Pattern Recognition 66 (2017) 82–94

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Group-aware deep feature learning for facial age estimation

Hao Liu[a,b], Jiwen Lu[a,b,*], Jianjiang Feng[a,b], Jie Zhou[a,b]

[a] Department of Automation, Tsinghua University, Beijing 100084, China
[b] Tsinghua National Laboratory for Information Science and Technology, Beijing 10084, China

ARTICLE INFO

Keywords:
Facial age estimation
Deep learning
Feature learning
Biometrics

ABSTRACT

In this paper, we propose a group-aware deep feature learning (GA-DFL) approach for facial age estimation. Unlike most existing methods which utilize hand-crafted descriptors for face representation, our GA-DFL method learns a discriminative feature descriptor per image directly from raw pixels for face representation under the deep convolutional neural networks framework. Motivated by the fact that age labels are chronologically correlated and the facial aging datasets are usually lack of labeled data for each person in a long range of ages, we split ordinal ages into a set of discrete groups and learn deep feature transformations across age groups to project each face pair into the new feature space, where the intra-group variances of positive face pairs from the training set are minimized and the inter-group variances of negative face pairs are maximized, simultaneously. Moreover, we employ an overlapped coupled learning method to exploit the smoothness for adjacent age groups. To further enhance the discriminative capacity of face representation, we design a multi-path CNN approach to integrate the complementary information from multi-scale perspectives. Experimental results show that our approach achieves very competitive performance compared with most state-of-the-arts on three public face aging datasets that were captured under both controlled and uncontrolled environments.

1. Introduction

Facial age estimation attempts to predict the real age value or age group based on facial images, which has widely potential applications such as facial bio-metrics, human-computer interaction, social media analysis and entertainments [1–4]. While extensive efforts have been devoted, facial age estimation still remains a challenging problem due to two aspects: 1) lack of sufficient training data where each person should contain multiple images in a wide range of ages, 2) large variations such as lighting, occlusion and cluttered background of face images which were usually captured in wild conditions.

Most existing facial age estimation systems usually consist of two key modules: face representation and age estimation. Representative face representation approaches include holistic subspace features [5,6], active appearance model (AAM) [7], Gabor wavelets [7], local binary pattern (LBP) [8] and bio-inspired feature (BIF) [9]. Having obtained face representations, age estimation can be addressed as a classification or regression problem [9–11]. However, the face representations employed most existing methods are hand-crafted, which requires strong prior knowledge to engineer it by hand. To address this problem, learning-based feature representation methods [5,12,13,3] have been made to learn discriminative feature representation directly from raw pixels. However, these methods aim to learn linear feature filters to project face images into another feature space such that they may not be powerful enough to exploit the nonlinear relationship of data. To address this nonlinear issue, deep learning-based methods have been adopted to learn a series of nonlinear mapping functions between face image and age label [14–16,16–18]. Unfortunately, these deep models cannot explicitly achieve the ordinal relationship among the chronological ages, which are still far from the practical satisfactory in most cases because they usually encounter unbalanced and insufficient training data for each age label.

Notice that age labels are chronologically correlated, so that it is desirable to employ nonlinear discriminative methods to exploit the correlated order information from facing images. Unlike existing deep learning-based facial age estimation methods that ignored the ordinal information of face aging data, we proposed a group-aware deep feature learning method (GA-DFL) under deep convolutional neural networks (CNN), by learning discriminative face representations directly from image pixels and exploiting the aging order information. Since facial aging datasets usually lack of face images from the same person covering a wide range of ages, our proposed GA-DFL first separates the chronological aging progress into several overlapped groups and then learns a series of hierarchical nonlinear mapping

* Corresponding author at: Department of Automation, Tsinghua University, Beijing 100084, China.
E-mail address: elujiwen@gmail.com (J. Lu).

http://dx.doi.org/10.1016/j.patcog.2016.10.026
Received 13 July 2016; Received in revised form 29 September 2016; Accepted 20 October 2016
Available online 26 October 2016

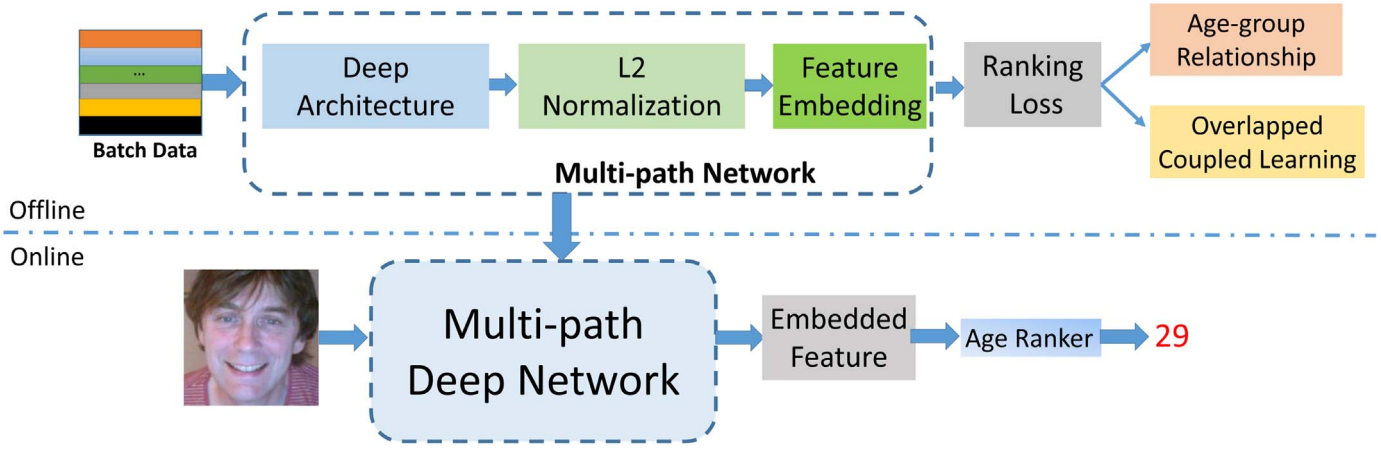

Fig. 1. The pipeline of the proposed facial age estimation approach. During the offline phase, we enforce two criterions on modeling aging progress to learn face representation: 1) the inter-group variances are maximized while the intra-group variances are minimized. 2) the separated age groups should be smoothed on the age-group specific overlaps. Then the parameters of the designed network are optimized by back-propagation. During online phase, we feed face image into the network to obtain face representation and the final age label is performed by an age ranker.

functions to project raw pixel values into another common feature space, so that face pairs in the same age groups are projected as close as possible while those in different age groups are projected as far as possible. Moreover, we link every discrete groups by overlapping structures and develop an overlapped coupled learning method, which aims to smooth the age differences lying on the overlaps of the adjacent age groups. We also propose a multi-path CNN architecture to enhance the capacity of feature representation to integrate complementary information from multiple scales to improve the performance. Fig. 1 illustrates the main procedure of our proposed approach. To evaluate the effectiveness of our proposed GA-DFL, we conducted experiments on three widely used facial age estimation datasets that were captured in both constrained and unconstrained environments. Experimental results show that our proposed GA-DFL obtains superior performance compared with most state-of-the-art facial age estimation methods.

The contributions of this work are summarized as follows:

(1) We develop a deep feature learning method to discriminatively learn a face representation directly from raw pixels. With the learned nonlinear filters, the chronological age information can be well exploited with a perspective of age groups in the obtained face descriptor.
(2) We propose an overlapped coupled learning method to achieve the smoothness on the neighboring age groups. With this learning strategy, the age difference information on the age-group specific overlaps can be well measured.
(3) We employ a multi-path deep CNN architecture to integrate multiple scale information into the learned face presentation.

The rest of this paper is organized as follows: Section 2 reviews some backgrounds. Section 3 details the proposed GA-DFL method. Section 4 provides the experimental results and Section 5 concludes this paper.

2. Related work

In this section, we briefly review two related topics: facial age estimation and deep learning.

2.1. Facial age estimation

Numerous facial age estimation methods [19–22,12,23–27] have been proposed over the past two decades. As one of the earliest studies, Lanitis et al. [25] applied a quadratic function to predict facial age. Thereafter, several works [28,23] were proposed to incorporated with correlated age labels to model practical human aging progress with different degrees of improvements. In particular, Chang et al. [28] presented an ordinal hyperplane ranking method (OHRank) which divided age classification as a series of sub-problems of binary classification. Geng et al. [23] proposed a label distribution learning approach to model the relationship between face images and age labels. Besides, Guo and Mu [29] showed human gender and race are used to exploit the complementary information for age estimation. However, most of these methods utilize hand-crafted features, which require strong prior knowledge by-hand and usually encounters time-consuming. To address this, several studies have been made to learn a discriminative face representation by using advanced feature learning approaches [30,24,3]. For example, Guo et al. [30] proposed a holistic feature learning approach utilizing manifold learning technique. Lu et al. [3] employed a local binary feature learning method to learn a face descriptor robust to local illumination, which has achieved considerable performances on facial age estimation. Nevertheless, these methods focus on learning linear filters so that they are not powerful enough to describe the age-informative facial appearances because there are large variances on collected face data due to scaling, occlusion and cluttered background especially captured in wild conditions. In contrast to these previous works, we propose a deep learning method from a perspective of feature learning with a feed-forward neural networks to exploit the nonlinear relationship of data.

