Pattern Recognition 75 (2018) 51–62

Contents lists available at ScienceDirect

Pattern Recognition

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

Distance metric optimization driven convolutional neural network for age invariant face recognition

Ya Li[a,*], Guangrun Wang[b], Lin Nie[b], Qing Wang[b,*], Wenwei Tan[c]

[a] Guangzhou University, Higher Education Mega Center, Guangzhou 510006, PR China
[b] Sun Yat-Sen University, Higher Education Mega Center, Guangzhou 510006, PR China
[c] Hisilicon Technologies Co., Ltd., Longgang District, Shenzhen 518129, China

ARTICLE INFO

Article history:
Received 14 November 2016
Revised 30 September 2017
Accepted 7 October 2017
Available online 17 October 2017

Keywords:
Age invariant
Face recognition
Distance metric
Deep CNN
Joint learning

ABSTRACT

Despite the great advances in face-related works in recent years, face recognition across age remains a challenging problem. The traditional approaches to this problem usually include two basic steps: feature extraction and the application of a distance metric, sometimes common space projection is also involved. On the one hand, handling these steps separately ignores the interactions of these components, and on the other hand, the fixed-distance threshold of measurement affects the model's robustness. In this paper, we present a novel distance metric optimization driven learning approach that integrates these traditional steps via a deep convolutional neural network, which learns feature representations and the decision function in an end-to-end way. Given the labelled training images, we first generate a large number of pairs with a certain proportion of matched and unmatched pairs. For matched pairs, we try to select as many different age instances as possible for each person to learn the identification information that is not affected by age. Then, taking these pairs as input, we aim to enlarge the differences between the unmatched pairs while reducing the variations between the matched pairs, and we update the model parameters by using the mini-batch stochastic gradient descent (SGD) algorithm. Specifically, the distance matrix is used as the top fully connected layer, and the bottom layers representing the image features are integrated with it seamlessly. Thus, the image features and the distance metric can be optimized simultaneously by backward propagation. In particular, we introduce several training strategies to reduce the computational cost and overcome insufficient memory capacity. We evaluate our method on three tasks: age-invariant face identification on the MORPH database, age-invariant face retrieval on the CACD database and age-invariant face verification on CACD-VS database. The experimental results demonstrate the effectiveness of our approach.

© 2017 Published by Elsevier Ltd.

1. Introduction

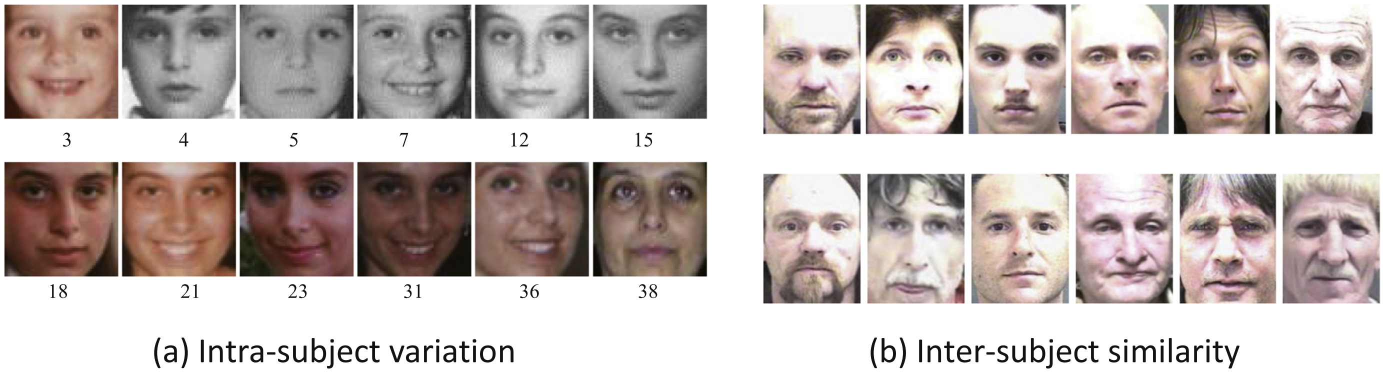
Age-related research has become attractive in recent years due to its wide range of application scenarios, such as age-based face image retrieval, age-specific human-computer interaction, automatic face simulation, security surveillance monitoring and intelligent advertisement systems.

Despite the great advances in face-related works in recent years, face recognition across age remains a challenging problem. The challenges include large intra-subject variation and great inter-subject similarity [1]. The human facial appearance changes greatly with the aging process. From birth to adulthood, the greatest change is craniofacial growth, which involves a change in shape; while from adulthood to old age, the most perceptible change turns skin aging, which involves texture change [2]. The changes of the same person are the intra-subject variations. Meanwhile, different persons in same age period may look similar, which is the inter-subject similarity. Fig. 1 illustrates the intra-subject variations and inter-subject similarity.

Therefore, reducing the intra-subject variations while enlarging the inter-subject differences is a crucial goal in metric-based age-invariant recognition. There are several traditional approaches have realized this goal for general face recognition, such as linear discriminant analysis (LDA) [4], Bayesian face [5,6] and metric learning [7]. However, they are limited by their linear nature. Some improvements have been made to address this limit. Li et al. [8] proposed learning a locally adaptive decision function for the face-matching problem, but this model is limited by its shallow structure.

[*] Corresponding authors.
E-mail addresses: liya@gzhu.edu.cn (Y. Li), wangq79@mail.sysu.edu.cn (Q. Wang).

https://doi.org/10.1016/j.patcog.2017.10.015
0031-3203/© 2017 Published by Elsevier Ltd.



(a) Intra-subject variation

(b) Inter-subject similarity

Fig. 1. Illustration of intra-subject variations and inter-subject similarity. (a) Shows changes in the facial appearance of one person with the aging process. Images come from the FG-NET database [3]. (b) Shows that different persons in the same age period may look similar.

In contrast, the conventional pipeline of face recognition usually includes two basic steps: extracting features and measuring similarity for matching. Some hand crafted features are extracted first in face recognition, such as Gabor wavelets, SIFT [9], LBP [10], HOG [11], PCANet [12], HD-LBP [13]. Then, a certain distance function is used to measure the similarity. To make the features comparable, an extra projection step in which the features are projected into a common space is added between the basic steps in some works. Altogether, models with these two or three separate steps ignore the interaction between the feature representation and the similarity measure.

In this paper, we present a novel similarity measure and a distance metric optimization driven learning approach that integrates feature leaning and distance metric learning via a deep convolutional neural network(CNN) to address the issues mentioned above. The key idea of metric learning methods is to learn a parametric distance/similarity metric between two images x and y. In most cases it takes the Mahalanobis distance metric in form of $s(x,y) = (x-y)^T\mathbf{M}(x-y) = x^T\mathbf{M}x - 2x^T\mathbf{M}y + y^T\mathbf{M}y$, where $\mathbf{M}$ is a positive semi-definite (PSD) matrix. Considering the metric itself is quadratic and similarity is symmetric, we define our similarity model as a binary quadratic polynomial, which is in a more generalized form and can be obtained by the translation of quadratic form. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors for images x and y. Our similarity model is defined as:

$$S(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} + \mathbf{x}^T\mathbf{B}\mathbf{y} + \mathbf{c}^T(\mathbf{x}+\mathbf{y}) + b, \tag{1}$$

where $\mathbf{A}$ is a positive semi-definite (PSD) matrix, $\mathbf{B}$ is negative semi-definite (NSD) matrix, $\mathbf{c}$ is a vector and b is the bias term. Factoring $\mathbf{A}$ and $\mathbf{B}$ as $\mathbf{L}_\mathbf{A}^T\mathbf{L}_\mathbf{A}$ and $-\mathbf{L}_\mathbf{B}^T\mathbf{L}_\mathbf{B}$, our final similarity model can be written as:

$$\begin{aligned} S(\mathbf{x},\mathbf{y}) \quad &= \frac{1}{2}(\mathbf{L}_\mathbf{A}\mathbf{x})^T(\mathbf{L}_\mathbf{A}\mathbf{x}) + \frac{1}{2}(\mathbf{L}_\mathbf{A}\mathbf{y})^T(\mathbf{L}_\mathbf{A}\mathbf{y}) \\ &-(\mathbf{L}_\mathbf{B}\mathbf{x})^T(\mathbf{L}_\mathbf{B}\mathbf{y}) + \mathbf{c}^T\mathbf{x} + \mathbf{c}^T\mathbf{y} + b. \end{aligned} \tag{2}$$

Thus, the image-matching problem is cast into the problem of computing the decision function, Eq. (2), after learning $\mathbf{L}_\mathbf{A}$, $\mathbf{L}_\mathbf{B}$, $\mathbf{c}$, and b. The components of $\mathbf{L}_\mathbf{A}\mathbf{x}$, $\mathbf{L}_\mathbf{B}\mathbf{x}$ and $\mathbf{c}\mathbf{x}$ for x and $\mathbf{L}_\mathbf{A}\mathbf{y}$, $\mathbf{L}_\mathbf{B}\mathbf{y}$ and $\mathbf{c}\mathbf{y}$ for y can be modelled as the weights of the fully connected layer that connect the neurons in deep CNN. Therefore, feature learning and distance-metric learning are integrated seamlessly in an end-to-end way.

