Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition

Yandong Wen[1,2], Zhifeng Li[2]*, Yu Qiao[2]
[1]School of Electronic and Information Engineering, South China University of Technology
[2]Shenzhen Key Lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China
yd.wen@siat.ac.cn, zhifeng.li@siat.ac.cn, yu.qiao@siat.ac.cn

Abstract

While considerable progresses have been made on face recognition, age-invariant face recognition (AIFR) still remains a major challenge in real world applications of face recognition systems. The major difficulty of AIFR arises from the fact that the facial appearance is subject to significant intra-personal changes caused by the aging process over time. In order to address this problem, we propose a novel deep face recognition framework to learn the age-invariant deep face features through a carefully designed CNN model. To the best of our knowledge, this is the first attempt to show the effectiveness of deep CNNs in advancing the state-of-the-art of AIFR. Extensive experiments are conducted on several public domain face aging datasets (MORPH Album2, FGNET, and CACD-VS) to demonstrate the effectiveness of the proposed model over the state-of-the-art. We also verify the excellent generalization of our new model on the famous LFW dataset.

1. Introduction

Face recognition is one of the most active areas in computer vision community. It has been studied for several decades with substantial progresses. Most of the existing research focuses on general face recognition. There are very limited work directly on age-invariant face recognition (AIFR), which aims to address the face matching problem in the presence of remarkable aging variations [26].

AIFR has many useful and practical applications. For instance, it can be applied in finding missing children after years or checking whether the same person has been issued multiple government documents in different ages. However, it still remains a challenging problem in real world applications of face recognition systems. The major challenge of AIFR is mostly attributed to the great changes in face appearance caused by aging process over time. Figure 1 is a typical example, in which the cross-age face images from the same person have significant intra-personal changes.

Figure 1. Cross-age face images for one of the subjects in the FGNET dataset [1]. One can see the significant intra-personal variation therein.

The previous research on AIFR falls into two categories: generative approaches and discriminative approaches. The generative approaches [6, 15, 24] first synthesis face that matches target age and then perform recognition. Due to the strong parametric assumptions and the complexity in modeling aging process, these methods are computationally expensive and the results are often unstable in real-world face recognition scenarios. Recently discriminative approaches [13, 18, 19, 22, 4, 16, 17] draw increasing attentions. However, the features they use still contain age information, which is detrimental to the subsequent classification. To separate the age and the identity information, [7] proposes the hidden factor analysis (HFA) method. It formulates face feature as the linear combination of an identity component and an age component. Then the identity component is used for face recognition. [8] reports a more effective maximum entropy feature descriptor for AIFR, and proposes a more robust identity matching framework. However, most of the existing methods in AIFR rely heavily on the hand-crafted feature descriptors to extract the dense features for age-invariant face recognition, which may limit the performance of these methods. Designing an effective age-invariant face features still remains an open problem in AIFR.

As one of the most promising feature learning tools

*Corresponding author.

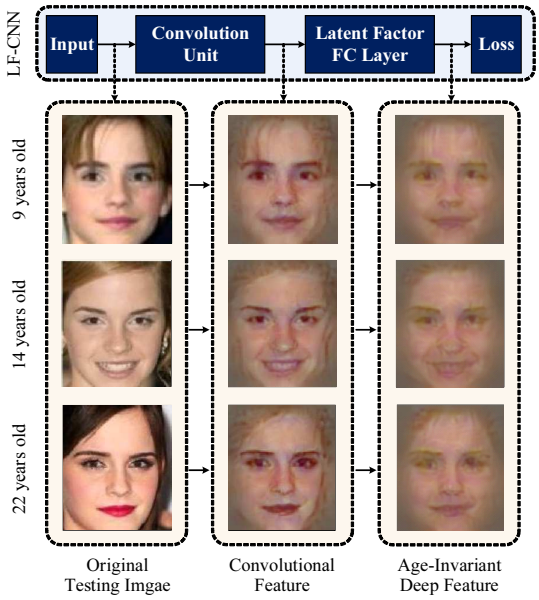

Figure 2. Cross-age faces processed by the proposed LF-CNNs. We visualize [35] the convolutional features and the age-invariant features.

nowadays, deep convolution neural networks (CNNs) have been successfully applied to a variety of problems in computer vision, including object detection and classification [14, 23, 28], and face recognition [33, 32, 29, 20, 36], etc. So it is desirable to use the deep learning model to address the AIFR problem. Surprisingly, there is no such work showing the superiority of deep learning on AIFR in the literature, to the best of our knowledge. A possible reason is the lack of a very suitable face aging dataset that can be used to train a robust deep learning model specifically for AIFR. For all the existing face aging datasets, each subject has very limited number of training samples across different ages, which are not suitable to serve as the training data in deep models. If we use the large scale web-collected face images to train the deep CNNs, the learned deep features will inevitably contain both identity-related component (e.g. ethnicity, gender) and identity-unrelated component (e.g. age, noise) [32]. Ideally, we expect the resulting deep feature contains only the identity-related components, reducing the variations caused by the aging process as much as possible.

