging on biometric identity verification," *Int. J. Biometrics*, vol. 2, no. 1, pp. 34–52, 2010.

[12] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2872–2879.

[13] K. Ricanek, Jr., and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. FGR*, Apr. 2006, pp. 341–345.

[14] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 804–815, Jun. 2015.

[15] T. Zheng, W. Deng, and J. Hu. (Aug. 2017). "Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments." [Online]. Available: https://arxiv.org/abs/1708.08197

[16] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[17] U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 947–954, May 2010.

[18] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[19] J. Suo, S.-C. Zhu, S. Shan, and X. Chen, "A compositional and dynamic model for face aging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 385–401, Mar. 2010.

[20] Z. Li, U. Park, and A. K. Jain, "A discriminative model for age invariant face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 1028–1037, Sep. 2011.

[21] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, "Face verification across age progression using discriminative methods," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 1, pp. 82–91, Mar. 2010.

[22] C. Otto, H. Han, and A. Jain, "How does aging affect facial components?" in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 189–198.

[23] L. Du and H. Ling, "Cross-age face verification by coordinating with cross-face age verification," in *Proc. CVPR*, Jun. 2015, pp. 2329–2338.

[24] D. Bouchaffra, "Nonlinear topological component analysis: Application to age-invariant face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1375–1387, Jul. 2015.

[25] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[26] Y. Sun, X. Wang, and X. Tang. (Jun. 2014). "Deep learning face representation by joint identification-verification." [Online]. Available: https://arxiv.org/abs/1406.4773

[27] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[28] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016.

[29] L. Liu, X. Chao, H. Zhang, Z. Niu, M. Yang, and S. Yan, "Deep aging face verification with large gaps," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 64–75, Jan. 2016.

[30] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4893–4901.

[31] H. Li, H. Zou, and H. Hu, "Modified hidden factor analysis for cross-age face recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 465–469, Apr. 2017.

[32] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li, "A maximum entropy feature descriptor for age invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5289–5297.

[33] K. He, X. Zhang, S. Ren, and J. Sun. (Aug. 2017). "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification." [Online]. Available: https://arxiv.org/abs/1502.01852

[34] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[35] M. Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/about/bib

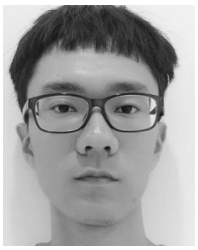
[36] C. Dong, C. Xudong, W. Fang, and S. Jian, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3025–3032.

[37] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, p. 6.

[38] Y. Wen, Z. Li, and Y. Qiao, "Noisy softmax: Improving the generalization ability of DCNN via postponing the early softmax saturation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 4021–4030.

Haoxi Li (S'16) received the B.S. degree from the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, where he is currently pursuing the M.S. degree. His research interests include the development and use of machine learning techniques for computer vision problems.

Haifeng Hu received the Ph.D. degree from Sun Yat-sen University in 2004. He has been an Associate Professor with the School of Electronics and Information Engineering, Sun Yat-sen University, since 2009. He is currently a Visiting Professor with the Robotics Institute, Carnegie Mellon University. His research interests are in computer vision, pattern recognition, image processing, and neural computation. He has authored about 80 papers since 2000.



Chitung Yip is currently pursuing the degree with the School of Electronics and Information Engineering, Sun Yat-sen University, China. Her research interests include pattern recognition, machine learning, and computer vision, with specific interest in face recognition.