2.2. Deep learning

In the recent literature, deep learning has received much attention in the research field of machine learning and computer vision due to its superior performance in learning a series of nonlinear feature mapping functions directly from raw pixels. A number of deep learning approaches such as restricted Boltzmann machine (RBM) [31], stacked denoising auto-encoder (SDAE) [32], deep convolutional neural networks (CNN) [33] have been successfully employed in many visual analysis tasks such as handwritten digit recognition [34], object detection [35], visual tracking [36] and scene labeling [37]. More recently, deep learning methods have been applied to face analysis tasks including face detection [38], face alignment [39] and face recognition [40,41]. Specifically, Zhang et al. [39] proposed a deep learning method with stacked auto-encoder networks to estimate facial landmarks in a coarse-to-fine manner, Sun et al. [40] developed DeepID2 network to reduce the personalized inter-covariance joint by identification and verification, and Parkhi et al. [41] employed a very deep architecture VGG-16 Face Net pre-trained by a large scale face dataset to perform face recognition.

Inspired by the aforementioned works which learns task-adaptive face representation, deep learning has been also used to learn a set of nonlinear feature transformations for facial age estimation [42,43,14–16,44]. For example, Yi et al. [45] employed a multi-scale CNN to predict the age value with additional gender and ethnicity information. Levi et al. [46] jointly conducted age estimation and gender classification with CNN. Yang et al. [47] deployed a deep scattering network to predict facial age via category-wise rankers. However, these deep learning-based models usually require a very large face aging dataset to learn offline feature representation in most cases such that they usually suffer from insufficient face data because densely collecting face images in a wide range of ages is difficult and impractical. To address this problem, Liu et al. [48] and Yang et al. [16] employed their defined loss functions on the top layer of pretrained deep model. While significant performances have been obtained, these deep models ignore the ordinal relationship of age labels. In this work, we present a group-aware deep feature learning approach, which learns discriminative face representation for facial age estimation and exploits the aging order information simultaneously.

3. Proposed approach

In this section, we present the proposed model GA-DFL and multi-path network architecture.

3.1. Model

Modeling real-world age progress requires sufficient facial data for the same person covering a widely range of age labels. However, densely collecting abundant face images per person is difficult and even impractical because face aging data encounters missing label problem. Fortunately, face images in short-interval (e.g. age labels covers smaller than 10 years old) are usually available on existing face aging datasets. To address this, recent study [4] has proposed a face aging method, which splits the aging progress into a set of groups and then learned the corresponding dictionaries to characterize the aging patterns for different age groups. However, their goal is to reconstruct aging face sequence by learning a set of age-group specific dictionaries, which cannot be directly applied to our age estimation problem. Intuitively, facial images with neighboring age values are generally similar to each other. For example, the appearance of a person of 50 years old is more similar to those of 47–52 years old than those bellow 30 years old. Motivated by this fact, we split the total age process into several discrete groups and the age-group specific relationships are exploited as follows: 1) the distance between face pairs from the similar group should be minimized, 2) the distance between face pairs that come from different groups should be maximized as far as possible. To achieve this goal, we propose a group-aware deep feature learning (GA-DFL) method to learn a new appearance space, where face pairs within the similar group are as close as possible and those from different age groups are pushed as far as possible simultaneously.

3.1.1. Group-aware formulation

Let $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the training set which contains N samples, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes i-th face image of d pixels. We first construct age groups on the wide range of age labels. To be specific, we divide the age progress into G groups according to the age values, where each age group consists of α age labels (e.g. α is assigned to 10 and thus each group distributes as 0–10, 11–20, 21–30, etc.). Sequentially, we compute feature representation $f(\mathbf{x}_i)$ for each face image $\mathbf{x}_i$ based on VGG-16 architecture [41]. As is illustrated in Fig. 2, our network architecture consists of M layers including convolution, ReLU non-linearity, pooling and fully connected layers. We firstly feed the face image to the convolutional network and obtain the immediate feature representation as:

$$f(\mathbf{x}_i) = \mathbf{h}_i^{(m)} = \mathrm{pool}(\mathrm{ReLU}(\mathbf{W}^{(m)} \otimes \mathbf{x}_i + \mathbf{b}^{(m)})) \tag{1}$$

where $\mathrm{pool}(\cdot)$ denotes the max pooling operation, $\mathrm{ReLU}(\cdot)$ denotes the nonlinear ReLU function and $m = \{1, 2, \ldots, M-2\}$.

To exploit our defined group-aware relationship, we define two-layer deep neural network, where the output of the most top layer can be computed as:

$$f(\mathbf{x}_i) = \mathbf{h}_i^{(M)} = \delta(\mathbf{W}^{(M)}\mathbf{x}_i + \mathbf{b}^{(M)}) \tag{2}$$

where $\mathbf{W}^{(M)}$ and $\mathbf{b}^{(M)}$ denote the weights and bias of the top layer, respectively, $\delta(\cdot)$ is the nonlinear function *tanh* function in fully connected layers. To sum up the total weights, we collect $m = \{1, 2, \ldots, M\}$ to train the whole deep neural networks in a globally tuned manner.

In this paper, our goal is to learn a face representation based on the similarity or dissimilarity of face pairs of training set. Given each face pair $\mathbf{x}_i$ and $\mathbf{x}_j$, they can be represented as $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ at the top layer of the designed network, and the Euclidean distance between the face pair $\mathbf{x}_i$ and $\mathbf{x}_j$ can be computed as:

$$d_f^2(\mathbf{x}_i, \mathbf{x}_j) = \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2. \tag{3}$$

In order to preserve the geometric structure of face data, we enforce the marginal fisher analysis criterion [49] on the outputs of training points in the nearest neighbour space at the most top level of our deep architecture. Hence, we formulate our goal to minimize the following objective function:

$$\min_f = J_1 - \lambda_1 J_2 + \lambda_2 J_3 = \frac{1}{2Nk_1}\sum_{i}^{N}\sum_{j}^{N}\sum_{g}^{G} Q_g^{ij} d_f^2(\mathbf{x}_i^g, \mathbf{x}_j^g) - \lambda_1 \frac{1}{2Nk_2}\sum_{i}^{N}\sum_{j}^{N}\sum_{g_1 \neq g_2}^{G} S_{g_1,g_2}^{ij} d_f^2(\mathbf{x}_i^{g_1}, \mathbf{x}_j^{g_2}) + \lambda_2 \frac{1}{2}\sum_{m}^{M}(\|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2), \tag{4}$$

where J_1 denotes the intra-group compactness which aims to minimize the distance of pairs in the same group, J_2 denotes the inter-group separability which enforces the distance of pairs coming from different groups, λ_1 is employed to balance term J_1 and term J_2, k_1 and k_2 are parameters to define the size of the intra-group and inter-group nearest neighbours, $\|\mathbf{W}^{(m)}\|_F^2$ denotes the Frobenius norm of matrix $\mathbf{W}^{(m)}$, Q_g^{ij} and S_{g_1,g_2}^{ij} are affinity matrices to measure the similarity of pairs, which are defined as follows:

$$Q_g^{ij} = \begin{cases} 1, & \text{if} \quad \mathbf{x}_i^g \in \mathcal{N}_{k_1}(\mathbf{x}_j^g) \quad \text{or} \quad \mathbf{x}_j^g \in \mathcal{N}_{k_1}(\mathbf{x}_i^g). \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

$$S_{g_1,g_2}^{ij} = \begin{cases} 1, & \text{if} \quad \mathbf{x}_i^{g_1} \in \mathcal{N}_{k_2}(\mathbf{x}_j^{g_2}) \quad \text{or} \quad \mathbf{x}_j^{g_2} \in \mathcal{N}_{k_2}(\mathbf{x}_i^{g_1}) \\ & \text{and} \quad g_1 \neq g_2. \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where g and $\mathcal{N}_k$ denote the label of the defined age group and the k nearest neighbours of $\mathbf{x}_i$.

While (4) addresses the group-wise relationship, the smoothness between the neighboring age groups cannot be implicitly modeled under the framework of separated age groups, which can degrade the estimation performance, especially for the images which locate near the boundary of age group. To address this issue, we present an overlapping coupled learning method to exploit the age difference information on the overlapping range of adjacent age groups (Fig. 3).

3.1.2. Overlapping coupled learning

Since age labels are continuous in chronological order so that the defined discrete age groups should be smoothed across adjacent age groups, illustrated as Fig. 3, we construct overlapping structures, which cover across every neighboring age groups with the stride of κ.

Assuming that we have a face pair denoted by x_{p1} and x_{p2} from a coupled overlapping regions O^g covering both the age groups g and

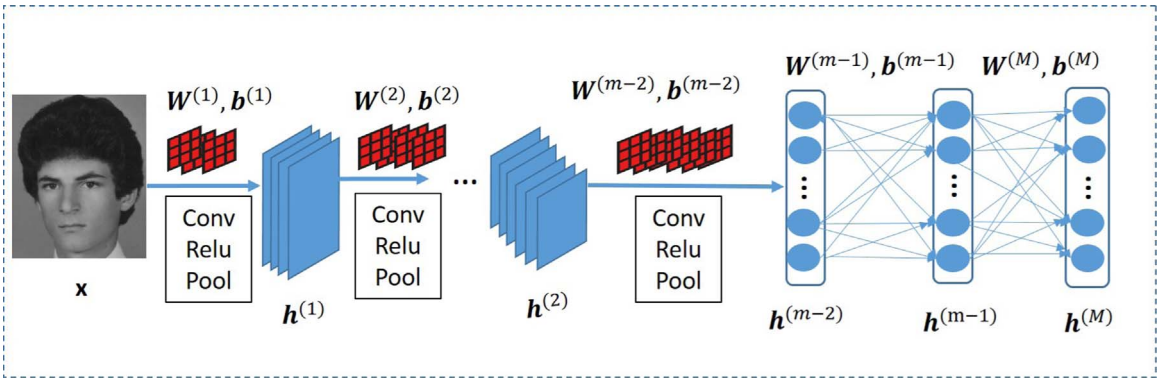

Fig. 2. The network architecture. Our network consists of a series of convolution layers in size of 3×3, ReLU layers and max pooling layers, summing up to M layers, where the model parameters consists of $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^{M-2}$. Following the convolutional layers, we discard the fully connected layers employed in VGG and then added two additional two-layer neural networks consisting fully connections parameterized by $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=M-1}^{M}$ and *tanh* nonlinear functions. We jointly learn the model parameters by back-propagation.