Given labelled training images, we first generate a large number of pairs, where matched and unmatched pairs are in a certain proportion. We call two images matched if they belong to the same person. For matched pairs, we try to select as many different age instances as possible for each person to learn the identification information that is not affected by age. Then, taking these pairs as input, we aim to enlarge the differences between the unmatched pairs while reducing the variations between matched pairs, and we update the model parameters using the mini-batch stochastic gradient descent (SGD) algorithm. Note that the similarity components of $\mathbf{x}$ and $\mathbf{y}$ in Eq. (2) do not interact with each other, so we can compute $[\mathbf{L}_\mathbf{A}, \mathbf{L}_\mathbf{B}, \mathbf{c}]^T\mathbf{x}$ and $[\mathbf{L}_\mathbf{A}, \mathbf{L}_\mathbf{B}, \mathbf{c}]^T\mathbf{y}$ separately. Furthermore, we calculate the gradients based on the small number of images in each batch instead of the large number of image pairs. We know that one image can be assigned to more than one training pairs, and in sample-based way the computational cost is reduced due to avoiding the repeated calculation of gradients.

Instead of using handcrafted features, deep CNN learns features from raw images directly, and it has made impressive progress in image classification, object detection, semantic segmentation, scene labelling, face recognition, person re-identification and many other vision tasks. Among the deep CNN approaches Deep Ranking [14], deep Filter-Pairing Neural Network (FPNN) [15] and the method proposed in paper [1] are closest to our model. They are also joint models, but there are obvious differences compared with ours.

Our model is different from the Deep Ranking model, which takes *triplets* as input. The data structure of the triplet is $\langle p_i, {p_i}^+, {p_i}^- \rangle$, where ${p_i}^+$ is the positive sample, and it shares the same class label as the probe sample p_i, and ${p_i}^-$ is the negative sample which include two types: one type is the out-of-class negative sample, which is in a different category than p_i, and the other type is the in-class negative sample, which is in a same category as p_i but is less relevant to p_i than ${p_i}^+$. In spite of the good performance achieved by learning fine-grained image similarities, Deep Ranking is not optimal for age-invariant face recognition. Because many factors, such as expression and head pose influence the appearance, two images of one person with a bigger age gap can look more similar than two images with a lower age gap. In this circumstance, the relative distance between the positive and negative samples is inaccurate, which will affect the robustness accordingly.

Our model is different from FPNN, although FPNN also uses image pairs as input. First, our model uses only one network to extract features, whereas FPNN uses two networks. Second, in FPNN a *patch-matching-layer* is added to address the large differences in body parts of pedestrians by dividing pedestrian images into several horizontal stripes and matching image patches only within the same stripe; however, this is unnecessary for face images. Third, we use the fully connected layer to learn a distance metric, whereas FPNN uses it to learn a binary determination.

Different from our work, paper [1] models the similarity limited to the thresholds. The fluctuation of the thresholds definitely affects the judgement of whether two faces are from the same person. On the contrary, our formulation learns a more general distance metrics, which is not be influenced by the thresholds. In addition, the similarity function of the newly proposed method can be modelled as the weights of fully connected layer, therefore, feature learning and metric learning are integrated together in an end-to-end way. Furthermore, we have tried different network structures, and the experimental results show that our method is also applicable in deeper network; while the work [1] have not demonstrated accuracy gains with increased depth. Finally, we evaluate the generalization ability of our approach on more databases than [1].

There are three major contributions in our work. (1) We propose a new distance metric optimization driven deep-learning model for age-invariant face recognition, that learns feature representations and the similarity measure simultaneously in an end-to-end way. (2) We train the proposed network using a novel optimization method and carefully designed training strategies. (3) We evaluate our model on several databases for age-invariant face recognition, and the experimental results demonstrate that the performance of our method is close to those of other state-of-the-art approaches.

2. Related work

In the past, most age-related works focused on age estimation [3,16–23], including exact-age estimation and age-group estimation. In recent years, many works focused on age-invariant recognition. In [24], Park et al. proposed a 3D facial-aging model and simulation method, which used a generative approach and tried to compensate for the lack of age information for the 3D facial-images before recognition. Li et al. [25] proposed a multi-feature discriminant-analysis (MFDA) method, in which each face was represented by patch-based SIFT and LBP descriptors and faces were recognized using a variation of random-subspace LDA. Sungatullina et al. [26] also used the multi-feature descriptor and further proposed a multi-view discriminative-learning (MDL) method. Bereta et al. [27] quantified abilities of local descriptors commonly used in face recognition, and gave a comparison among them in the context of age discrimination. Gong et al. [28] proposed a hidden-factor analysis (HFA) method, in which they supposed that there are two hidden factors, age and identity, that influence facial appearance. They further proposed a maximum-entropy feature descriptor (MEFD) using the identity-factor analysis (IFA) matching method in [29] to improve the HFA. It is worth noting that in 2014, Chen et al. [30,31] released a new age-related face dataset named CACD; since then, almost all the works on age-invariant face recognition have performed their experiments using it. Bouchaffra [32] introduced a novel formalism that performed dimensionality reduction and captured topological features to conduct pattern classification. More recently, several works have applied the hierarchical and deep models for age-invariant face recognition. Li et al. [33] proposed a two-level hierarchical-learning model with a new feature descriptor called local pattern selection (LPS) to address this problem. Two newer works [34,35] adopting deep models both represented the face images using three components as [28]; paper [34] used a coupled auto-encoder networks (CAN); and paper [35] used CNNs to obtain identity features for face recognition. Although the work in [35] used CNN, the parameters in the convolutional unit and in the fully connected layer were updated separately; to some extent, the CNN was only used to extract features. Our work is different from this in an essential aspect: in our model, feature learning and distance-metric learning are integrated seamlessly in an end-to-end way and the parameters are updated jointly via SGD.

Metric learning has been a hot research topic, influenced by the pioneering work of Xing et al. [36]. Metric-learning approaches usually focus on the linear metric Mahalanobis distance $d(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{M}(\mathbf{x}-\mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are feature vectors of images x and y, respectively, and $\mathbf{M}$ is a PSD matrix. One of the most widely used Mahalanobis-distance learning methods is Large-Margin Nearest Neighbours (LMNN), introduced by Weinberger et al. [37], which defines the constraints in a local way and has many extensions. Information-Theoretic Metric Learning (ITML) is another important work, proposed by Davis et al. [38], which uses LogDet divergence regularization to learn Mahalanobis distance. Qi et al. [39] exploited the sparse nature of the high-dimensional feature space and considered the case of high-dimensional data together with few training samples, for which they used LogDet divergence and L_1 regularization together. Mirror-Descent Metric Learning (MDML), proposed by Kunapuli et al. [40], is a general online Mahalanobis-distance learning framework. Non-Mahalanobis based metric-learning methods have also been proposed, such as kernelizations of the linear method [41] and the nonlinear method [42]. Deep CNN has already made impressive progress on many computer vision problems, and recently, several deep-CNN based models have been explored for metric learning. Deep Ranking, proposed by Wang et al. [14], learns fine-grained image similarity; and FPNN, proposed by Li et al. [15], handles misalignment and photometric and geometric transforms via deep networks. Our deep model is partially motivated by these works but differs from them.