In this paper, we explore the use of deep CNNs in AIFR and propose a *latent factor guided CNN (LF-CNN)* framework to learn the age-invariant deep face features. Specifically, we extract the age-invariant deep features from convolutional features by a carefully designed fully connected layer, termed as latent factor fully connected (LF-FC) layer. For this purpose, we develop a latent variable model, called latent identity analysis (LIA), to separate the variations caused by the aging process from the identity-related components in the convolutional features. The parameters of the LIA model are used to update the parameters of LF-FC layer. Moreover, the LIA model and the loss function in CNNs constitute the age-invariant identity loss, which is used to guide the learning of the LF-CNNs. In this way, our model is more adapted to age-invariant face recognition problem, as supported by our experimental results in Section 4. Figure 2 is a visualization example of the age-invariant deep features and the convolutional features, from which we can clearly see that the convolutional features are still age sensitive, while the age-invariant features are robust to aging process.

The major contributions of this paper are summarized as follows:

- We propose a robust age-invariant deep face recognition framework. To the best of our knowledge, it is the first work to show the effectiveness of deep CNNs in AIFR and achieve the best results to date.

- Instead of directly applying deep learning model to AIFR, we propose a new model called latent factor guided convolutional neural network (LF-CNN) to specifically address the AIFR task. By coupled learning the parameters in CNNs and LIA, the age-invariant deep face features can be extracted, which are more robust to the variations caused by the aging process over the time.

- Extensive experiments have shown that the proposed approach significantly outperforms the state-of-the-art on all the three face aging datasets (MORPH Album2 [27], FG-NET [1] and CACD-VS [4]), even beating the human voting performance on the CACD-VS dataset. We further demonstrate the excellent generalizability of our approach on the famous LFW [11] dataset.

2. Related Work

2.1. Convolutional Networks for Face Recognition

CNNs play a significant role in recent advances of face recognition. DeepFace [33] reports that a deeply-learned face representation achieves the accuracy close to human-level performance on LFW dataset [11]. [32] learns a deep CNN with the identification-verification supervisory signal and further adds supervision to early convolutional layers, greatly boosting the face recognition accuracy. FaceNet [29] 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, [20] 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.

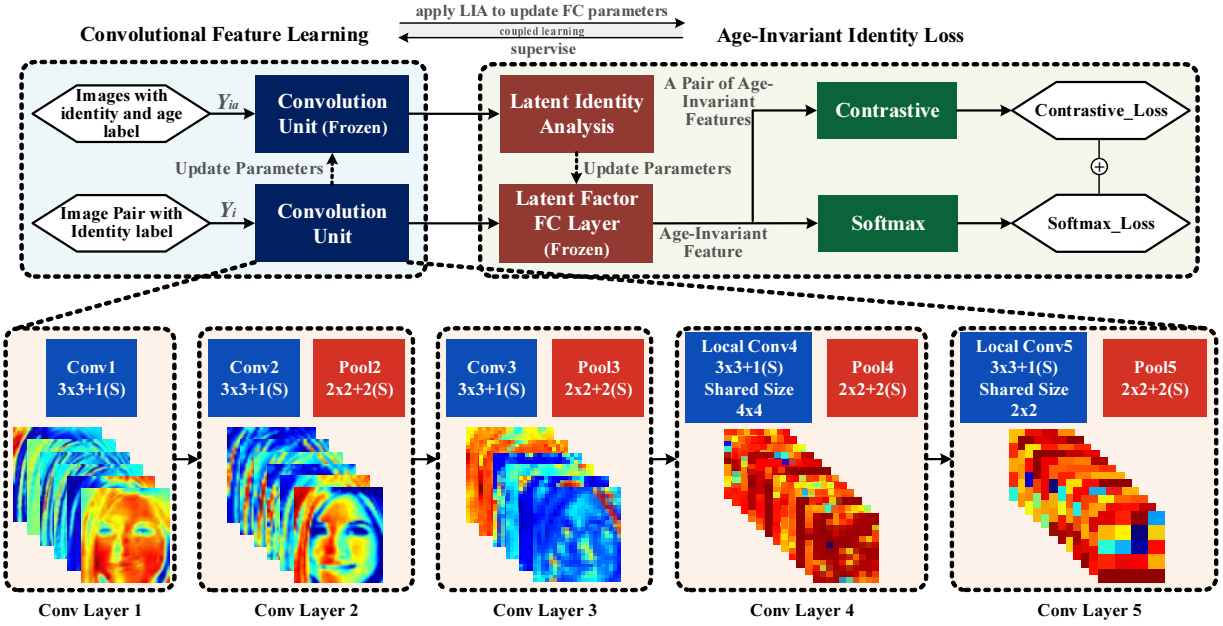

Figure 3. The architecture of the proposed LF-CNNs and its training process. *Frozen* layer only performs regular forward and backward calculations, but does not update their parameters (in other words, the parameters of this layers are fixed). The outside data $\boldsymbol{Y}_{\text{ia}}$ and the training data $\boldsymbol{Y}_{\text{ia}}$ are trained differently. Specifically, $\boldsymbol{Y}_{\text{ia}}$ and $\boldsymbol{Y}_{\text{ia}}$ are used for training the convolutional unit and the LF-FC layer respectively, following different pipelines. The *two parallel* convolution units are corresponding to a physical module in two stages (frozen and not frozen).

2.2. Latent Variable Model

The Latent variable model finds the latent variable that are not directly observed but inferred through a statistical model from observations. Latent variable models are widely used in recommended systems [10]. A few recent work [7, 8] apply the latent variable model to face recognition and achieves impressive performance. However, how to smoothly apply latent variable model in CNNs to achieve robust recognition still remains to be explored.