$g+1$, our goal is to dynamically treat the similarities among face pairs according to the age gap, which can be formulated to minimize the following objective function:

$$\sum_{p}^{P}\sum_{g}^{G} \ell_{p1,p2}^{O_g}(\tau - d_f^2(\mathbf{x}_{p1}, \mathbf{x}_{p2})\cdot\omega(y_1, y_2)), \tag{7}$$

where $\ell_{p1,p2}^{O_g}$ denotes the indicator that is assigned to 1 in which x_{p1} and x_{p2} are from the same overlapping region O^g, and assigned to 0 in other cases (as shown in Fig. 4). y_1 and y_2 denote the age label of the face pair, τ is the corresponding margins (basically, assigned to 1), respectively. $\omega(y_1, y_2)$ is the age-sensitive weighting function, where the distance of the face pair is weighted according to the age-related gap. For example, the similarity for a face pair with a small age gap should be weighted smaller than that for a large age gap. We apply the Gaussian function as the weighting function, which is defined as follows:

$$\omega(y_1, y_2) = \begin{cases} 1 - \exp^{\frac{-(y_1-y_2)^2}{L^2}}, & |y_1 - y_2| \le L. \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

where L is the age gap in the overlapping region (we set L as 4 years old and illustrate the influence of the weighting function).

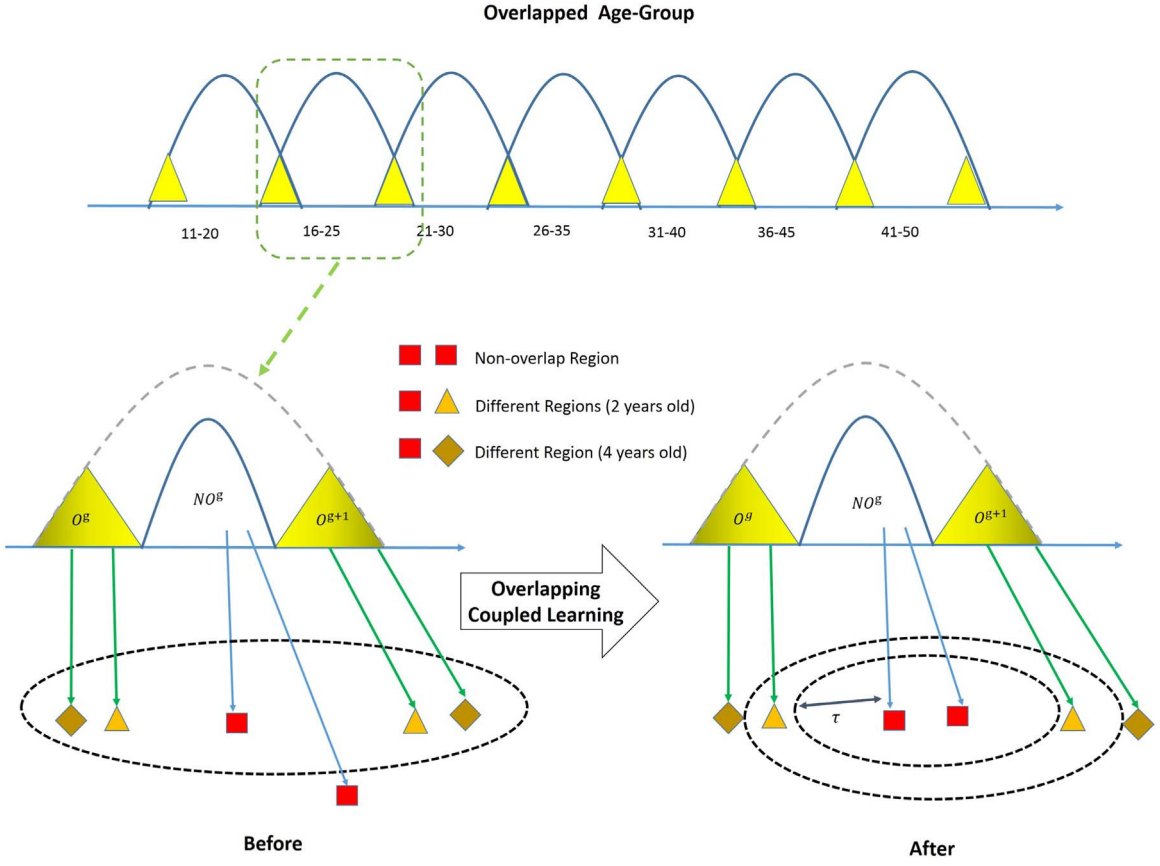

Fig. 3. The basic idea of the proposed overlapping coupled learning. To achieve group overlaps, one age group can be split into three regions: one non-overlap region in the middle colored in white and two symmetrical overlapping regions adjacent with neighboring age groups colored in yellow. Taken six face samples as an example, we expect the distance of a pair with a small age gap where one sample in overlap region (red square) and the other in non-overlap region (yellow triangle) should be smaller than that (red square and yellow diamond) with a large age gap in the learned feature space. (Best view in the color PDF file.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

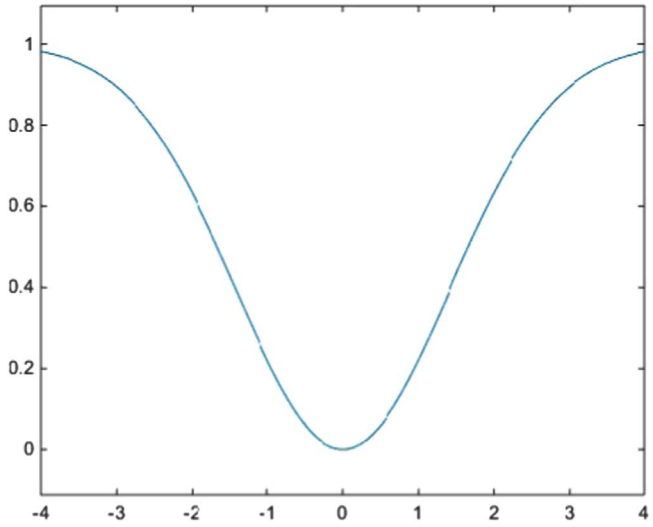

Fig. 4. Illustration of weighting function to define $\omega(y_1, y_2)$. In this figure, we set $|y_1 - y_2| \le 4$. According to the function curve, When y_1 is close to y_2, the weighting value is small. This is because our goal attempts to enforce the mis-estimation should cost small between the nearer samples y_1 and y_2.

There are two key objectives for (7):

(1) The distance between obtained face descriptors with a small age gap is smaller than that with a large age gap because the estimation error of a face pair from the overlapping region is less than that with a larger gap.
(2) With the learned nonlinear feature embedding, different weights can be assigned to different face pairs in overlapping region with different age gaps, so that the similarities of face pairs covering adjacent age groups can be treated dynamically because the neighboring age labels encounter smoothness in a wide aging progress.

By combining (4) and (7), we rewrite the objective function as follow:

$$\min_f J = J_1 - \lambda_1 J_2 + \lambda_2 J_3 + \lambda_3 J_3 = \frac{1}{2Nk_1}\sum_i^N\sum_j^N\sum_g^G Q_g^{ij} d_f^2(\mathbf{x}_i^g, \mathbf{x}_j^g) - \lambda_1\frac{1}{2Nk_2}\sum_i^N\sum_j^N\sum_{g_1\neq g_2}^G S_{g_1,g_2}^{ij} d_f^2(\mathbf{x}_i^{g_1}, \mathbf{x}_j^{g_2}) + \lambda_2\frac{1}{2}\sum_p^P\sum_g^G \ell_{p1,p2}^{O_g}(\tau - d_f^2(\mathbf{x}_{p1}, \mathbf{x}_{p2}))\cdot\omega(y_1, y_2)) + \lambda_3\frac{1}{2}\sum_m^M (\|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2), \tag{9}$$

where λ_1, λ_2, λ_3 are regularization parameters.