3. Our approach

3.1. Optimization objective

Our approach is to seek a function $S(\mathbf{x},\mathbf{y})$ as shown in Section 1. It is worth noting that $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors for images x and y, respectively, and that instead of learning a similarity function on hand-crafted features, we integrate feature learning and distance-metric learning seamlessly via a deep model. Recall that if we learn $\mathbf{L_A}$, $\mathbf{L_B}$ and $\mathbf{c}$ in Eq. (2), the similarity components of x and y can be obtained, and the matching problem is solved accordingly. We denote $[\mathbf{L_A}, \mathbf{L_B}, \mathbf{c}]^T$ by $\mathbf{\Phi}$ and the parameters of the feature representation by $\mathbf{W}$.

The objective of age-invariant face recognition is as follows. For an age instance z of person P, we wish to learn a reidentification model to successfully identify another age instance z' of the same person P; that is, we wish to learn the identification information that is not affected by age. Therefore, we select as many different age instances as possible for each person to construct training-image pairs to make the model more robust. A positive sample is an image pair of the same person at different ages, and a negative sample is an image pair of different people.

Given a training set $Z = \{(z_i, y_i)\}_{i=1}^{N}$, where N is the total number of images, and y_i is the person ID, i.e., class label, we define an image-pair set $\Omega = \{\Omega_k = (z_i, z_j)\}$, where Ω_k is the kth image pair in Ω. Define $l(\Omega_k)$ to be the label of image pair Ω_k: $l(\Omega_k) = 1$, if z_i and z_j come from the same person, otherwise, $l(\Omega_k) = -1$. That is,

$$l(\Omega_k) = l(z_i, z_j) = \begin{cases} 1, & \text{if } l(z_i) = l(z_j) \\ -1, & \text{otherwise} \end{cases}. \tag{3}$$

Taking the image-pair set Ω as input, we can make the $S(\Omega_k)$ for the positive sample as large as possible while making it as small as possible for the negative sample by maximizing the sum of $l(\Omega_k) \times S(\Omega_k)$. An image pair with very large similarity has little effect on the model's improvement. The extreme case is when the two images in a pair are the same image. Thus, we can define a certain threshold and ignore pairs for which the similarity is greater than the threshold. We define the threshold as 1; here, the choice of the constant 1 is arbitrary and is not important (any other positive constant c only results in the elements in the matrices being multiplied by c). Our hinge-loss-like objective function is

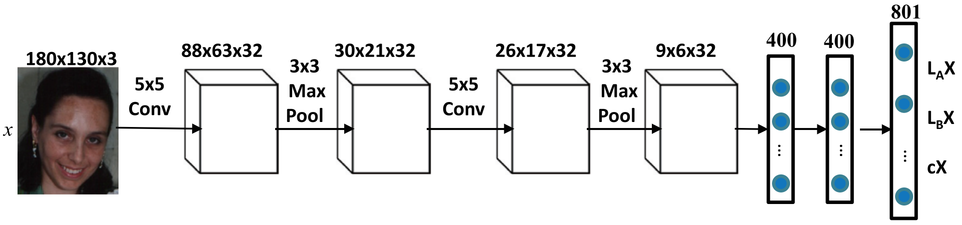


Fig. 2. Architecture of our model. There are two convolution-pooling layers and three fully connected layers. We use batch normalization and ReLU for neurons after the convolutional layers.

Table 1
The configuration of our model. It takes the $180\times130\times3$ images as input and generates the 801-dimensional similarity component for recognition. Note that BN and FC are the abbreviations of batch normalization and fully connection, respectively.

layer type	Kernel size/stride	Output size
Convolution	$5\times5/2$	$88\times63\times32$
BN & ReLU	–	$88\times63\times32$
Max-pooling	$3\times3/3$	$30\times21\times32$
Convolution	$5\times5/1$	$26\times17\times32$
BN & ReLU	–	$26\times17\times32$
Max-pooling	$3\times3/1$	$9\times6\times32$
FC	–	$1\times1\times400$
FC	–	$1\times1\times400$
FC	–	$1\times1\times801$

$$H(\mathbf{W},\boldsymbol{\Phi}) = \sum_{\Omega}\max\{0, 1 - l(\Omega_k)\times S(\Omega_k)\} + \zeta(\mathbf{W},\boldsymbol{\Phi}) \qquad k = 1, 2, \ldots, N^2. \tag{4}$$

The first term is the empirical loss, which is occasionally written as $\sum_{\Omega}[1 - l(\Omega_k)\times S(\Omega_k)]_+$. The empirical-loss function takes a positive value, and if $1 - l(\Omega_k)\times S(\Omega_k) \le 0$, the empirical loss is 0. The second term is a regularization term, $\zeta(\mathbf{W},\boldsymbol{\Phi}) = \frac{\lambda}{2}(\|\mathbf{W}\|^2 + \|\boldsymbol{\Phi}\|^2)$, which denotes the regularizer on the parameters of the similarity model and feature representation. To simplify the notation, we denote by $L(\mathbf{W},\boldsymbol{\Phi})$ the first item's empirical loss, and the above objective function can be further simplified:

$$H(\mathbf{W},\boldsymbol{\Phi}) = L(\mathbf{W},\boldsymbol{\Phi}) + \zeta(\mathbf{W},\boldsymbol{\Phi}). \tag{5}$$

3.2. Deep architecture

Our model learns the features and distance metric jointly as described above. The combination is achieved by deep CNN, where $\mathbf{L_A}$, $\mathbf{L_B}$ and $\mathbf{c}$ can be considered as the weights of the fully connected layer, and all parameters can be optimized using network-propagation algorithms.

Due to the small number of images in public datasets with person-age information, the network that we used is relatively shallow. Fig. 2 illustrates the overall network architecture. It takes the image pairs as input and contains two convolution-pooling layers and three fully connected layers. Table 1 shows the network configuration. Note that after the convolutional layer, we adopt batch normalization and rectified linear units (ReLU) for neurons. The metric learning is implemented by the third fully connected layer with 801 dimensional outputs, which learns matrices $\mathbf{L_A}$ and $\mathbf{L_B}$ and vector $\mathbf{c}$ simultaneously. For the input images x and y, their outputs of the whole network are $[\mathbf{L_A x}, \mathbf{L_B x}, \mathbf{cx}]^T$ and $[\mathbf{L_A y}, \mathbf{L_B y}, \mathbf{cy}]^T$, where $\mathbf{x}$ and $\mathbf{y}$ are the second fully connected layer's activations, respectively. In this way, the metric learning is tightly integrated with the feature learning, and they can be jointly optimized during model training. For simplicity of notation, we denote the output layer's activations of images x and y by $\tilde{x}$ and $\tilde{y}$, that is, $\tilde{x} \triangleq [\mathbf{L_A x}, \mathbf{L_B x}, \mathbf{cx}]^T$ and $\tilde{y} \triangleq [\mathbf{L_A y}, \mathbf{L_B y}, \mathbf{cy}]^T$. The value of b affects only the learning convergence in our experiments and has no impact on the matching performance, so we empirically set $b = -1.9$ in our experiments. By substituting the output into Eq. (2), we can obtain the matching result.

3.3. Training strategies

3.3.1. Mini-batch selection

We apply a mini-batch learning strategy to optimize the parameters, due to insufficient memory. For q training images, the number of pairs is $O(q^2)$. It is impossible to load all image pairs into memory, even for a moderately sized dataset. Therefore, a subset of the pairs must be chosen for use in batch process. A simple way to select the pairs is to generate pairs randomly. However, this method results in far more the negative pairs (unmatched pairs) than positive pairs (matched pairs). For example, given a labelled dataset with p persons in which each person has u images on average, the number of negative pairs is p^2 times the number of positive pairs. Random pair selection causes the positive and negative training samples to be seriously poorly-balanced.

To solve this problem, we propose the following pair-generation rule. In each iteration, we select a fixed number of persons and generate image pairs using only these persons. We ensure exhaustive positive pairs and randomly select a certain ratio of negative pairs.