3. The Proposed Method

3.1. The LF-CNNs Model

The LF-CNN model is composed of two key components: convolution unit for convolution feature learning and LF-FC layer for age-invariant deep feature learning. The architecture of the LF-CNNs is shown in Figure 3.

The structure of the convolution unit in LF-CNNs follows typical CNNs, alternatively stacking convolution layer, nonlinearity layer, and optional pooling layer. We construct the convolution unit with 5 convolution layers as shown in Figure 3. The convolution kernel size and stride are set as 3×3 and 1 respectively, to capture more facial details in raw images [30]. In the 4th and 5th layer, the weights of convolution are locally shared to learn different mid-level and high-level features from different regions [33]. The five convolution layers output 128, 128, 128, 256 and 256 feature maps respectively. The nonlinear function is the Parametric Rectified Linear Unit (PReLU) [9], which improves model fitting. The max pooling is used for enhancing the robustness to potential translation and subsampling.

Next we focus on the construction of the LF-FC layer. Notice that the FC layer is equivalent to the matrix multiplication: $\boldsymbol{F}_{fc} = \boldsymbol{W}\boldsymbol{F}_{conv} + \boldsymbol{b}$ where $\boldsymbol{F}_{fc}$ is the output of FC layer, $\boldsymbol{F}_{conv}$ is the convolutional feature, and $\boldsymbol{W}, \boldsymbol{b}$ are the parameters of the FC layer. We leverage such equivalence property to design a set of $\boldsymbol{W}, \boldsymbol{b}$ that can extract the age-invariant feature from $\boldsymbol{F}_{conv}$. Note that, no nonlinearity layer will cascade to the LF-FC layer according to the LIA model. The feature dimension of $\boldsymbol{F}_{fc}$ is 512. Instead of iteratively updating the all the parameters in LF-CNNs by stochastic gradient descent (SGD), we design an *Latent Identity Analysis (LIA)* method to learn $\boldsymbol{W}$ and $\boldsymbol{b}$ for the LF-CNN model, as elaborated in the following subsection. For the parameters of the convolution unit, we fix $\boldsymbol{W}, \boldsymbol{b}$ to update them via standard SGD.

3.2. Latent Identity Analysis

The latent identity analysis model can infer the unobserved latent factors (one of them being the identity factor) from the observed features in a supervised fashion. The general model of latent identity analysis is formulated as

$$\boldsymbol{v} = \sum_{i=1}^{d} \boldsymbol{U}_i \boldsymbol{x}_i + \bar{\boldsymbol{v}} \tag{1}$$

where $\boldsymbol{v} \in \mathbb{R}^{n\times 1}$ denotes the observed facial features, $\boldsymbol{U}_i \in \mathbb{R}^{n\times p_i}$ is the corresponding matrix whose columns span the subspace of different variation and $\boldsymbol{x}_i \in \mathbb{R}^{p_i\times 1}$ denotes the latent variable with prior zero-mean Gaussian distribution. $\bar{\boldsymbol{v}} \in \mathbb{R}^{n\times 1}$ is the mean of all the facial features. The intuition behind this model is very clear. Each facial feature is viewed as the combination of different components according to different supervised signals. Such an idea is useful to achieve robust face recognition in practical. In the cross-age face recognition, we usually decompose the facial features into two latent components and a noise variable. So the model can be simplified as $\boldsymbol{v} = \boldsymbol{U}_{\text{id}}\boldsymbol{x}_{\text{id}} + \boldsymbol{U}_{\text{ag}}\boldsymbol{x}_{\text{ag}} + \boldsymbol{U}_{\text{e}}\boldsymbol{x}_{\text{e}} + \bar{\boldsymbol{v}}$ where $\boldsymbol{x}_{\text{id}}, \boldsymbol{x}_{\text{ag}}$ satisfy standard Gaussian distribution $\mathcal{N}(0, I)$ and $\boldsymbol{x}_{\text{e}}$ satisfies $\mathcal{N}(0, \sigma^2 I)$. Note that $\boldsymbol{U}_{\text{e}}$ is set to be a unix matrix since $\boldsymbol{x}_{\text{e}}$ stands for noise. $\boldsymbol{U}_{\text{id}}\boldsymbol{x}_{\text{id}}$ is the identity-related component which is key to achieve age-invariant face recognition. $\boldsymbol{U}_{\text{ag}}\boldsymbol{x}_{\text{ag}}$ is the age-related component, representing the age variations.

A set of model parameters $\theta = \{\boldsymbol{U}_{\text{id}}, \boldsymbol{U}_{\text{ag}}, \sigma^2, \bar{v}\}$ can be learned by maximizing the following maximum likelihood function $\mathcal{L}(\theta)$:

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{i,j} \mathcal{L}_{i,j}(\theta) \\ &= \sum_{i,j} \ln P(\boldsymbol{v}_i^j, \boldsymbol{x}_{\text{id},j}, \boldsymbol{x}_{\text{ag},j}|\theta)\end{aligned} \tag{2}$$

where $\boldsymbol{v}_i^j$ is the feature of the ith subject at jth age group. $\boldsymbol{x}_{\text{id},i}$ and $\boldsymbol{x}_{\text{ag},j}$ are the corresponding identity and age factors respectively. The summation is over all the available training samples from all subjects at all age groups. We use the expectation-maximization (EM) algorithm to estimate these model parameters. To perform the EM algorithm, the Q function is given by

$$\begin{aligned}\mathcal{Q}(\theta, \theta^{(i)}) &= \sum_{i,j} E\big\{ \ln P(\boldsymbol{v}_i^j, \boldsymbol{x}_{\text{id},j}, \boldsymbol{x}_{\text{ag},j}|\theta)\big\} \\ &= \sum_{i,j} \int_{\boldsymbol{x}_{\text{id,i}}, \boldsymbol{x}_{\text{ag,j}}} P(\boldsymbol{x}_{\text{id},j}, \boldsymbol{x}_{\text{ag},j}|\theta^{(i)}, \boldsymbol{V})\mathcal{L}_{i,j}(\theta)\end{aligned} \tag{3}$$

where $\boldsymbol{V}$ denotes the observed features of all the training samples, $\theta^{(i)}$ is the given model parameters and θ is the parameters to be estimated. With the given parameter $\theta^{(i)}$, we can compute the posterior distribution of the latent variables $P(\boldsymbol{x}_{\text{id},j}, \boldsymbol{x}_{\text{ag},j}|\theta^{(i)}, \boldsymbol{V})$. With the given posterior distribution, we can maximize the Q function to obtain a new θ. The EM algorithm is performed in an iterative fashion.

E Step. Given the model parameter $\theta^{(i)}$ and training data $\boldsymbol{V} = \{\boldsymbol{v}_i^j\}_{i=1,\cdots,N;j=1,\cdots,M}$, we first compute all the necessary first and second conditional moments of $P(\boldsymbol{x}_{\text{id},j}|\theta^{(i)}, \boldsymbol{V})$ and $P(\boldsymbol{x}_{\text{ag},j}|\theta^{(i)}, \boldsymbol{V})$ for the posterior distribution $P(\boldsymbol{x}_{\text{id},j}, \boldsymbol{x}_{\text{ag},j}|\theta^{(i)}, \boldsymbol{V})$:

$$\mu_1(\boldsymbol{x}_{\text{id},i}) = \frac{\boldsymbol{U}_{\text{id}}\Sigma^{-1}}{N_i}\sum_{k=1}^{N_i}(\boldsymbol{v}_i^k - \bar{\boldsymbol{v}}) \tag{4}$$

$$\mu_1(\boldsymbol{x}_{\text{ag},j}) = \frac{\boldsymbol{U}_{\text{ag}}\Sigma^{-1}}{M_j}\sum_{k=1}^{M_j}(\boldsymbol{v}_k^j - \bar{\boldsymbol{v}}) \tag{5}$$

$$\mu_2(\boldsymbol{x}_{\text{id},i}, \boldsymbol{x}_{\text{id},i}) = \frac{\boldsymbol{I} - \boldsymbol{U}_{\text{id}}^T\Sigma^{-1}\boldsymbol{U}_{\text{id}}}{N_i} + \mu_1(\boldsymbol{x}_{\text{id},i})\big(\mu_1(\boldsymbol{x}_{\text{id},i})\big)^T \tag{6}$$

$$\mu_2(\boldsymbol{x}_{\text{ag},j}, \boldsymbol{x}_{\text{ag},j}) = \frac{\boldsymbol{I} - \boldsymbol{U}_{\text{ag}}^T\Sigma^{-1}\boldsymbol{U}_{\text{ag}}}{M_j} + \mu_1(\boldsymbol{x}_{\text{ag},j})\big(\mu_1(\boldsymbol{x}_{\text{ag},j})\big)^T \tag{7}$$

$$\mu_2(\boldsymbol{x}_{\text{id},i}, \boldsymbol{x}_{\text{ag},j}) = -\frac{\boldsymbol{U}_{\text{ag}}^T\Sigma^{-1}\boldsymbol{U}_{\text{id}}}{\sqrt{N_i M_j}} + \mu_1(\boldsymbol{x}_{\text{id},i})\big(\mu_1(\boldsymbol{x}_{\text{ag},j})\big)^T \tag{8}$$

$$\mu_2(\boldsymbol{x}_{\text{ag},j}, \boldsymbol{x}_{\text{id},i}) = -\frac{\boldsymbol{U}_{\text{id}}^T\Sigma^{-1}\boldsymbol{U}_{\text{ag}}}{\sqrt{N_i M_j}} + \mu_1(\boldsymbol{x}_{\text{ag},j})\big(\mu_1(\boldsymbol{x}_{\text{id},i})\big)^T \tag{9}$$

where $\Sigma = \sigma^2\boldsymbol{I} + \boldsymbol{U}_{\text{id}}\boldsymbol{U}_{\text{id}}^T + \boldsymbol{U}_{\text{ag}}\boldsymbol{U}_{\text{id}}^T$, N_i and M_j are the numbers of training samples for the ith subject and the kth age group, respectively.