3.1.3. Optimization

To optimize (9), we employ the stochastic gradient decent method to obtain the parameters $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}$, where $m = 1, 2, \ldots, M$. The gradients of objective function with respect to the parameters $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}$ can be computed as follows:

$$\frac{\partial J}{\partial \mathbf{W}^{(m)}} = \frac{1}{Nk_1}\sum_i^N\sum_j^N\sum_g^G Q_g^{ij}(\Delta_{ij}^{(m)}\mathbf{h}_i^{(m-1)^T} + \Delta_{ji}^{(m)}\mathbf{h}_j^{(m-1)^T}) - \lambda_1\frac{1}{Nk_2}\sum_i^N\sum_j^N\sum_{g_1\neq g_2}^G S_{g_1,g_2}^{ij}(\Delta_{ij}^{(m)}\mathbf{h}_i^{(m-1)^T} + \Delta_{ji}^{(m)}\mathbf{h}_j^{(m-1)^T}) + \lambda_2\sum_p^P\sum_g^G \ell_{p1,p2}^{O_g}(\Delta_{p1,p2}^{(m)}\mathbf{h}_{p1}^{(m-1)^T} + \Delta_{p2,p1}^{(m)}\mathbf{h}_{p2}^{(m-1)^T})\cdot\omega_{y_{p1},y_{p2}} + \lambda_3\mathbf{W}^{(m)}, \tag{10}$$

where

$$\frac{\partial J}{\partial \mathbf{b}^{(m)}} = \frac{1}{Nk_1}\sum_i^N\sum_j^N\sum_g^G Q_g^{ij}(\Delta_{ij}^{(m)} + \Delta_{ji}^{(m)}) - \lambda_1\frac{1}{Nk_2}\sum_i^N\sum_j^N\sum_{g_1\neq g_2}^G S_{g_1,g_2}^{ij}(\Delta_{ij}^{(m)} + \Delta_{ji}^{(m)}) + \lambda_2\sum_p^P\sum_g^G \ell_{p1,p2}^{O_g}(\Delta_{p1,p2}^{(m)} + \Delta_{p2,p1}^{(m)})\cdot\omega_{y_{p1},y_{p2}} \quad + \lambda_3\mathbf{b}^{(m)}, \tag{11}$$

where the updating equations are computed as follows:

$\Delta_{ij}^{(M)} = (\mathbf{h}_i^{(M)} - \mathbf{h}_j^{(M)}) \odot \varphi'(\mathbf{z}_i^{(M)})$, $\Delta_{ji}^{(M)} = (\mathbf{h}_j^{(M)} - \mathbf{h}_i^{(M)}) \odot \varphi'(\mathbf{z}_j^{(M)})$,
$\Delta_{1p,2p}^{(M)} = (\mathbf{h}_{1p}^{(M)} - \mathbf{h}_{2p}^{(M)}) \odot \varphi'(\mathbf{z}_{1p}^{(M)})$, $\Delta_{2p,1p}^{(M)} = (\mathbf{h}_{2p}^{(M)} - \mathbf{h}_{1p}^{(M)}) \odot \varphi'(\mathbf{z}_{2p}^{(M)})$,
$\Delta_{ij}^{(m)} = (\mathbf{W}^{(m+1)^T}\Delta_{ij}^{(m+1)}) \odot \varphi'(\mathbf{z}_i^{(m)})$, $\Delta_{ji}^{(m)} = (\mathbf{W}^{(m+1)^T}\Delta_{ji}^{(m+1)}) \odot \varphi'(\mathbf{z}_j^{(m)})$,
$\Delta_{1p,2p}^{(m)} = (\mathbf{W}^{(m+1)^T}\Delta_{1p,2p}^{(m+1)}) \odot \varphi'(\mathbf{z}_{1p}^{(m)})$,
$\Delta_{2p,1p}^{(m)} = (\mathbf{W}^{(m+1)^T}\Delta_{2p,1p}^{(m+1)}) \odot \varphi'(\mathbf{z}_{2p}^{(m)})$, $\mathbf{z}_i^{(m)} = \mathbf{W}^{(m)}\mathbf{h}_i^{(m-1)} + \mathbf{b}^{(m)}$.

where $m = 1, 2, \ldots, M-1$ and $\odot$ denote the element-wise multiplication. Then, $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ can be updated as follows until convergence:

$$\mathbf{W}^{(m)} = \mathbf{W}^{(m)} - \rho\frac{\partial J}{\partial \mathbf{W}^{(m)}}, \quad \mathbf{b}^{(m)} = \mathbf{b}^{(m)} - \rho\frac{\partial J}{\partial \mathbf{b}^{(m)}}. \tag{12}$$

where ρ is the learning rate, which controls the convergence speed of objective function J. Algorithm 1 summarizes the detailed procedure of optimization for GA-DFL.

Algorithm 1. The optimization of GA-DFL

Input: Training set X; Parameters: λ_1, λ_2, λ_3, γ, learning rate ρ, total iterative number Γ, and convergence error ε.
Output: Parameters: $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^M$.
Initialize $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^M$ according to(13).
// Optimization by back-prorogation:
for $t = 1, 2, \ldots, \Gamma$ **do**
  Randomly select a batch of $\mathbf{X}$.
  // Forward propagation:
  For $1, 2, \ldots, M$ **do**
  | Perform forward propagation to get $\mathbf{h}_i^{(m)}$, $\mathbf{h}_j^{(m)}$.
  end
  // Perform back propagation and compute the gradients:
  for $M, M-1, \ldots, 1$ **do**
  | Compute gradients $\frac{\partial J_t}{\partial \mathbf{W}^{(m)}}$ and $\frac{\partial J_t}{\partial \mathbf{b}^{(m)}}$ according to(10) and (11).
  end
  // Update parameters:
  Perform update for parameter set $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}$.
  for $m = 1, 2, \ldots, M$ **do**
  | Update$\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ according to (12)
  end
  Calculate J_t using (9).
  If $t > 1$ and $|J_t - J_{t-1}| < \varepsilon$, go to **Return**.
end
Return: $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^M$.

3.2. Multi-path CNN

In this subsection, we define a multi-path CNN, which is fine-tuned by the proposed loss function (9) to enhance the capacity of the learned face representation.

Conventional deep learning-based methods deployed the age estimation-specific loss function on the top of one-scale CNN and the parameters of the networks are optimized by back-propagation. Moreover, some well-pretrained deep models can be fine-tuned by limited face aging data, which has gained significant improvements [48,16]. However, the original scales of face data across different datasets are largely various, whereas the strong invariance of the model can be harmful for the singe scale face input. To address this, we propose a multi-path CNN architecture to capture multi-level semantic invariance at different scales and deploy the objective loss (9) on the top of the defined CNN to perform the joint fine-tuning. To be specific, our network starts with an efficient VGG-16 Face Net [41] and two shallower CNNs to capture the fine-grain feature representation with exploiting multiple scale information. Having obtained 4096-hidden layers from VGG and two subnets, we concatenate them into a long vector and then take it as feature input into an learned age estimator. Fig. 5 illuminates the details of the proposed network architecture. Since our proposed loss is based on the similarity of face pairs on Euclidean space, we perform the ℓ_2 normalization to make the feature normalized and measurable.

3.3. Implementation details

We normalized face image in size of 224×224 and then downscaled it in 64×64 and 32×32 as multi-scale input to the defined multi-path CNN. For the VGG-16 Face Net, we preserved the main architecture detailed in [41] except for the top loss layer. For subnet-1, the architecture consists of a series of convolution, max pooling, ReLU and fully connected functions. The deep structure of subnet-2 is similar with subnet-1, except for the input size. More details are tabulated in Table 1.

It is important to initialize the network parameters $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$, where m denotes the layer number of the deep network. In our experiments, we applied the well-initialized parameters of VGG-16 Face Net and employed the normalized random initialization method to initialize $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ for the proposed subnets. To be specific, the layer-wise weight $\mathbf{W}^{(m)}$ was initialized by the uniform distribution as:

$$\mathbf{W}^{(m)} \sim \left[-\frac{\sqrt{6}}{\sqrt{r^{(m)} + r^{(m-1)}}}, \frac{\sqrt{6}}{\sqrt{r^{(m)} + r^{(m-1)}}} \right] \tag{13}$$

where the bias $\mathbf{b}^{(m)}$ in this layer was set as $\mathbf{0}$, and $r^{(m)}$ was the size of m-th input layer.

In terms of the parameter setting employed in our method, we set $\alpha = 10$ years to each group and assigned the overlap region to $\kappa = 4$ for overlapping coupled learning method. The balance parameter was determined as $\{\lambda_1 = 0.2, \lambda_2 = 0.5, \lambda_3 = 0.001\}$ by cross-validation. We assigned the values of the weight decay, moment parameter and learning rate empirically to 0.0001, 0.9, and 0.01 respectively for training stage. The whole training procedure of our GA-DFL needed 20 iterations to convergence.

4. Experiments

In this section, we conducted facial age estimation experiments on the widely used FG-NET [25], MORPH (Album2) [50] and Chalearn Challenge Dataset [51] to show the effectiveness of the proposed GA-DFL. The followings describe the details of experimental settings and results.

4.1. Experimental settings

We resized each face images from training set in 224×224 and fed them into the defined network. Having obtained the learned face representation, we employed OHRank [20] as the age estimator to gain outperformed performances for facial age estimation.

Before the evaluation of our method, we performed a face-processing for all images. Specifically, we detected the face bounding box and facial landmarks on the origin image. All the face detection and alignment were handled by the open source library Dlib [52]. We utilized three landmarks including two centers of eyes and nose base to align the detected face into the canonical coordinate system by using similar transformation.

We utilized the mean absolutely error (MAE) [21,5,1,6,24,27,47] to measure the error between the predicted age and the ground-truth, which was normalized and defined as follows:

$$MAE = \frac{\|\hat{y} - y^*\|_2}{N} \tag{14}$$

where $\hat{y}$ and y^* denote predicted and ground-truth age value. N denotes the number of the testing samples.