3.3.2. Data augmentation

We apply two forms of data augmentation to reduce overfitting. The first form of data augmentation involves extracting patches for positive pairs and training the network by using these patches. The patch width and height are set to a certain percentage of the dimensions of the original input image. For example, if the input image size is 150×200 and the percentage is 90%, the patch size is 135×180, and the number of matched image pairs is increased by a factor of 300. The second form of data augmentation is horizontal reflection on positive image pairs. After these two forms of augmentation, the number of positive pairs is increased by a factor of 600, thus alleviating the over-fitting problems to some extent.

3.4. Learning algorithm

In this section, we discuss the learning method for our distance metric optimization model's training. It is impossible to load all images into memory, so we use the mini-batch learning approach; that is, in each training iteration, a small subset of the image pairs is fed into the neural network for model optimization.

To incorporate $\tilde{x}$ and $\tilde{y}$ into the similarity model, given in Eq. (2), we need three transformation matrices:

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{I}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{(2r+1)\times(2r+1)}$$

$$\mathbf{P}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 0 \end{bmatrix}_{(2r+1)\times(2r+1)}$$

$$\mathbf{P}_3 = \begin{bmatrix} \mathbf{0}_{2r\times 2r} & 0 \\ 0 & 1 \end{bmatrix}_{(2r+1)\times(2r+1)},$$

where **I** is the identity matrix and r is the dimension of the second fully connected layer's output. Thus, Eq. (2) can be written as

$$\begin{aligned} \tilde{S}(\mathbf{x},\mathbf{y}) &= \tfrac{1}{2}(\mathbf{P}_1\tilde{x})^T(\mathbf{P}_1\tilde{x}) + \tfrac{1}{2}(\mathbf{P}_1\tilde{y})^T(\mathbf{P}_1\tilde{y}) \\ &\quad -(\mathbf{P}_2\tilde{x})^T(\mathbf{P}_2\tilde{y}) + (\mathbf{P}_3\tilde{x}) + (\mathbf{P}_3\tilde{y}) + b. \end{aligned} \tag{6}$$

Suppose that Ω' is the image-pair set of a mini-batch, and the total number of pairs is n. We know that Ω' is a subset of Ω. Recall that $l(\Omega_k)$ is the label of image pair Ω_k, so $l(\Omega'_k)$ is the label of the kth element in Ω', which we denote by l_k for simplicity. Incorporating Eq. (6) into $L(\mathbf{W}, \mathbf{\Phi})$, we have the following pair-based form of the empirical loss for a mini-batch.

$$\begin{aligned} L(\mathbf{W},\mathbf{\Phi}) &= \sum_{k=1}^{n}\Big\{1 - l_k \times \Big[\tfrac{1}{2}(\mathbf{P}_1\tilde{x}_k)^T(\mathbf{P}_1\tilde{x}_k) \\ &\quad +\tfrac{1}{2}(\mathbf{P}_1\tilde{y}_k)^T(\mathbf{P}_1\tilde{y}_k) - (\mathbf{P}_2\tilde{x}_k)^T(\mathbf{P}_2\tilde{y}_k) \\ &\quad +(\mathbf{P}_3\tilde{x}_k) + (\mathbf{P}_3\tilde{y}_k) + b\Big]\Big\}_+, \end{aligned} \tag{7}$$

where $<x_k, y_k>$ is the kth image pair in Ω', $\tilde{x}_k$ and $\tilde{y}_k$ are the outputs of the two images. Given $\mathbf{G} = (\mathbf{W}, \mathbf{\Phi})$, Eq. (5) can be written as

$$H(\mathbf{G}) = L(\mathbf{G}) + \zeta(\mathbf{G}). \tag{8}$$

Under the mini-batch training strategy, the objective can be optimized using the mini-batch back propagation algorithm. The key step is updating the model parameters by stochastic-gradient descent(SGD). The formulation is

$$\mathbf{G} = \mathbf{G} - \alpha\frac{\partial}{\partial\mathbf{G}}\mathbf{H}(\mathbf{G}), \tag{9}$$

where α is the learning rate.

The key problem in solving the above equation is calculating $\frac{\partial}{\partial\mathbf{G}}\mathbf{L}(\mathbf{G})$. A straightforward method to obtain this gradient is to calculate the sum of all of the pairs' gradients. However, in this pair-based approach, the gradient of an image appears in more than one pair is considered repeatedly. It is inefficient without gradient sharing. Fortunately, the loss function for each pair is defined by two images' outputs, so we can optimize this process by computing the gradients of distinct images rather than image pairs. Thus, in this sample-based way, we can reuse the forward and backward propagations of images that appear in several pairs. Calculating the gradient for all pairs is equivalent to summing up the gradient of all samples. Therefore, we rewrite Eq. (7) in a sample-based form:

$$\begin{aligned} L(\mathbf{W},\mathbf{\Phi}) &= \sum_{i=1}^{m}\sum_{k=1}^{t}\{1 - l_k \times [\tfrac{1}{2}(\mathbf{P}_1\tilde{z}_i^k)^T(\mathbf{P}_1\tilde{z}_i^k) \\ &\quad +\tfrac{1}{2}(\mathbf{P}_1\tilde{z}_j^k)^T(\mathbf{P}_1\tilde{z}_j^k) - (\mathbf{P}_2\tilde{z}_i^k)^T(\mathbf{P}_2\tilde{z}_j^k) \\ &\quad +(\mathbf{P}_3\tilde{z}_i^k) + (\mathbf{P}_3\tilde{z}_j^k) + b]\}_+, \end{aligned} \tag{10}$$

where $<z_i^k, z_j^k>$ is the kth pair containing image z_i in Ω', $\tilde{z}_i^k$ and $\tilde{z}_j^k$ are the outputs of the two images, m denotes the number of different images in a mini-batch, t denotes the number of pairs containing image z_i, and $i, j = 1, \ldots, m,\ k = 1, \ldots, t$. Note that the subscript i of z_i^k denotes the index of samples in the mini-batch and the superscript k of z_i^k denotes the index of sample pairs containing image z_i.

Furthermore, we denote by $\mathbf{G}^{(l)}$ the parameters in lth layer, by ${Z_i}^l$ the feature tensor in the lth layer for image z_i and by $\tilde{z}_i$ the final output. Therefore, the gradients in all layers for image z_i can be calculated using the following four steps:

1. Perform a forward propagation to compute each layer's activation $\mathbf{Z}_i^{(2)}$, $\mathbf{Z}_i^{(3)}$, and so on, up to the output layer $\tilde{z}_i$.
2. Compute the partial derivatives of the output layer's activation for every sample. Recall that for input image z_i the output is $\tilde{z}_i$, so for each sample z_i, the partial derivatives of its output layer's activation are written as follows.

   $$\delta_i^{(o)} = \frac{\partial L(\mathbf{G})}{\partial\tilde{z}_i} = -\sum_{k=1}^{t} l_k(\mathbf{P}_1\tilde{z}_i - \mathbf{P}_2\tilde{z}_j + \mathbf{P}_3), \tag{11}$$

   where o is the output layer, t is the number of pairs including image z_i, l_k is the pair label, and z_j is the other image in the pair.
3. Compute the partial derivatives of the hidden layers' activation $\delta^{(l)}$, according to the formulation formula.

   $$\delta_i^{(l)} = (\mathbf{G}^{(l)})^T\delta_i^{(l+1)}. \tag{12}$$

4. Compute the partial derivatives $\frac{\partial L(\mathbf{G})}{\partial\mathbf{G}^{(l)}}$ of image z_i. We calculate them using the chain rule $\frac{\partial L(\mathbf{G})}{\partial\mathbf{Z}_i^{(l+1)}}\cdot\frac{\partial\mathbf{Z}_i^{(l+1)}}{\partial\mathbf{G}^{(l)}}$, where $\frac{\partial\mathbf{Z}_i^{(l+1)}}{\partial\mathbf{G}^{(l)}}$ is the next layer's output $\mathbf{Z}_i^{(l)}$, and $\frac{\partial L(\mathbf{G})}{\partial\mathbf{Z}_i^{(l+1)}}$ is the previous layer's partial derivative. Therefore the whole formulation is

   $$\frac{\partial L(\mathbf{G})}{\partial\mathbf{G}^{(l)}} = \delta_i^{(l+1)}({\mathbf{Z}_i}^{(l)})^T. \tag{13}$$

   This process is summarized in Algorithm 1.