M Step. We maximize the Q function to estimate the $\theta^{(i+1)}$. The maximization is shown as follows:

$$\theta^{(i+1)} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{(i)}) \tag{10}$$

To solve this optimization, the model parameter $\theta^{(i+1)}$ is given by

$$\begin{aligned}\boldsymbol{U}_{\text{id}} &= (\boldsymbol{C} - \boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{E})(\boldsymbol{A} - \boldsymbol{F}\boldsymbol{B}^{-1}\boldsymbol{E})^{-1} \\ \boldsymbol{U}_{\text{ag}} &= (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{F})(\boldsymbol{B} - \boldsymbol{E}\boldsymbol{A}^{-1}\boldsymbol{F})^{-1} \\ \sigma^2 &= \frac{1}{Nn}\sum_{i,j}\big\{(\boldsymbol{v}_i^j - \bar{\boldsymbol{v}} - \boldsymbol{U}_{\text{id}}\mu_1(\boldsymbol{x}_{\text{id},i}) \\ &\qquad - \boldsymbol{U}_{\text{ag}}\mu_1(\boldsymbol{x}_{\text{ag},j}))^T(\boldsymbol{v}_i^j - \bar{\boldsymbol{v}})\big\}\end{aligned} \tag{11}$$

in which

$$\begin{aligned}&\boldsymbol{A} = \sum_{ij}\mu_2(\boldsymbol{x}_{\text{id},i}, \boldsymbol{x}_{\text{id},i}), \boldsymbol{B} = \sum_{i,j}\mu_2(\boldsymbol{x}_{\text{ag},j}, \boldsymbol{x}_{\text{ag},j}), \\ &\boldsymbol{C} = \sum_{ij}(\boldsymbol{v}_i^j - \bar{\boldsymbol{v}})\big(\mu_1(\boldsymbol{x}_{\text{id},i})\big)^T, \boldsymbol{D} = \sum_{ij}(\boldsymbol{v}_i^j - \bar{\boldsymbol{v}})\big(\mu_1(\boldsymbol{x}_{\text{ag},j})\big)^T, \\ &\boldsymbol{E} = \sum_{ij}\mu_2(\boldsymbol{x}_{\text{ag},j}, \boldsymbol{x}_{\text{id},i}), \boldsymbol{F} = \sum_{i,j}\mu_2(\boldsymbol{x}_{\text{id},i}, \boldsymbol{x}_{\text{ag},j}).\end{aligned} \tag{12}$$

After the model parameters and the posterior distribution are estimated by the EM algorithm, the identity factor for the ith subject can be inferred by the first moment of $\boldsymbol{x}_{\text{id},i}$, namely $\mu_1(\boldsymbol{x}_{\text{id},i})$. One can observe that the form of $\mu_1(\boldsymbol{x}_{\text{id},i})$ is actually identical to an LF-FC layer with the following parameters:

$$\boldsymbol{W} = \boldsymbol{U}_{\text{id}}^T\Sigma^{-1},\ \boldsymbol{b} = -\boldsymbol{U}_{\text{id}}^T\Sigma^{-1}\bar{\boldsymbol{v}} \tag{13}$$

where $\bar{\boldsymbol{v}} = \frac{1}{N}\sum_{i,j}\boldsymbol{v}_i^j$ and $\Sigma = \sigma^2\boldsymbol{I} + \boldsymbol{U}_{\text{id}}\boldsymbol{U}_{\text{id}}^T + \boldsymbol{U}_{\text{ag}}\boldsymbol{U}_{\text{id}}^T$. The feature $\boldsymbol{v}$ comes from the convolution unit. Because $\boldsymbol{v}$ is updated with new parameters of the identity factor guided FC layer, $\boldsymbol{W}, \boldsymbol{b}$ are also updated using the new $\boldsymbol{v}$. The procedure iteratively goes on. The detailed learning framework is elaborated in the next subsection.

Algorithm 1 Coupled Learning Algorithm for LF-CNNs

Input: Outside training data $\boldsymbol{Y}_{\text{i}}$ with identity label, cross-age training data $\boldsymbol{Y}_{\text{ia}}$ with both age and identity label.

Output: The parameters θ_f and θ_g.

1: $i \leftarrow 0$.
2: Initialize the parameters $\theta_f^{(1)}$ by Xaiver filter.
3: Initialize $\theta_g^{(1)} = \{\boldsymbol{U}_{\text{id}}, \boldsymbol{U}_{\text{ag}}, \sigma^2, \bar{v}\}$ where $\sigma^2 = 0.1, \bar{v} = \boldsymbol{0}$ and $\boldsymbol{U}_{\text{id}}, \boldsymbol{U}_{\text{ag}}$ are randomly initialized from -0.1 to 0.1.
4: Compute $\boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}$ for the LF-FC layer via Eq. (13).
5: **while** not converge **do**
6: $i \leftarrow i + 1$.
7: Fix the $\theta_g^{(i)}$, and train the LF-CNNs with the outside data $\boldsymbol{Y}_{\text{i}}$ to update $\theta_f^{(i+1)}$ from $\theta_f^{(i)}$.
8: Fix the $\theta_f^{(i+1)}$, and input the training data $\boldsymbol{Y}_{\text{ia}}$ to obtain the convolutional features $\boldsymbol{F}_{conv}$, which are taken as the observed features $\boldsymbol{V}$.
9: Update the parameters $\theta_g^{(i+1)}$ from $\theta_g^{(i)}$ via Eq. (3).
10: Compute the parameters $\boldsymbol{W}^{(i+1)}, \boldsymbol{b}^{(i+1)}$ for the LF-FC layer via the Eq. (13).
11: **end while**

3.3. Learning Framework

In LF-CNN model, the convolution unit maps a raw input image $\boldsymbol{F}_{img}$ to convolutional feature $\boldsymbol{F}_{conv}$ by $\boldsymbol{F}_{conv} = f(\boldsymbol{F}_{img})$, and then the LF-FC layer computes the age-invariant feature $\boldsymbol{F}_{fc}$ via $\boldsymbol{F}_{fc} = g(\boldsymbol{F}_{conv})$. The age-invariant feature $\boldsymbol{F}_{fc}$ is used for the AIFR. The parameters in the convolution unit and the LF-FC layer are denoted by θ_f, θ_g respectively. In our framework, $f(\cdot)$ and $g(\cdot)$ characterize different properties of AIFR and the original learning methods based on CNNs are not appropriate, we adopt a coupled learning framework to optimize the LF-CNN model. Concretely, after initializing the networks, the parameters θ_f of convolutional unit is updated by SGD with the fixed θ_g. Then we fix θ_f and use the LIA model to learn the parameters θ_g for the LF-FC layer. We alternatively update θ_f and θ_g until the stopping condition is satisfied. The procedure is summarized in Algorithm 1.