We also applied the cumulative score (CS) [21,5,1,6,24,27,47]

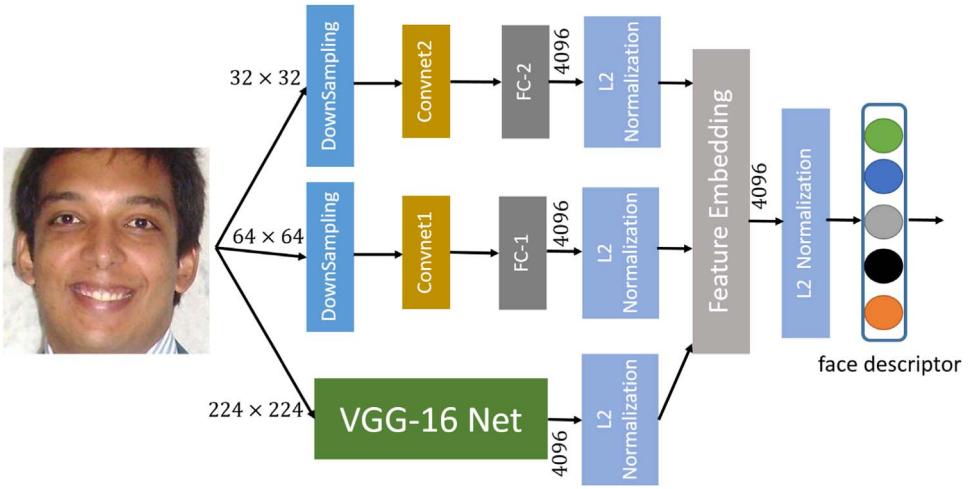

Fig. 5. The multi-path network architecture. Each face image starts with three scales of 224×224, 64×64, 32×32, respectively. The green box VGG-16 Net is the same as the employed architecture in [41]. Besides, there are two lower scale paths including convnet1, FC1, convnet2 and FC2. The number on the top of the architecture is the input scales, and 4096 denotes the output dimension of each subnet. Finally, we normalized the embedded feature also in dimension 4096. More details are listed in Table 1.

Table 1
The multi-path network architecture.

SubNet1	1	2	3	4	5	6	7	8	9	10	11
	conv	relu	pool	conv	relu	pool	conv	relu	pool	conv	conv
filt dim	5	–	3	5	–	3	5	–	3	4	1
num filt	32	–	–	64	–	–	128	–	–	256	512
stride	1	–	2	1	–	2	1	–	2	1	1
pad	2	–	1	2	–	1	2	–	1	0	0
SubNet2	1	2	3	4	5	6	7	8	9	10	11
	conv	relu	pool	conv	relu	pool	conv	relu	pool	conv	conv
filt dim	5	–	3	5	–	3	5	–	3	2	1
num filt	256	–	–	512	–	–	512	–	–	512	1024
stride	1	–	2	1	–	2	1	–	2	2	1
pad	2	–	1	2	–	1	2	–	1	0	0

curves that were demonstrated to quantitatively evaluate the performance of age estimation methods. We provided the cumulative prediction accuracy at the error θ, which is defined as:

$$CS(\theta) = \frac{N_{e\leq\theta}}{n} \times 100\% \tag{15}$$

where $N_{e\leq\theta}$ is the number of images on which the error θ is no less than e. Basically, θ starts from 0.

4.2. Visualization of the proposed networks

To illustrate what we have learned in our multi-path network, we visualized the learned filters and the corresponding feature maps. Having obtained the trained network, we show the first convolution layers in the triple subnets to see the learned filters, showed in Fig. 7. To visualize the learned face representation, we fed the network with normalized face image and figured the feature maps for their corresponding filters. To achieve this, we conducted heat maps for each maps and showed the visualizations of feature maps are showed Fig. 6. We have made three observations from the visualizations:

(1) Both the VGG-16 Net and two other SubNets detect the meaningful representation of facial properties such as eye corners, beard and nasolabial folds, which clearly shows the age-informative details.
(2) VGG-16 is referred to capture the detailed visual appearance, while two low-resolution subnets characterize the coarse contours for age information.
(3) The learned features are robust to the large invariance in such cases that some face images encounter cluttered environment such as wearing glasses.

4.3. Experiments on the FG-NET dataset

There are 1002 images from 82 persons in FG-NET Dataset [25] and there exists averaging 12 samples for each person. The age range in this dataset covers from 0 to 69. The FG-NET dataset encounters large variations in pose, illumination and expression.

We employed the leave-one-person-out (LOPO) strategy to conduct the age estimation experiments. Specifically, we randomly selected face images from one person as testing images, and the remaining were used for model training. Thus the whole experiments should be performed 82 times for cross-validation. Finally, we averaged the 82 folds results as the final age estimation results.

4.3.1. Comparisons with different facial age estimation approaches

Table 2 demonstrates the MAE performance compared with the state-of-the-art methods. The results were cropped from the original paper. Fig. 8 shows the CS curves with different facial age estimation. From the results, without any additional datasets, our GA-DFL outperforms other methods. This is because our model aims at learning a series nonlinear mapping functions while considering the age-group specific relationship. Moreover, the extensive model GA-DFL integrated with multi-path network (described in Section 3.2) obtains higher MAE and CS performance than GA-DFL, which shows that the multi-path network has provided complementary scale information to predict facial age value (Fig. 9).

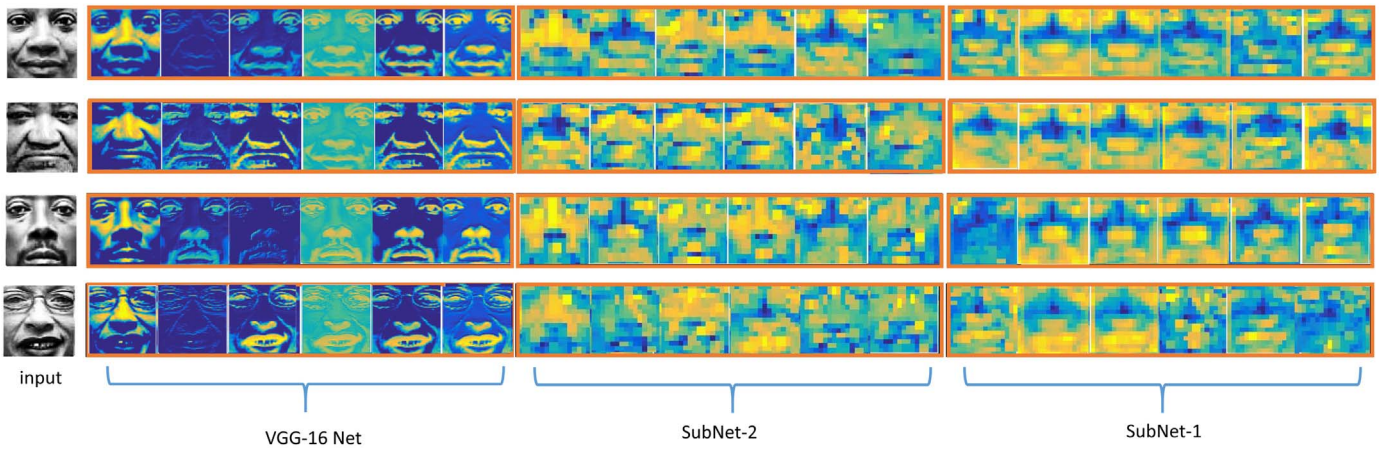

Fig. 6. The visualization of learned feature maps on MORPH Dataset. The fist column denotes the face input for deep network, and the remaining are feature maps of the first convolutional layer in each subnets: VGG-16 Net, subnet-1, subnet-2.

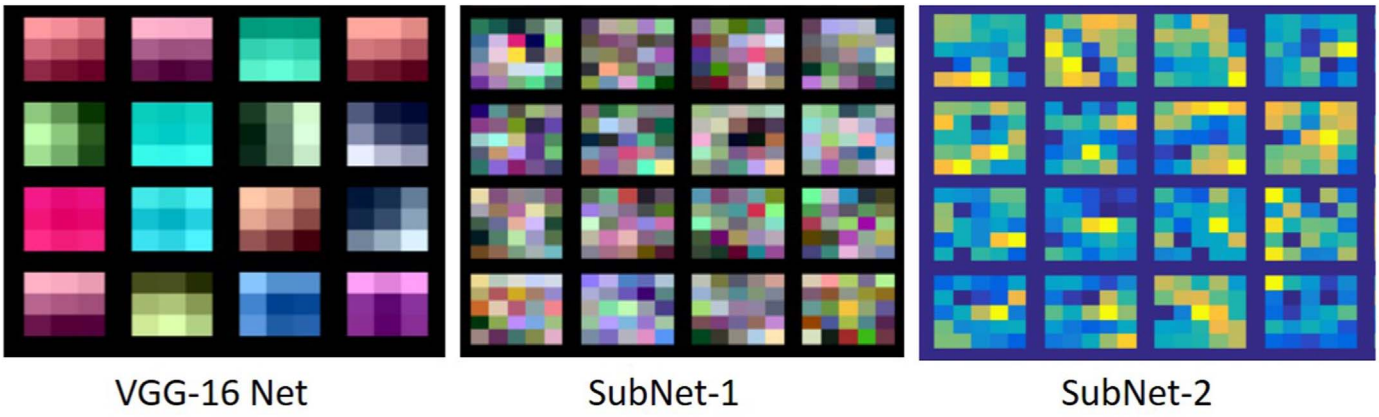

Fig. 7. The visualization of the first convolutional filters for the corresponding VGG-16 net and the other two subnets. For each filter bank, the selected convolution filter weights are illustrated in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of averaged errors with state-of-the-art approaches on FG-NET.