Algorithm 1 Calculate each layer's gradient for image z_i using back propagation.

```
Input:
    Image pairs set $\Omega'$ in each mini-batch
Output:
    Each layer's gradient $\frac{\partial L(\mathbf{G})}{\partial\mathbf{G}^{(l)}}$ for image $z_i$
Perform a forward propagation to compute each layer's activation: $\mathbf{Z}_i^{(2)}, \mathbf{Z}_i^{(3)}, \ldots, \tilde{z}_i$.
for all image pairs including image $z_i$ do
    if $\tilde{S}(\Omega'_k) \le 1$ then
        $\delta_i^{(o)} += -l_k(\mathbf{P}_1\tilde{z}_i - \mathbf{P}_2\tilde{z}_j + \mathbf{P}_3)$(the other image in a pair is $z_j$)
    end if
end for
for $l = o - 1 : 1$ do
    Compute hidden layers' partial derivatives: $\delta_i^{(l)} = (\mathbf{G}^{(l)})^T\delta_i^{(l+1)}$
    Compute gradient: $\frac{\partial L(\mathbf{G})}{\partial\mathbf{G}^{(l)}} = \delta_i^{(l+1)}({\mathbf{Z}_i}^{(l)})^T$
end for
```

For m different images in each mini-batch, following the chain rule and summing the gradients for each image, we obtain the gradient for all pairs:

$$\frac{\partial H(\mathbf{G})}{\partial\mathbf{G}^{(l)}} = \sum_{i=1}^{m}\delta_i^{(l+1)}({\mathbf{Z}_i}^{(l)})^T + \lambda\mathbf{G}^{(l)}. \tag{14}$$

Then, we can update parameters using the gradient descent. There are N persons, and the number of images of each person is different. For the balance of positive and negative training samples, we exhaust positive pairs and randomly select a certain ratio of negative pairs. After each image's gradients are obtained by Algorithm 1, the total gradients are computed using Eq. (14), and the parameters can be updated according to Eq. (9). The whole joint learning process is described in Algorithm 2.

Algorithm 2 The whole distance metric optimization driven joint-learning process.

```
Input:
    Training-sample set $U = \{(z_i, l_i)\}$, initial parameters $G$, learning rate $\alpha$
Output:
    Network parameters $\mathbf{G}$
$\Delta \mathbf{G} = 0$
for $t = 1 : T$ do
    Construct mini-batch $\Omega'$
    for all different images $\{z_i\}$ in the mini-batch do
        Compute gradient $\frac{\partial L(\mathbf{G})}{\partial \mathbf{G}^{(l)}}$ using algorithm 1
        $\Delta \mathbf{G} += \frac{\partial L(\mathbf{G})}{\partial \mathbf{G}} + \lambda \mathbf{G}$
    end for
    $\mathbf{G}^t = \mathbf{G}^{t-1} - \alpha \Delta \mathbf{G}$
end for
```

3.5. Age-invariant recognition

Age-invariant recognition is realized based on image matching. For a test image, we first input it into the deep network obtain an 801 dimensional vector, $(L_A x,\ L_B x,\ cx)^T$. Then, we compute the matching values of this probe image with other gallery images by using Eq. (2), assuming that all gallery images have passed through this network and the outputs were obtained in advance. By sorting the match values we obtain the top-k results, and by counting the top-1 matching results of the probe images we achieve strict accuracy. We observed that the feature representation and similarity components of the gallery images stored in the database can be precomputed because the matching images in one pair are independent, as shown in Eq. (2). This property makes the matching task efficient.

4. Experiment

We evaluate the performance of our approach on three different databases, they are the MORPH-II[1] database [43], the CACD[2] database [30] and the FG-NET database [3].

4.1. Databases and methodology

MORPH-II contains more than 55,000 face images of more than 13,000 individuals, whose ages range from 16 to 77. The average number of images per individual is 4. The training set consists of 20,000 face images from 10,000 subjects, with each subject having two images with the largest age gap. The test data are composed of a probe set and a gallery set from the remaining 3000 subjects. The probe set is composed of the oldest face images of each subject. The gallery set is composed of the youngest face images of each subject. This experimental setting is the same as that used in [30] and [28]. There is only one image in the gallery for each subject. We use the Cumulative Match Characteristic (CMC) curve to measure the performance. The CMC curve is plotted using the top-k ranked points $(i, a_i), i = 1, 2, \ldots k,$ where i indicates the top-I points and a_i is the recognition accuracy using the top-I points. The output is accurate if the top-i results include the person in the probe image.

The CACD database is a large-scale database, released in 2014, which contains more than 160,000 images of 2000 celebrities. The images in CACD are challenging because they vary in age, pose, illumination and occlusion. However, images of only 200 celebrities were manually checked originally. The studies using this dataset are based on these previously manually checked images. We extend the set of manually checked images to 500 celebrities. To make the results comparable with those of previous works, we adopt the original 200 individuals and along with 500 new individuals. We use the mean average precision (MAP) to measure the performance (see [31] for the MAP calculation). Based on the CACD database, Chen et al. [31] developed a verification subset called CACD-VS, which contains 2000 positive pairs and 2000 negative pairs. Two images of each positive pair come from the same person but with a large age interval. CACD-VS is used to evaluate the performance of face verification across age, which differs from the famous LFW database. We report verification results with receiver-operating characteristic (ROC) curves, which explores the trade off between falsely accepting unmatched pairs and falsely rejecting matched pairs.

We use FG-NET database to evaluate the generalization ability due to the large age span for each person. It includes 1002 images of 82 individuals, where the ages are distributed in the range from 0 to 69. The age range is more than 40 years for some of the individual.

We first evaluate the performance of our approach on face identification task on the medium scale database MORPH-II by a basic network. Then we conduct face retrieval task on CACD, which is large enough to be trained by a deeper network. Besides the comparison with other methods, we also analyse the effectiveness of our model including the depth and joint architecture. At last, the face verification on CACD-VS and the evaluation of generalization ability on FG-NET are performed using the model trained on CACD.

In our architecture, the initial parameters of the convolutional and fully connected layers are set according to two zero-mean Gaussian distributions with standard deviations of 0.01 and 0.001. We first detect the faces using the method proposed in paper [44] and resize the faces to a uniform size. We set learning rate α and weight decay λ to 0.001 and 0.0005 respectively. How to set them is present in Section 4.3.

4.2. Age-invariant face identification on MORPH

First, we perform experiments to study the influences of different training strategies, and we then compare our method with several state-of-the-art methods.

4.2.1. Influence of mini-batch selection

We perform a set of experiments to study the influence of mini-batch selection. First, we study the influences of the numbers of persons and pairs. The CMC curve is shown in Fig. 3, where the numbers in parentheses in the legend denote the number of persons and the number of pairs. The recognition accuracies of each strategy are obtained after the same number of iterations. In fact, the accuracies will be better, after many more iterations, but here, we terminate the experiments after the same number of iterations to quickly find the optimal strategy.

To balance each person, we average the number of times each person appears in pairs. For example, (60, 4800) denotes a mini-batch that includes 60 persons, each of whom appears in 80 pairs; thus, a total of 4800 pairs are selected. In this set of experiments, the ratio between the numbers of positive and negative samples is fixed at 1:1. We can see from Fig. 3 that the (80, 11200) mini-batch shows the best performance. In fact, in the top-10, the performance improvements achieved by selecting different persons in each mini-batch are relatively small.

We also study the influence of the positive to negative samples ratio. Fix the number of people and the number of pairs number in a mini-batch as (60, 4800) and change the ratio between the number of positive and negative samples. The CMC curve for this

[1] http://www.faceaginggroup.com/morph/.

[2] http://bcsiriuschen.github.io/CARC/.