Learning parameters for LF-FC layer. We use the training data $\boldsymbol{Y}_{\text{ia}}$ with both age and identity label to train the LF-FC layer. Specifically, the convolutional feature $\boldsymbol{F}_{fc}$ is taken as the observed features $\boldsymbol{V}$ in LIA. The LIA model learns the parameters $\theta_g = \{\boldsymbol{U}_{\text{id}}, \boldsymbol{U}_{\text{ag}}, \sigma^2, \bar{v}\}$, and then computes $\boldsymbol{W}, \boldsymbol{b}$ for the LF-FC layer via Eq. (13).

Learning parameters for convolution unit. To learn the parameters θ_f for the convolution unit, we need to fix the parameters θ_g of the LF-FC layer. Then we use the SGD to train the LF-CNNs with outside training data $\boldsymbol{Y}_{\text{i}}$. Note that, we use both the softmax loss and the contrastive loss to strength the supervision in learning, similar to [31].

3.4. Discussion

The learning process of LF-CNNs has the following advantages. First, the coupled learning is very beneficial to AIFR. The joint objective functions consists of minimizing the classification error (softmax loss and contrastive loss) and maximizing the likelihood probability that the training samples are generated by the latent factors. The former aims to learn discriminant feature representations for classification while the later improves the robustness of age-invariant features. Both of them consistently contribute to the gain in the AIFR task. Second, we update the LF-FC layer by LIA instead of SGD, which largely reduces the parameter scale and prevents potential overfitting. LIA plays an essential role in LF-CNNs by inferring the effective identity factor to guide the parameter estimation of the LF-FC layer.

4. Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of the proposed LF-CNNs on several challenging face aging databases, including MORPH Album 2 [27] (the largest face aging database available in the pubic domain), FG-NET [1] (a public-domain face aging dataset), and the subset of Cross-Age Celebrity Dataset (CACD-VS) [4]. To further demonstrate the generalization of our model, we also evaluate our model on the Labeled Faces in the Wild (LFW) database [11].

4.1. Implementation Details

Preprocessing. For each face image, we use the recently proposed algorithms [2, 37, 3] to detect the facial landmarks in images. 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.

Training Data. The training data used in this paper is composed of two parts: outside training data $\boldsymbol{Y}_{\text{i}}$ (that only contains the identity information) and face aging training data $\boldsymbol{Y}$ (that contains both the age and identity information). For the outside training data $\boldsymbol{Y}_{\text{i}}$, we use the large scale web-collected face data, including CASIA-WebFace [34], CACD [4], and Celebrity+ [21]). When testing our model on CACD-VS, we remove all the identities in CACD from the training data. The images are horizontally flipped for data augmentation. For the face aging training data $\boldsymbol{Y}_{\text{ia}}$, we use the MORPH Album 2 dataset, as described in Section 4.2.

Detailed setting in LF-CNN model. We implement the LF-CNN model using the Caffe library [12] with our modifications. Unless otherwise specified, the batch size is 150. When training convolution unit with outside data, the learning rate is 1e-1, 1e-2, 1e-3 and is switched when the error plateaus. The total number of epochs is about 12 for our model.

Classifier. To better evaluate the performance of the proposed age-invariant deep face features, our model uses the simple Euclidean Distance and the Nearest Neighbor rule as the classifier.

4.2. Experiments on the MORPH Album 2 Dataset

The MORPH Album 2 database is the largest face aging dataset available in the public domain, consisting of about 78,000 face images of 20,000 persons with age ranging from 16 to 77. Following the same training and testing split scheme in [7], 10,000 subjects are used for training and the remaining 10,000 subjects are used for testing. There is no overlapping subject between the training set and the testing set. For each subject, two face images with the youngest age and the oldest age are selected as gallery and probe set respectively. For fair comparison, we also train a baseline CNN model with the same networks as LF-CNNs, and learn all the parameters by SGD. The experimental results are shown in Table 1.

Method	Rank-1 Identification Rates
HFA (2013) [7]	91.14%
CARC (2014) [4]	92.80%
MEFA (2015) [8]	93.80%
MEFA+SIFT+MLBP (2015) [8]	94.59%
Method (2015) in [16]	87.13%
LPS+HFA (2016) [17]	94.87%
CNN-baseline	89.68%
CNN-baseline (fine-tuned by MORPH training data)	95.13%
LF-CNNs (fine-tuned by MORPH training data)	**97.51%**

Table 1. Performance of different methods on MORPH.