Method	Model Description	Year	MAE
KNN			8.24
SVM			7.25
MLP			6.95
RUN [53]	AAM + RUN	2007	5.78
AGES [1]	AAM + Aging pattern subspace	2007	6.77
LARR [6]	AAM + Locally adjusted robust regression	2008	5.07
PFA [54]	AAM + Probabilistic fusion approach	2008	4.97
KAGES [55]	AAM + Kernel AGES	2008	6.18
MSA [56]	AAM + Multilinear subspace analysis	2009	5.36
SSE [57]	AAM + Submanifold embedding	2009	5.21
mKNN [58]	AAM + Metric Learning	2009	5.21
MTWGP [27]	AAM + Multi-task warped GPR	2010	4.83
RED-SVM [28]	AAM + Red SVM	2010	5.21
OHRanker [20]	AAM + Ordinal hyperplanes ranker	2011	4.48
PLO [26]	Feature selection + OHRanker	2012	4.82
IIS-LLD [23]	AAM/BIF + learning from label distribution	2013	5.77
CPNN [23]	AAM/BIF + learning from label distribution	2013	4.76
CA-SVR [59]	AAM + cumulative/joint attribute learning	2013	4.67
CS-LBFL [3]	Feature learning + OHRanker	2015	4.43
CS-LBMFL [3]	Multiple feature learning + OHRanker	2015	4.36
GA-DFL	Deep feature learning + OHRanker		**4.16**
GA-DFL (MP-CNN)	Deep feature learning + OHRanker		**3.93**

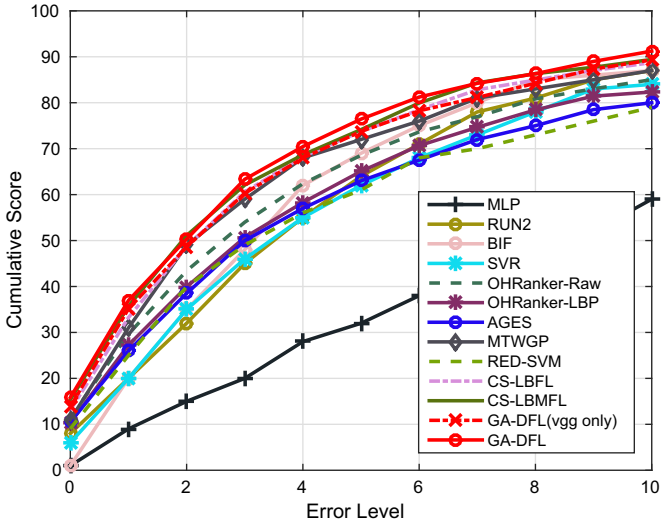

Fig. 8. The CS curves compared with different facial age estimation methods on FG-NET.

4.3.2. Comparisons with several deep learning approaches

To evaluate the effectiveness of the proposed deep learning framework, we conducted several comparisons under different deep features and loss functions. Firstly, we directly extracted features from the pretrained VGG net and further employed KNN classifier to construct a baseline method. In terms of unsupervised setting, we fed the extracted feature descriptors to OHRanker without fine-tuning VGG. For supervised setting, we deployed the Softmax loss [63] for age classification and the linear regression for age regression on the top of VGG-16 Net and then fine-tuned the whole deep networks. Table 4 shows the MAE performances. According to the demonstrated results, we have drawn some conclusions: Our model obtains outperformed performance compared with different deep learning methods. This is because we explicitly modeled the age-related information under the deep learning framework. By employing VGG Net, we have obtained an improved performance on facial age estimation with carefully tuning the deep networks by classification and regression specific loss. Unsupervised VGG feature improved the MAE performance compared with raw pixel, which shows that the VGG Net could capture the visual appearance for facial patterns, which provides main cues for age estimation (Fig. 10).

4.3.3. Comparisons with existing feature learning methods

We compared our GA-DFL approach with existing feature learning approach LQP [60], DFD [61], RICA [62] and CS-LBFL [3]. For RICA and DFD, we use the released code for experiments. For LQP and CS-LBFL, we implemented it from the paper details. We also extend CS-LBFL under a two-layer deep neural networks, which consists of 500 – 256 – 128 multi-layer neural networks followed by a nonlinear function ReLU [63]. Table 3 shows the MAEs of different feature learning approaches. According to the performance, we have gained the highest performance, even outperformed the deep extension of the feature learning based method CS-LBFL.

To evaluate more fair experiments, we implemented multi-scale feature learning approach based on LQP, DFD and RICA and CS-LBMFL. The MAEs of results are also showed in Table 3. As can be seen, our proposed GA-DFL performs better than the other multi-scale feature learning approaches. This is because our learned deep features can represent the nonlinear mapping between face image and age-informative targets.

4.3.4. Comparisons with different age estimators

We investigated the effectiveness of different facial age estimators with our learned feature. Specifically, we compared with support vector regression (SVR) [64] and OHRanker and computed the MAEs for final performance. We also implemented common losses such as softmax and regression as age estimator under deep architecture. As the results are demonstrated in Table 5, we see our model with OHRanker

Fig. 9. The sampled examples of two persons from the FG-NET dataset. Each row represents one person identity and the number below each face image is the age value. The dataset encounters crowded background and large invariance due to expressions and aspect ratio.

performs better than deep learning based age estimators. Moreover, our model with OHRanker outperforms SVR methods, but the difference is not large.

4.3.5. Performance analysis of different factors

We conducted experiments on performance analysis of different factors of our methods, with OHRanker as the age predictor. Table 6 is demonstrated the MAEs performances. First, we set each age group covering years of $\alpha = \{2, 5, 10\}$ and assigned $\kappa = \{2, 4\}$ to the overlapped stride. According to the results, the MAE under group 5 performs the best. Moreover, we investigated the influences of with and without overlapped coupled learning strategies. Specifically, we denotes the without overlap setting by directly setting γ to 0 in (7) while to 0.3 as with overlapped coupled learning. From the results, we see that under the same age group number, the overlapped coupled learning can improve the prediction performance. It is because our model achieves in exploiting the age-sensitive relationship, while the age group without overlapped learning degrades the performance due to the misclassified samples on the bound of the age group.

4.4. Experiments on the MORPH dataset

MORPH (Album 2) dataset [50] contains 55608 face images from about 13000 subjects. The age range lies from 16 to 77 years old and there exists averaging 4 samples per person. Since MORPH dataset contains thousands of persons, LOPO cross-validation is time-consuming. Thus, we performed 10-folds cross-validation for performance evaluation. Specifically, we first divided the whole dataset into equally size of 10 folds. We randomly selected one fold as testing set and the remaining folds as training set. Sequentially, we repeated whole

Table 3
Comparison of MAE with different feature learning based approaches on FG-NET.

Method	MAE
LQP [60]	4.70
DFD [61]	4.57
RICA [62]	6.09
CS-LBFL [3]	4.43
Deep-CS-LBFL	4.40
GA-DFL	**4.16**
MDFD	5.35
MRICA	5.65
CS-LBMFL	4.36
Deep-CS-LBMFL	4.22
GA-DFL (MP-CNN) + OHRanker	**3.93**

Fig. 10. The sampled examples from the MORPH dataset. The number below face image is the age value for each person.

Table 4
Comparison of MAE with different deep learning methods on FG-NET.

Method	MAE
Unsupervised pVGG feature + KNN	6.54
VGG feature + Regression	4.88
VGG feature + Softmaxloss	4.72
Unsupervised VGG features + OHRanker	4.44
GA-DFL (VGG only) + OHRanker	4.16
GA-DFL (MP-CNN) + OHRanker	**3.93**

Table 5
Comparison of MAE with different age estimators on FG-NET.

Method	MAE
GA-DFL (VGG only) + SVR	4.32
GA-DFL + SVR	4.47
GA-DFL (VGG only) + OHRanker	4.16
GA-DFL (MP-CNN) + OHRanker	**3.93**

Table 6
MAEs Comparisons with Different Learning Strategies on the FG-NET Dataset.

Method	Overlap-Stride κ	Group-Capacity α	MAE
GA-DFL	without overlap	2	5.22
GA-DFL	without overlap	10	5.77
GA-DFL	without overlap	5	5.08
GA-DFL	2	5	4.24
GA-DFL	2	10	4.11
GA-DFL	4	10	**3.93**

procedure 10 times and averaged the performances as final result.

4.4.1. Comparisons with different state-of-the-art methods

We compared our model with several different state-of-the-art facial age estimation approaches. The experimental results are demonstrated in Table 7. All comparable results are reported from the original paper. Fig. 11 shows the CS curves with different facial age estimation methods. As can been seen that the experimental results show that our DA-DFL outperforms most state-of-the-art methods. In particular, we compared our method with deep learning based model DeepRank [69] and OrdinalCNN [44]. According to the results, we inferred that our model performed a very competitive performance, which shows the effectiveness of the proposed method.

Table 7
Comparison of MAEs with different state-of-the-art approaches on MORPH (Album 2).