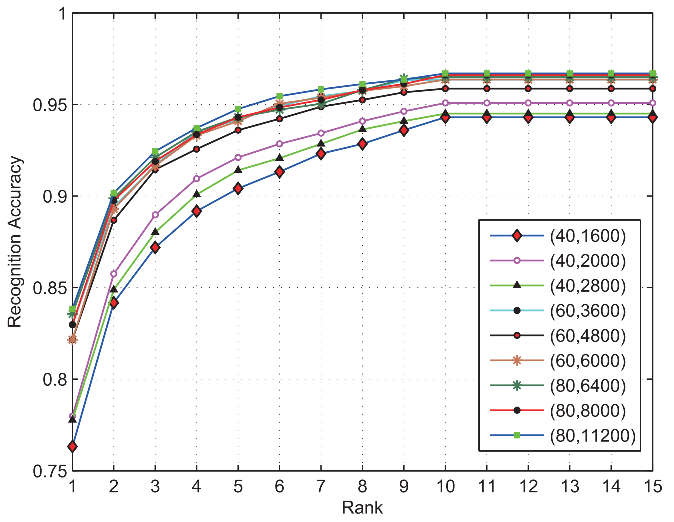


Fig. 3. Influence of the number of individuals and the number of pairs in a mini-batch. The two numbers in parentheses in the legend denote the number of individuals and the number of pairs, respectively.

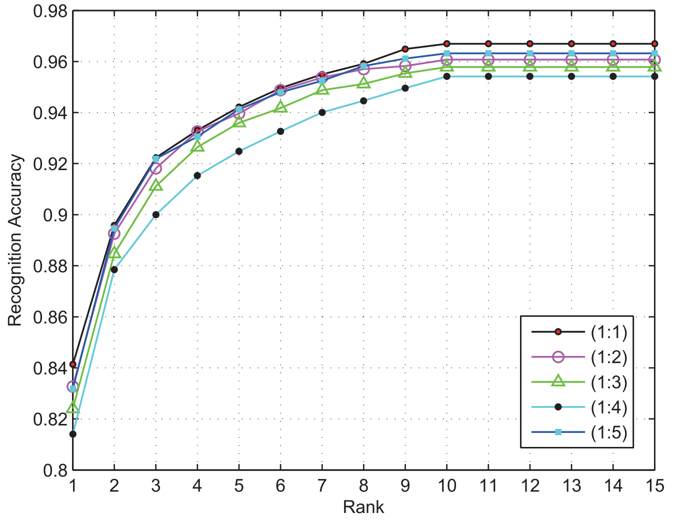


Fig. 4. Influence of positive and negative samples ratio in a mini-batch.

case is shown in Fig. 4. We can see from Fig. 4 that the ratio of 1:1 shows the best performance, which is slightly better than that for the ratio of 1:5.

4.2.2. Comparison with state-of-the-art methods

We compare our deep CNN model with several state-of-the-art methods for age-invariant face recognition on MORPH-II, including MFDA [25], HFA [28], CARC [30], LPS+HFA [33] and the method proposed in paper [24]. The comparison of the results is reported in Table 2. Note that we train our model only on the MORPH dataset for fair comparison, while the newest deep model on age invariant face recognition proposed in [35] is pre-trained on several other face datasets. It is difficult to distinguish the performance improvement comes from the method itself or the large-scale training data with different scale of training set. So we do not compare our results with it. It is encouraging to see that our approach achieves very competitive results compared to other state-of-the-art methods trained only on MORPH. For the top-10 and top-20 cases, our model achieves recognition accuracies of 98.8% and 99.34%, respectively. The CMC curve is shown in Fig. 5. It should be noted that the result is better than the results in Figs. 3 and4 because the recognition accuracies in Figs. 3 and4 are obtained after the same number of iterations for each different strategy in mini-batch selection.

Table 2
Rank-1 identification rates on the MORPH database. Our method achieves the highest recognition rate compared to other state-of-the-art methods.

Method	Recognition rate (%)
Park et al. [24]	79.8
MFDA [25]	83.9
HFA [28]	91.1
CARC [30]	92.8
LPS+HFA [33]	94.8
Ours	**93.6**

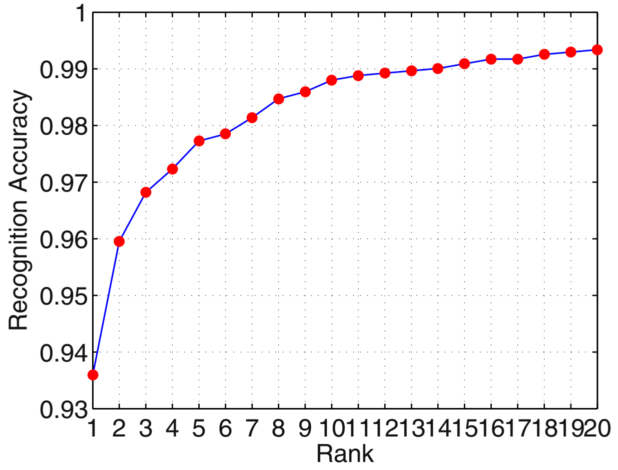


Fig. 5. The CMC curve of our method on the MORPH database.

Table 3
The selection of weight decay λ and learning rate α. It is determined by a set of models trained on a small subset of CACD, where the accuracies are the verification results on the CACD-VS.

Parameters	Values	Iteration times (K)	Accuracy (%)
λ	0.0001	8	70.79
	0.0005	8	71.60
	0.001	12	71.07
α	0.0005	14	70.66
	0.001	8	71.60
	0.005	16	68.43

4.3. Age-invariant face retrieval on CACD

Age-invariant face retrieval means finding those images that are matched to the probe image in gallery, where the age gap between the probe image and the matched images in gallery is large. We use images taken in 2013 as query images and find the same person's images from among the rest of the images; that is, images taken in 2013 are probe images, and the rest of the images are gallery images. The experimental setting is the same as in [30], and we divide the rest of the images into three folders: images taken in 2004–2006, in 2007–2009 and in 2010–2012. We compare our deep-CNN model with several state-of-the-art methods, including CARC [30] and HFA [28].

Parameters setting. The weight decay λ and learning rate α are determined via control variable method. We fix the learning rate $\alpha = 0.001$ first. Given several candidates of λ we train a set of models on a small subset of CACD (only contains ten persons' images) and select the value showing the best performance on validation set CACD-VS. The verification results are shown in Table 3.

Table 4
The influence of the number of training samples. We use images taken in 2013 as query images and retrieve the same person's images in "2010-2012" folder. When gradually increasing the number of training samples from 160 to 460 individuals, the MAPs increase accordingly.

#Training individuals	MAP(2010–2012) (%)
160	62.7
210	64.2
260	65.8
360	68.6
460	71.2

Table 5
Verification accuracy on the CACD-VS. Our method outperforms other state-of-the-art methods.

Method	Verification accuracy (%)
SepaModel	72.8
HD-LBP	81.6
HFA [28]	84.4
CARC [31]	87.6
Ours-2	**89.3**
Ours-3	**91.0**
Ours-2(Aligned)	**89.33**
Ours-3(Aligned)	**91.1**

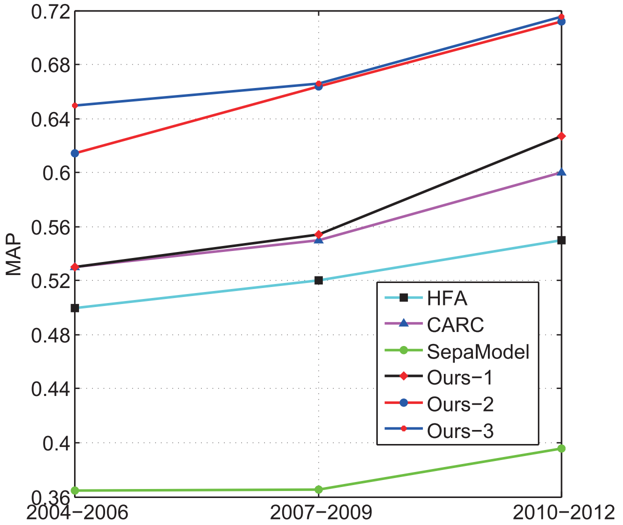


Fig. 6. The retrieval performance of different methods on the CACD database. Our method outperforms the current best-performing method, CARC, significantly by improving the MAP to 71.6%. Ours-1 and Ours-2 are different in the number of training samples. Ours-1 denotes the model trained on small dataset, Ours-2 is trained on large dataset, Ours-3 is the model with deeper network, and SepaModel is the model with two separate training stages of feature and metric.

We can observe that when $\lambda = 0.0005$ it reach the best verification accuracy 71.60%. The selection process of α is similar to λ. We fix the weight decay $\lambda = 0.0005$ and change the values of α, and we select the value showing the best performance on validation set CACD-VS as well. We can observe that when $\alpha = 0.001$ it reach the best performance.