In Table 1, we compare our LF-CNN model against (i) the CNN-baseline model, (ii) the CNN-baseline model (fine-tuned by MORPH training data), and (iii) several recently developed top-performing AIFR algorithms in the literature. From these results, we have the following observations. First, the result of the CNN-baseline is only 89.68%, which is inferior to the other results in Table 1. This confirms that directly applying the deep CNN model to address the AIFR problem is indeed not a good choice. Second, the performance of CNN-baseline can be improved to 95.13% by fine-tuning with the additional MORPH training data. However, this result (95.13%) is near to the top-performing result in the literature (94.59%). The lack of a significant improvement over the state-of-the-art reflects the limitation of the CNN-baseline model. So it is desirable to design a new deep CNN model to address the AIFR problem. Finally, it is encouraging to see that the proposed LF-CNN model obtains a significant performance improvement over the other results in Table 1, demonstrating a new state-of-the-art (97.51%) on the MORPH Album 2 database.

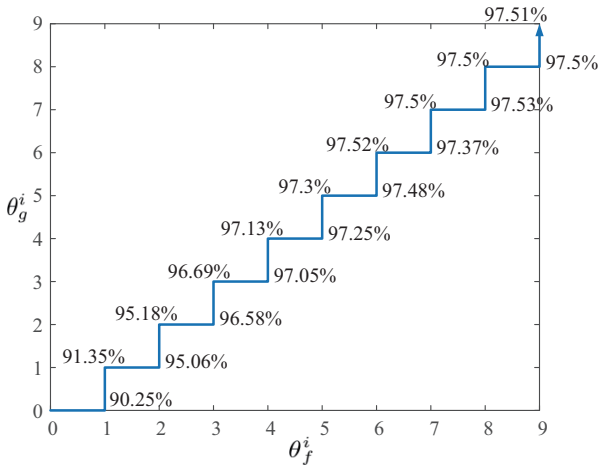

Figure 4. The recognition rates in each iteration of the coupled learning algorithm.

To further evaluate the performance of our LF-CNN model, we design an experiment to report the recognition results of our model in each iteration of the coupled learning process, as illustrated in Figure 4. The two parameters θ_f and θ_g are updated in a step-wise manner. With one fixed, update the other one, and vice versa. Figure 4 clearly shows that the coupled learning process consistently contributes to the performance improvement of AIFR, converging to a good result quickly.

Figure 5 shows some examples of the failed retrievals using our approach in MORPH Album 2 dataset. We can see that although our results are incorrect in these cases, the probe images appear to be more similar to the incorrect retrievals than the gallery images.

4.3. Experiments on the FG-NET Dataset

The FG-NET dataset consists of 1002 face images from 82 different subjects, with each subject having multiple face images at different ages (ranging from 0 to 69). Following the testing scheme in [18], we compare our LF-CNNs with the state-of-the-art approaches on this dataset. The comparative results are reported in Table 2.

Method	Rank-1 Identification Rates
Park et al. (2010) [24]	37.4%
Li et al. (2011) [18]	47.5%
HFA (2013) [7]	69.0%
MEFA (2015) [8]	76.2%
CNN-baseline	84.4%
LF-CNNs	**88.1%**

Table 2. Performance of different methods on FG-NET.

As can be seen from Table 2, LF-CNNs achieve the highest recognition accuracy (88.1%) among all the results, sig-

Figure 5. Some failed retrievals in MORPH Album 2. The first row: the probe images. The second row: the incorrect retrievals using our approach. The third row: the corresponding gallery images for the probe images.

nificantly outperforming the top-performing result (76.2%) in [8] by 11.9%. Moreover, the proposed LF-CNNs outperform the CNN-baseline method by a clear margin. This confirms what we observe from the MORPH Album 2 dataset. Interestingly, MORPH and FG-NET have different age distributions. In FGNET, roughly 61% samples are less than 16 years old. But for the MORPH dataset, all the persons are more than 16 years old. So it is desirable to exploit the influence of different age distributions on the proposed approach. In Table 3 we give the rank-1 identification rates in different age groups.

Age group	Amount	CNN-baseline	LF-CNNs
0 - 4	193	51.81%	60.10%
5 - 10	218	84.86%	88.53%
11 - 16	201	91.04%	94.03%
17 - 24	182	94.51%	97.80%
25 - 69	208	99.04%	99.52%
0 - 16	**612**	**76.47%**	**81.37%**
17 - 69	**390**	**96.93%**	**98.72%**

Table 3. Performance of different age groups on FG-NET.

The results in Table 3 show that the proposed LF-CNNs consistently outperforms the CNN-baseline model on all the age groups. This further confirms the advantage of our LF-CNN model over the CNN-baseline model in AIFR task.

4.4. Experiments on the CACD Verification Subset

The CACD dataset is a recently released dataset for AIFR, 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. Following the state-of-the-art configuration [4], we test LF-CNNs on a subset of CACD [4], CACD-VS, which consists of 4000 image pairs (2000 positive pairs and 2000 negative pairs) and have been carefully annotated. We follow the same training strategy as in section 4.3. Note that, the identities in CACD-VS are excluded from the outside training data in this experiment. So only 400,000 training samples are used. According to the ten-fold cross-validation rule, we calculate the Euclidean Distance of each pairs and choose the best threshold using nine training folds, then testing on the leftover fold. We compute the face verification rate and compare our result with the existing methods in this dataset, as shown in Figure 6 and Table 4.