Method	MAE
KNN	9.64
SVM	7.34
AGES [1]	8.83
MTWGP [27]	6.28
OHRanker [20]	6.49
IIS-LLD [23]	5.69
CPNN [23]	5.67
CA-SVR [59]	4.87
MFOR [65]	5.88
BIF+OLPP [66]	4.20
CS-LDA [67]	6.03
CS-FS [68]	6.59
CS-LBFL [3]	4.52
CS-LBMFL [3]	4.37
rKCCA [29]	3.98
rKCCA + SVM [29]	3.91
CPLF [45]	3.63
DeepRank [47]	3.57
DeepRank+ [47]	3.49
OrdinalCNN [44]	3.27
GA-DFL	**3.37**
GA-DFL (MP-CNN)	**3.25**

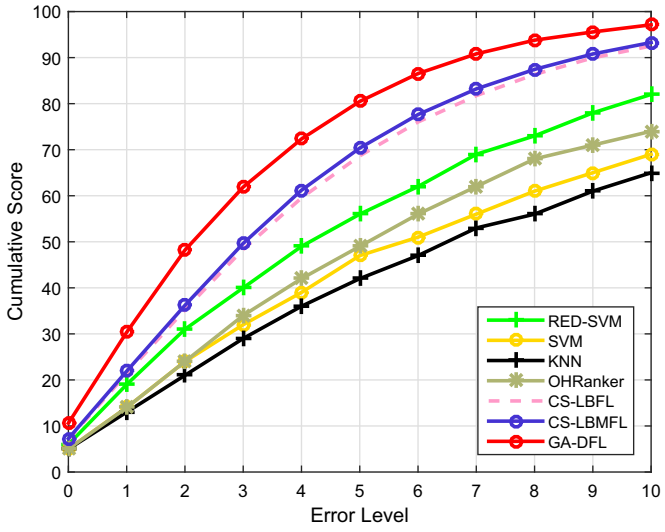

Fig. 11. The CS curves compared with different facial age estimation methods on Morph.

4.4.2. Timing-cost

Our approach was implemented on the Matlab platform with the MatConvnet [70] deep learning toolbox. We trained our model with a speed-up parallel computing technique by one single GPU with Tesla K40. For large amounts of training data in MORPH Dataset, we investigated the training timing-cost for different depths of the VGG-16 Net. Table 8 tabulates the time (frames per second) during training procedure. From these results, we see that the fine-tuning depths directly determines the training timing-cost, this is because a set of the convolution computations can be time-consuming. However, the MAEs becomes higher while more layers are trained, which shows the effectiveness of depths of network model depend on the amount of training data. Lastly, we investigated the testing time with several different feature learning methods on a PC with a i7-CPU@3.40 GHz and a 16 Gb RAM in Table 9. We see that our proposed methods satisfies the real-time requirements.

4.5. Experiments on apparent age estimation dataset

In addition to these four datasets, we presented results of our proposed method on the ICCV 2015 'Looking at People-Age Estimation' Challenge Dataset [51] (Chalearn Challenge Dataset), which is the first dataset on apparent age estimation containing annotations. There are 4112 images for training and 1500 images for validation. The age range is from 0 to 100 years old. All the images were captured in the unconstrained condition with large variance of pose, aspect ratio and low quality.

We initially evaluated the performance of our proposed method by utilizing MAE and CS. By following the protocol provided by the Chalearn Challenge Dataset [51], we also computed the error rate as follows:

Table 8
Training time (fps) performed in MORPH Dataset.

Fine-tuning Depths	MAE	Time
1	4.29	120
2	3.88	100
5	3.69	80
Total	**3.37**	40

Table 9
Computation time (Second) comparison of different feature learning methods in MORPH Dataset.

Method	Time
DFD	0.60
LQP	0.10
RICA	0.35
CS-LBFL	0.06
CS-LBMFL	0.18
GA-DFL	**0.32**
GA-DFL (MP-CNN)	**0.38**
GA-DFL with GPU	**0.02**
GA-DFL (MP-CNN) with GPU	**0.04**

Table 10
Comparison of MAEs and Gaussian errors with different feature learning based approaches on Chalearn challenge dataset.

Method	Model Description	MAE	Gaussian Error
G-LR [19]	Age estimation by facial landmarks	7.09	0.620
OHRanker-SIFT	SIFT + OHRanker	7.18	0.582
OHRanker-LBP	LBP + OHRanker	7.00	0.563
OHRanker-UnsupVGG	Unsupervised VGG feature + OHRanker	7.24	0.593
Deep Regression	VGG Face Net fine-tuned by Linear Regression	5.05	0.456
Deep Softmax	VGG Face Net fine-tuned by Softmaxloss	4.58	0.423
GA-DFL(vgg only)	GA-DFL(vgg only) + OHRanker	**4.39**	**0.393**
GA-DFL	GA-DFL(MP-CNN) + OHRanker	**4.21**	**0.369**

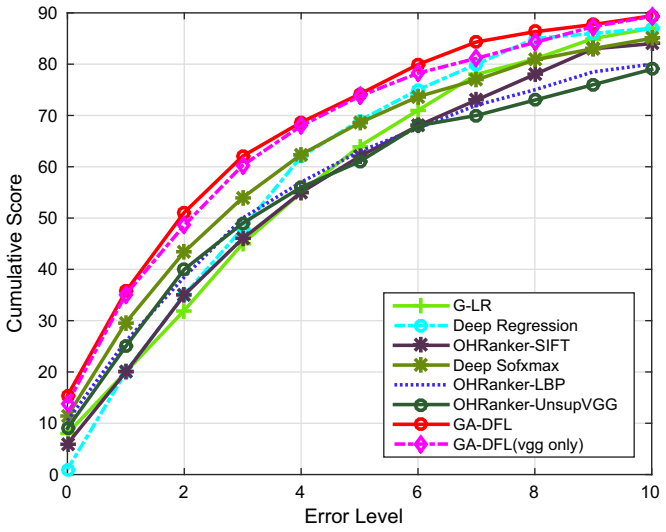

Fig. 12. The CS curves compared with different facial age estimation methods on Chalearn Challenge Dataset.

$$\epsilon = 1 - e^{-\frac{(\hat{y}-\mu)^2}{2\sigma^2}} \tag{16}$$

where $\hat{y}$ denotes the predicted age value, μ is the provided apparent age label which averaged by about 10 users opinions and σ is the standard deviation. To address the age recognition as a classification problem, we floored all the annotated age values to integers to fit on our method. Since the testing data has not been released publicly, we had to conduct experiments on the validation set.

To conduct fair experimental comparisons, we implemented two types of methods to be compared. First, we defined shallow model by utilizing hand-crafted features such as SIFT, LBP and then we performed facial age estimation by OHRanker. To better evaluate our method compared with different deep learning based approaches, we involved with different loss functions including softmax and linear regression by fine-tuning VGG Face Net. In addition, we compared our method with G-LR [19], which aims to predict facial age by jointly exploiting face identification by facial landmarks. Table 10 tabulates the MAEs and Gaussian errors, while Fig. 12 shows the CS curves with different methods. From the results, we can infer that our GA-DFL performs significant improvements on facial age estimation compared with the other deep learning based methods with limited training data (Fig. 13). Furthermore, we illustrated some face images, the CS of which are below 1 years old (showed in Fig. 14). By looking at the images, we can infer that our method is robust to the large invariance in unconstrained environment, especial for extremely pose and resolution changes to a certain extent. The performance of our method can be improved considerably if we train using more number of age labeled data.

4.6. Discussion

We make two observations from the experimental results compared with existing facial age estimation methods:

(1) Compared with the traditional shallow models [5–9,3], we have obtained outperformed experimental results with CNN framework on three public facial age estimation datasets that were captured under both constrained and unconstrained environments. This is because our GA-DFL has learned a set of nonlinear hierarchical feature transformations to capture the nonlinear relationship between face images and age values, while existing shallow models and hand-crafted features are not powerful enough to address this nonlinear problem.
(2) Compared with existing deep learning-based methods [45,47], our GA-DFL has achieve very competitive performances on three face aging datasets. These methods and our GA-DFL follow the deep learning architecture to learn powerful feature representation for facial age estimation. In contrast to them, the learned face representations of GA-DFL has discovered the ordinal relationship from face pair similarity with integrating the aging rank levels and age difference information.

5. Conclusions and future work

In this paper, we have proposed a group-aware deep feature learning (GA-DFL) for facial age estimation. Since the real-world age

Fig. 13. The sampled examples from the apparent age estimation dataset. The number below each face image is the corresponding apparent age.

Fig. 14. The selected examples from the Chalearn Challenge Dataset, where the CS errors are below 1 years old. These resulting face images present the robustness of learned feature representation to the large invariance of facial wearing glasses, pose and expressions.

labels are correlated and hand-crafted face descriptors are not powerful to model the relationship between face images and age values, we have defined a set of age groups to describe the aging order relationship of face data and implicitly achieved the ordinal age-group relationship. Moreover, we developed an overlapped coupled learning to smooth the adjacent age groups. To further improve the performance, we designed a multi-path CNN to capture age-informative appearances from different scale information. Experimental results on three released datasets have evaluated the effectiveness of the proposed GA-DFL compared with the state-of-the-art. How to learn personalized face descriptors for age estimation is an interesting future work.

Acknowledgement

This work is supported by the National Key Research and Development Program of China under Grant 2016YFB1001001, the National Natural Science Foundation of China under Grants 61225008, 61672306, 61572271, 61527808, 61373074 and 61373090, the National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program.