Influence from the number of training samples. Large scaled training samples for deep CNN have been shown in previous work are beneficial to the performance. The following set of experiments demonstrate above viewpoint. We still use images taken in 2013 as query images and retrieve the same person's images in "2010–2012" folder. We gradually increase the number of training samples from 160 to 460 individuals, and the MAPs increase accordingly. Detailed results are shown in Table 4. From the experimental results, we can observe that when the number of training samples added to 460 individuals, the MAPs significantly increased to 71.2%. The red and black lines represented in Fig. 6 illustrate the performance improvement on three folders, where the black line is the result using 200 original manually checked individuals (Ours-1), of which 160 individuals are training samples and the remaining 40 individuals are testing samples, and the red line is the result for 500 individuals (Ours-2), of which 460 individuals are training samples and the remaining 40 individuals are testing samples.

Influence from the depth of network. Deeper neural networks have been shown in previous work are beneficial to the performance. However, it isn't the deeper always better, the accuracy is stagnant or even reduced in some very deep attempts [45]. The CACD database is large enough to be trained by a deeper network. However considering the sufficient pairs in a mini-batch is crucial to our model and deeper network also requires more memory, we use the Alexnet [46] to train a new deeper model to investigate the performance improvement bring by deeper network. The results are reported in Fig. 6, where the blue and red lines present the comparison of different depth. The blue line denotes the result of the deeper model (Ours-3) with same training set as Ours-2. We can observe that on "2004–2006" fold the MAP of deeper model Ours-3 is significantly improved from 0.61 to 0.65, while on "2007–2009" fold the improvement is not obvious. Further, it is obvious that the performance of every method is improve as the age gap is made smaller, which also indicates the undoubted influence of age variation on the retrieval results.

The effectiveness of joint learning. Our approach integrates feature leaning and distance metric learning via a deep CNN in an end-to-end way. We also investigate the effectiveness of our similarity metric learning. We implement a model with the feature learning and distance metric learning stages performed separately. Using the identity information as the supervised signal, we first learn the feature with softmax loss, and then adopt LMNN [37] to learn Mahalanobis distance. We compare our joint model with the separated model and the experiment result illustrates the better performance of our proposed joint model. In Fig. 6, the green line denotes the separated model, where the MAPs are all lower than 0.4 on three folders, especially, on "2004-2006" fold the MAP is only 0.364. We also show some retrieval examples in Fig. 7. The leftmost column shows the probe images. The right fifteen columns correspond to top-15 retrieval results, and the number on top of each column indicates the rank. The incorrect retrieval results are highlighted by red rectangles. We can see that it is completely correct for the top-15 retrieval results of probe images 4, 5, 7, 9, 10 and 14, in which the expression, hairstyle, head pose, presence of moustache and illumination are varied. For probe-3 and probe-8, the retrieval results include image with make-up changes, and for probe-2 and probe-12, the results include images in which the individual is wearing glasses. These visualized results demonstrate that our method has amazing robustness to expression, hairstyle, head pose, moustache, illumination, make-up and wearing glasses.

4.4. Age-invariant face verification on CACD-VS

We use the model trained on CACD to conduct face verification on CACD-VS. We compare our model with CARC [30], HFA [28] and HD-LBP [13]. The verification accuracies of different methods on CACD-VS are given in Table 5. Our method outperforms other state-of-the-art methods, the performance of Ours-3 is 91.0%. Even for our original shallow model Ours-2, the performance of it is also better than other methods. Especially, the accuracy of the separated model we discussed in Section 4.3 is only 72.8%. In addition,

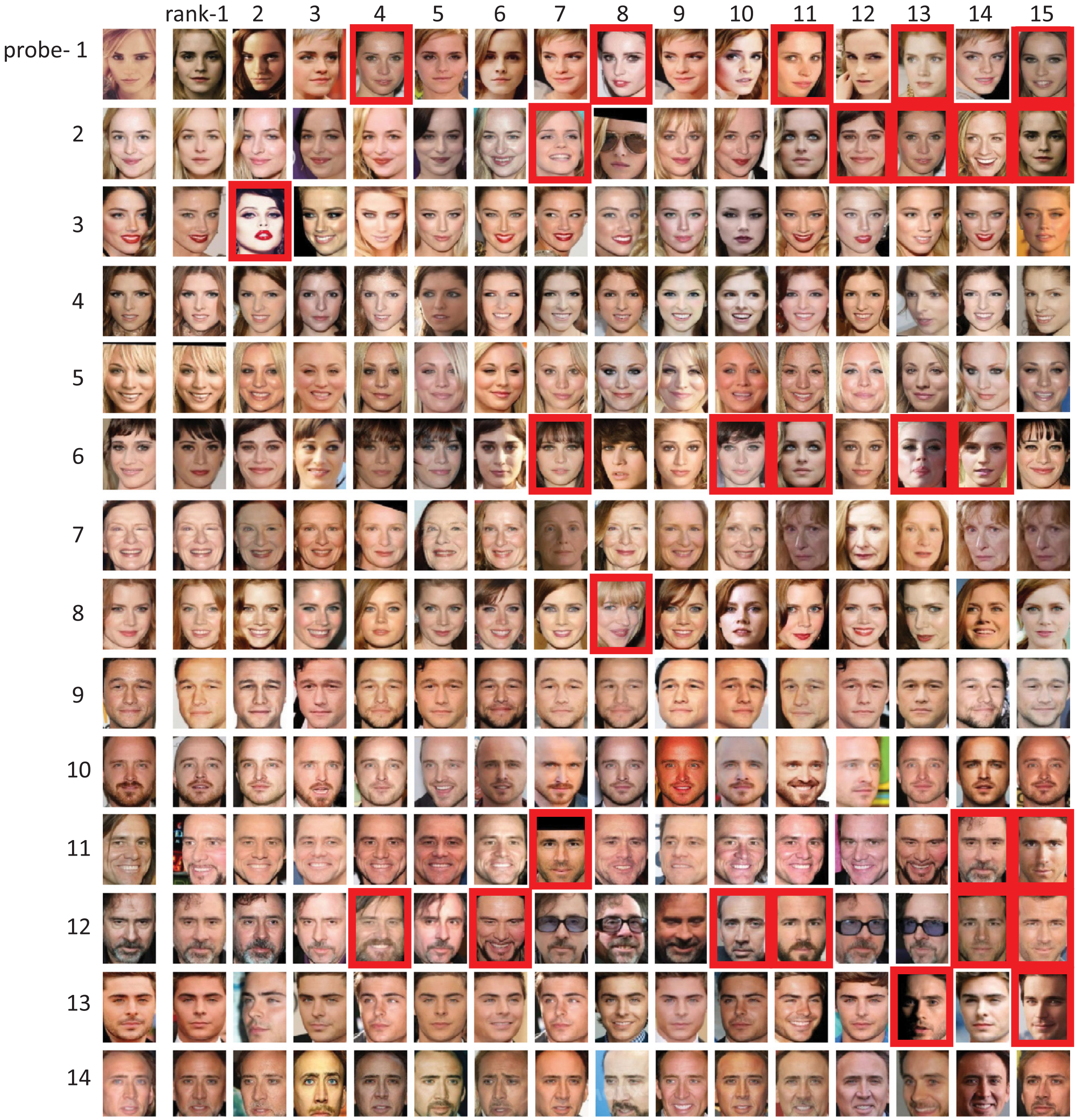


Fig. 7. The retrieval examples of our method on the CACD database. The incorrect retrieval results are highlighted by red rectangles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we also do experiments to investigate whether the preprocessing of face alignment can make the performance better. For testing images we detect the five facial landmarks first, they are two eyes centers, nose tip and the two mouth corners. Then according to the coordinates of the two eyes, we rotate the image make the two eyes line parallel to the horizontal line. Finally, we crop the images to the size of 200 × 150 pixels with two eyes locating at the coordinate (40, 80) and (110, 80). In Table 5, Ours-2(Aligned) and Ours-3(Aligned) are denoted as the model Ours-2 and Ours-3 using aligned testing images, and the accuracies are 89.33% and 91.1% respectively. We can observe that the performance is improved if using aligned testing images. However, the improvement bring by face alignment is not very significant. The accuracy of Ours-2(Aligned) and Ours-3(Aligned) are increased only 0.03 and 0.1% compared with Ours-2 and Ours-3. On the other hand, it confirms the effectiveness of the data augmentation, which makes the model more robust. The ROC curves are shown in Fig. 8. Although our method slightly underperformed the CARC, if the false positive rate (FPR) is less than 0.06, the AUC value of our method is the highest, which is 0.958. Furthermore, if the FPR is greater than 0.06, the true positive rate (TPR) of our method exceeds that of the CARC. We think that it makes sense to allow a certain FPR because generally, the false prediction of samples is better than the omission of positive samples in real applications.