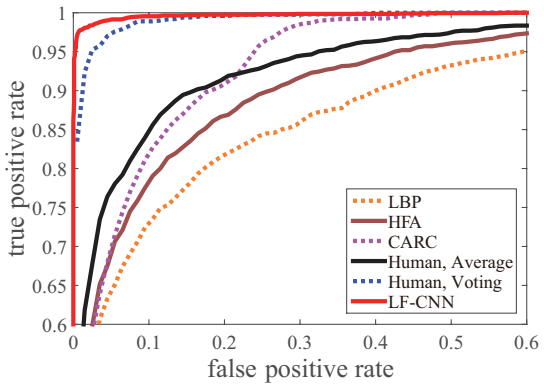

Figure 6. ROC comparisons of different methods on CACD-VS.

Again, the proposed LF-CNN model significantly outperforms all the published results in this dataset, even surpassing the human-level performance with a clear margin. It further demonstrates the effectiveness of the proposed age-invariant deep features.

Method	Acc.
High-Dimensional LBP (2013) [5]	81.6%
HFA (2013) [7]	84.4%
CARC (2014) [4]	87.6%
Human, Average (2015)	85.7%
Human, Voting (2015)	94.2%
LF-CNNs	**98.5%**

Table 4. Performance of different methods on CACD-VS.

4.5. Experiments on the LFW Dataset

To evaluate the generalization performance of LF-CNNs, we further conduct an experiment on the famous LFW dataset [11]. This dataset contains 13,233 face images from 5749 different subjects, collecting from uncontrolled conditions. Following the unrestricted with labeled outside data protocol [11], we train on the outside dataset and test on 6,000 face pairs. People overlapping between the outside training data and the LFW testing data are excluded. We respectively train 25 networks with 25 different image patches, and concatenate the output features from these networks into a long feature vector. We then apply PCA on the long feature vector to obtain a compact feature vector for classification. In Table 5 we compare our results against the recent

state-of-the-art results. From the results we can see that our approach can obtain comparable results to the state-of-the-art approaches using relatively small training data, demonstrating the excellent generalization ability of our approach.

Method	Images	Networks	Acc.
DeepFace (2014) [33]	4M	3	97.35%
DeepID-2+ (2015) [32]	-	25	99.47%
FaceNet (2015) [29]	200M	1	99.65%
Deep Embedding (2015) [20]	1.2M	10	99.77%
Deep FR (2015) [25]	2M	1	98.95%
LF-CNNs (single)	**700K**	**1**	**99.10%**
LF-CNNs (ensemble)	**700K**	**25**	**99.50%**

Table 5. Performance of different methods on LFW.

5. Conclusions

In this paper, we have proposed an age-invariant deep face recognition framework, referred to as LF-CNNs. Unlike the existing deep learning models in face recognition community, the proposed new model constructs a latent identity analysis (LIA) module to guide the learning of the CNNs parameters. By coupled learning the CNNs parameters and the LIA parameters, our model can extract the age-invariant deep face features, which are well suitable for the AIFR task. Extensive experiments are conducted on several public-domain face aging databases to demonstrate the significant performance improvement of our new model over the state-of-the-art. We have also performed experiments on the famous LFW dataset to demonstrate the excellent generalization ability of our new model.

6. Acknowledge

This work was supported by grants from Natural Science Foundation of Guangdong Province (2014A030313688), Shenzhen Research Program (JSGG20150925164740726, JCYJ20150925163005055, and CXZZ20150930104115529), Guangdong Research Program (2014B050505017 and 2015B010129013), the Key Laboratory of Human-Machine Intelligence-Synergy Systems through the Chinese Academy of Sciences, and National Natural Science Foundation of China (61103164).

References

[1] Fg-net aging database. In *http://www.fgnet.rsunit.com/*, 2010.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1859–1866. IEEE, 2014.

[4] B.-C. Chen, C.-S. Chen, and W. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE TMM*, 17(6):804–815, 2015.

[5] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032. IEEE, 2013.

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

[7] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *ICCV*, 2013.

[8] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li. A maximum entropy feature descriptor for age invariant face recognition. In *CVPR*, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[10] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *IJCAI*, 1999.

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

[12] 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.

[13] B. Klare and A. K. Jain. Face recognition across time lapse: On learning feature subspaces. In *IJCB*, 2011.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

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

[16] Z. Li, D. Gong, X. Li, and D. Tao. Learning compact feature descriptor and adaptive matching framework for face recognition. *Image Processing, IEEE Transactions on*, 24(9):2736–2745, 2015.

[17] Z. Li, D. Gong, X. Li, and D. Tao. Aging face recognition: A hierarchical learning model based on local patterns selection. *Image Processing, IEEE Transactions on*, 25(5):2146 – 2154, 2016.

[18] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *Information Forensics and Security, IEEE Transactions on*, 6(3):1028–1037, 2011.

[19] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs. Face verification across age progression using discriminative methods. *Information Forensics and Security, IEEE Transactions on*, 5(1):82–91, 2010.

[20] J. Liu, Y. Deng, B. Tao, W. Zhengping, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *arXiv preprint arXiv:1411.7766*, 2014.

[22] C. Otto, H. Han, and A. Jain. How does aging affect facial components? In *ECCV*, 2012.

[23] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In *CVPR*, 2015.

[24] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE TPAMI*, 32(5):947–954, 2010.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *Proceedings of the British Machine Vision*, 2015.

[26] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009.

[27] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FG*, 2006.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 1–42, 2014.

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

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

[31] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.

[32] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015.

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

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

[35] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.

[36] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015.

[37] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.