References

[1] X. Geng, Z. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2234–2240.
[2] R. He, W.-S. Zheng, B.-G. Hu, Maximum correntropy criterion for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1561–1576.
[3] J. Lu, V.E. Liong, J. Zhou, Cost-sensitive local binary feature learning for facial age estimation, IEEE Trans. Image Process. 24 (12) (2015) 5356–5368.
[4] X. Shu, J. Tang, H. Lai, L. Liu, S. Yan, Personalized age progression with aging dictionary, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3970–3978.
[5] Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Trans. Multimed. 10 (4) (2008) 578–584.
[6] G. Guo, Y. Fu, C.R. Dyer, T.S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, IEEE Trans. Image Process. 17 (7) (2008) 1178–1188.
[7] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[8] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[9] G. Guo, G. Mu, Y. Fu, T.S. Huang, Human age estimation using bio-inspired features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 112–119.
[10] Y.H. Kwon, N. da Vitoria Lobo, Age classification from facial images, Comput. Vis. Image Underst. 74 (1) (1999) 1–21.
[11] Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Trans. Multimed. 10 (4) (2008) 578–584.
[12] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1955–1976.
[13] R. He, W. Zheng, T. Tan, Z. Sun, Half-quadratic based iterative minimization for robust sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 261–275.
[14] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, X. Chen, Agenet: Deeply learned regressor and classifier for robust apparent age estimation, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 258–266.
[15] Z. Kuang, C. Huang, W. Zhang, Deeply learned rich coding for cross-dataset facial age estimation, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 338–343.
[16] X. Yang, B. Gao, C. Xing, Z. Huo, X. Wei, Y. Zhou, J. Wu, X. Geng, Deep label distribution learning for apparent age estimation, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 344–350.
[17] R. Ranjan, S. Zhou, J. Chen, A. Kumar, A. Alavi, V. M. Patel, R. Chellappa, Unconstrained age estimation with deep convolutional neural networks, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 351–359.
[18] X. Wang, R. Guo, C. Kambhamettu, Deeply-learned feature for age estimation, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 534–541.
[19] T. Wu, P. Turaga, R. Chellappa, Age estimation and face verification across aging using landmarks, IEEE Trans. Inf. Forensics Secur. 7 (6) (2012) 1780–1788.
[20] K. Chang, C. Chen, Y. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 585–592.
[21] Y. Chen, C. Hsu, Subspace learning for facial age estimation via pairwise age ranking, IEEE Trans. Inf. Forensics Secur. 8 (12) (2013) 2164–2176.
[22] Z. Li, U. Park, A.K. Jain, A discriminative model for age invariant face recognition, IEEE Trans. Inf. Forensics Secur. 6 (3–2) (2011) 1028–1037.
[23] X. Geng, C. Yin, Z. Zhou, Facial age estimation by learning from label distributions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2401–2412.
[24] G. Guo, G. Mu, Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 657–664.
[25] A. Lanitis, C.J. Taylor, T.F. Cootes, Toward automatic simulation of aging effects on face images, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 442–455.
[26] C. Li, Q. Liu, J. Liu, H. Lu, Learning ordinal discriminative features for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2570–2577.
[27] Y. Zhang, D. Yeung, Multi-task warped gaussian process for personalized age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2622–2629.
[28] K. Chang, C. Chen, Y. Hung, A ranking approach for human ages estimation based on face images, in: Proceedings of the International Conference on Pattern Recognition, 2010, pp. 3396–3399.
[29] G. Guo, G. Mu, A framework for joint estimation of age, gender and ethnicity on a large database, Image Vis. Comput. 32 (10) (2014) 761–770.
[30] G. Guo, G. Mu, Y. Fu, T.S. Huang, Human age estimation using bio-inspired features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 112–119.
[31] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, in: Neural Information Processing Systems, 2006, pp. 153–160.
[32] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[33] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Neural Information Processing Systems, 2012, pp. 1106–1114.
[34] D.C. Ciresan, U. Meier, L.M. Gambardella, J. Schmidhuber, Deep, big, simple

neural nets for handwritten digit recognition, Neural Comput. 22 (12) (2010) 3207–3220.
[35] Y. Zhang, K. Sohn, R. Villegas, G. Pan, H. Lee, Improving object detection with deep convolutional networks via bayesian optimization and structured prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 249–258.
[36] N. Wang, D. Yeung, Learning a deep compact image representation for visual tracking, in: Neural Information Processing Systems, 2013, pp. 809–817.
[37] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
[38] S. Yang, P. Luo, C.C. Loy, X. Tang, From facial parts responses to face detection: A deep learning approach, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3676–3684.
[39] J. Zhang, S. Shan, M. Kan, X. Chen, Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 1–16.
[40] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Neural Information Processing Systems, 2014, pp. 1988–1996.
[41] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Conference, 2015.
[42] Y. Dong, Y. Liu, S. Lian, Automatic age estimation based on deep learning algorithm, Neurocomputing 187 (2016) 4–10.
[43] I.H. Casado, C. Fernández, C. Segura, J. Hernando, A. Prati, A deep analysis on age estimation, Pattern Recognit. Lett. 68 (2015) 239–249.
[44] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4920–4928.
[45] D. Yi, Z. Lei, S.Z. Li, Age estimation by multi-scale convolutional network, in: Proceedings of the Asian Conference on Computer Vision, 2014, pp. 144–158.
[46] G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 34–42.
[47] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, C.-S. Chen, Automatic age estimation from face images via deep ranking, in: Proceedings of the British Machine Vision Conference, 2015.
[48] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, X. Chen, Agenet: Deeply learned regressor and classifier for robust apparent age estimation, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 258–266.
[49] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[50] K.R. Jr., T. Tesafaye, MORPH: A longitudinal image database of normal adult age-progression, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2006, pp. 341–345.
[51] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzàlez, H.J. Escalante, D. Misevic, U. Steiner, I. Guyon, Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 243–251.
[52] D.E. King, Dlib-ml: a machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[53] S. Yan, H. Wang, X. Tang, T.S. Huang, Learning auto-structured regressor from uncertain nonnegative labels, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[54] G. Guo, Y. Fu, C.R. Dyer, T.S. Huang, A probabilistic fusion approach to human age prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–6.
[55] X. Geng, K. Smith-Miles, Z.-H. Zhou, Facial age estimation by nonlinear aging pattern subspace, in: Proceedings of the ACM Multimedia Conference, 2008, pp. 721–724.
[56] X. Geng, K. Smith-Miles, Facial age estimation by multilinear subspace analysis, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 865–868.
[57] S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, T.S. Huang, Synchronized submanifold embedding for person-independent pose estimation and beyond, IEEE Trans. Image Process. 18 (1) (2009) 202–210.
[58] C. Li, Q. Liu, J. Liu, H. Lu, Learning distance metric regression for facial age estimation, in: Proceedings of the International Conference on Pattern Recognition, 2012, pp. 2327–2330.
[59] K. Chen, S. Gong, T. Xiang, C.C. Loy, Cumulative attribute space for age and crowd density estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2467–2474.
[60] S.U. Hussain, T. Napoléon, F. Jurie, Face recognition using local quantized patterns, in: Proceedings of the British Machine Vision Conference, 2012, pp. 11–pages.
[61] Z. Lei, M. Pietikainen, S.Z. Li, Learning discriminant face descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 289–302.
[62] Q.V. Le, A. Karpenko, J. Ngiam, A.Y. Ng, Ica with reconstruction cost for efficient overcomplete feature learning, in: Neural Information Processing Systems, 2011, pp. 1017–1025.
[63] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Neural Information Processing Systems, 2012, pp. 1097–1105.
[64] A. Smola, V. Vapnik, Support vector regression machines, in: Neural Information Processing Systems, Vol. 9, 1997, pp. 155–161.
[65] R. Weng, J. Lu, G. Yang, Y.-P. Tan, Multi-feature ordinal ranking for facial age estimation, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2013, pp. 1–6.
[66] G. Guo, G. Mu, Human age estimation: What is the influence across race and gender?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 71–78.
[67] J. Lu, Y.-P. Tan, Cost-sensitive subspace learning for human age estimation, in: Proceedings of the International Conference on Image Processing, 2010, pp. 1593–1596.
[68] L. Miao, M. Liu, D. Zhang, Cost-sensitive feature selection with application in software defect prediction, in: Proceedings of the International Conference on Pattern Recognition, 2012, pp. 967–970.
[69] K. Chang, C. Chen, A learning framework for age rank estimation based on face images with scattering transform, IEEE Trans. Image Process. 24 (3) (2015) 785–798.
[70] A. Vedaldi, K. Lenc, Matconvnet – convolutional neural networks for matlab, in: Proceedings of the ACM Multimedia Conference, 2015.

Hao Liu received the B.S. degree in software engineering from Sichuan University, China, in 2011 and the Engineering Master degree in computer technology from University of Chinese Academy of Sciences, China, in 2014. He is currently pursuing the Ph.D. degree at the department of automation, Tsinghua University. His research interests include face recognition, facial age estimation and deep learning.

Jiwen Lu received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from the Nanyang Technological University, Singapore, in 2003, 2006, and 2012, respectively. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. From March 2011 to November 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/co-authored over 130 scientific papers in these areas, where more than 50 papers are published in the IEEE Transactions journals and top-tier computer vision conferences. He serves/has served as an Associate Editor of Pattern Recognition Letters, Neurocomputing, and the IEEE Access, a Guest Editor of Pattern Recognition, Computer Vision and Image Understanding, Image and Vision Computing and Neurocomputing, and an elected member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is/was a Workshop Chair/Special Session Chair/Area Chair for more than 10 international conferences. He has given tutorials at several international conferences including ACCV'16, CVPR'15, FG'15, ACCV'14, ICME'14, and IJCB'14. He was a recipient of the First-Prize National Scholarship and the National Outstanding Student Award from the Ministry of Education of China in 2002 and 2003, the Best Student Paper Award from Pattern Recognition and Machine Intelligence Association of Singapore in 2012, the Top 10% Best Paper Award from IEEE International Workshop on Multimedia Signal Processing in 2014, and the National 1000 Young Talents Plan Program in 2015, respectively. He is a senior member of the IEEE.

Jianjiang Feng is an associate professor in the Department of Automation at Tsinghua University, Beijing. He received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007. From 2008 to 2009, he was a Post Doctoral researcher in the PRIP lab at Michigan State University. He is an Associate Editor of Image and Vision Computing. His research interests include fingerprint recognition and computer vision.

Jie Zhou received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the International Journal of Robotics and Automation and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.