4.5. Generalization ability on FG-NET

The evaluation of generalization ability of our approach are performed on FG-NET due to its large age span for each person. We use the model trained on CACD directly for the task of identifica-

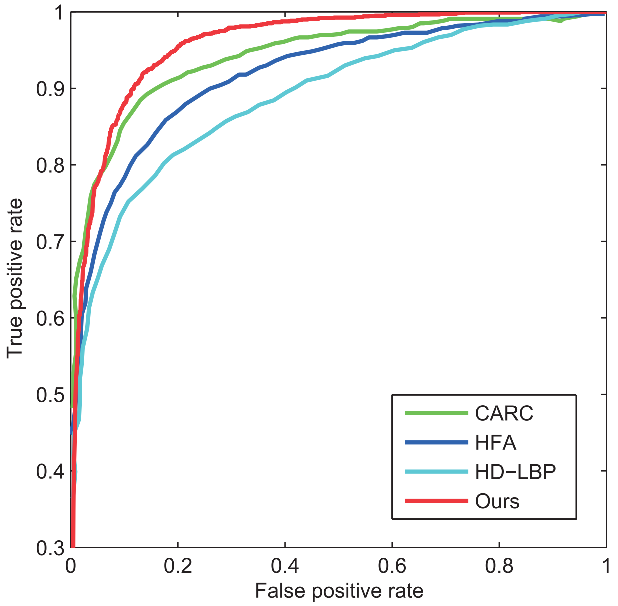


Fig. 8. The ROC curves of different methods on the CACD-VS database. If the FPR is greater than 0.06, the TPR of our method exceeds that of the current best method, CARC.

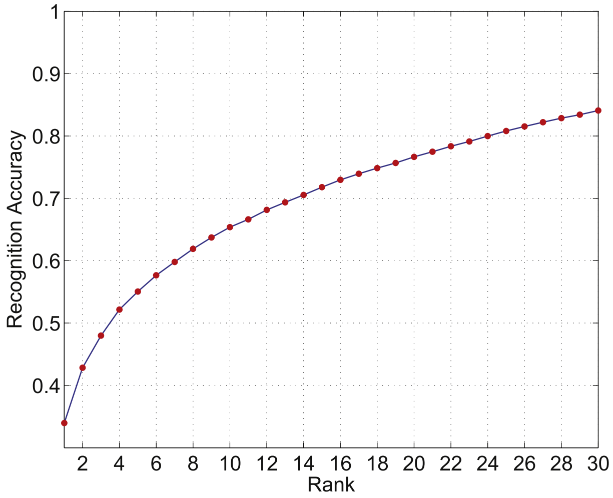


Fig. 9. Extensive study for the generalization ability of verification task for FG-NET using the model train on the CACD dataset.

tion on FG-NET, and the CACD-VS dataset is used as our gallery of distractors. Given a probe image, the gallery contains one image of the same person and distractors. The algorithm rank orders of all images in the gallery based on similarity to the probe. Specifically, the probe set includes *N* people, and for each person there are *M* images. We test each of the *M* images per individual by adding it into the gallery of distractors and use each of the other $M-1$ images as a probe. The experimental setting is the same as that used in [47]. Results are presented using CMC curves. As shown in Fig. 9, the rank-1 accuracy is 33.95%. It does make sense because we only train our model on CACD, while the best result presented in [47] is 38.2%, which is trained on 672 K identities, much more face data than our training samples.

5. Conclusion

In this paper, we propose a new distance metric optimization driven joint-learning model for age-invariant face recognition that can learn features and a distance metric simultaneously. The model's learning is performed in an end-to-end way via a deep-convolutional neural network. The experimental results show that our method achieves very competitive results compared to other state-of-the-art methods on both the MORPH-II,CACD, CACD-VS and FG-NET databases. In fact, age-related research is constrained by insufficient databases. If a database which possesses sufficient number of subjects and has large age spans, the model trained on it will be more robust. In the future, we will focus on building such new suitable database to improve the performance of age-related issues further. Furthermore, we will integrate the semi-supervised learning or active learning with CNN to reduce the annotation effort. We also intend to extend this work to other new applications(e.g. human actions and behaviours recognition).

Acknowledgments

This research is supported by the Research Project of Guangzhou Municipal Universities (No. 1201620302), the Science and Technology Planning Project of Guangdong Province (Nos. 2015B010128009, 2013B010406005). The authors would like to thank the reviewers for their comments and suggestions.

References

[1] Y. Li, G. Wang, L. Lin, H. Chang, A deep joint learning approach for age invariant face verification, in: Proceedings of the Computer Vision CCF Chinese Conference (CCCV), 2015, pp. 296–305.
[2] 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.
[3] 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.
[4] P.N. Belhumeur, J.P. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[5] B. Moghaddam, T. Jebara, A. Pentland, Bayesian face recognition, Pattern Recognit. 33 (11) (2000) 1771–1782.
[6] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: a joint formulation, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2012, pp. 566–579.
[7] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? Metric learning approaches for face identification, in: Proceedings of the International Conference on Computer Vision (ICCV), IEEE, 2009, pp. 498–505.
[8] Z. Li, S. Chang, F. Liang, T.S. Huang, L. Cao, J.R. Smith, Learning locally-adaptive decision functions for person verification, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 3610–3617.
[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[10] 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.
[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2005, pp. 886–893.
[12] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PCANet: a simple deep learning baseline for image classification? IEEE Trans. Image Process. 24 (12) (2015) 5017–5032.
[13] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 3025–3032.
[14] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1386–1393.
[15] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: deep filter pairing neural network for person re-identification, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 152–159.
[16] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2234–2240.
[17] W.L. Chao, J.Z. Liu, J.J. Ding, Facial age estimation based on label-sensitive learning and age-oriented regression, Pattern Recognit. 46 (3) (2013) 628–641.
[18] G. Guo, C. Zhang, A study on cross-population age estimation, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 4257–4263.
[19] X. Geng, C. Yin, Z.-H. Zhou, Facial age estimation by learning from label distributions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2401–2412.

[20] C. Yan, C. Lang, T. Wang, X. Du, C. Zhang, Age estimation based on convolutional neural network, in: Proceedings of the Pacific Rim Conference on Multimedia (PCM), Springer, 2014, pp. 211–220.

[21] C. Li, Q. Liu, J. Liu, H. Lu, Ordinal distance metric learning for image ranking, IEEE Trans. Neural Netw. Learn. Syst. 26 (7) (2015) 1551–1559.

[22] Y. Li, Z. Peng, D. Liang, H. Chang, Z. Cai, Facial age estimation by using stacked feature composition and selection, Vis. Comput. 32 (12) (2015) 1525–1536.

[23] J.K. Pontes, A.S. Britto, C. Fookes, A.L. Koerich, A flexible hierarchical approach for facial age estimation based on multiple features, Pattern Recognit. 54 (C) (2016) 34–51.

[24] U. Park, Y. Tong, A.K. Jain, Age-invariant face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (5) (2010) 947–954.

[25] Z. Li, U. Park, A.K. Jain, A discriminative model for age invariant face recognition, IEEE Trans. Inf. Forensics Secur. 6 (3) (2011) 1028–1037.

[26] D. Sungatullina, J. Lu, G. Wang, P. Moulin, Multiview discriminative learning for age-invariant face recognition, in: Proceedings of the International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–6.

[27] M. Bereta, P. Karczmarek, W. Pedrycz, M. Reformat, Local descriptors in application to the aging problem in face recognition, Pattern Recognit. 46 (10) (2013) 2634–2646.

[28] D. Gong, Z. Li, D. Lin, J. Liu, X. Tang, Hidden factor analysis for age invariant face recognition, in: Proceedings of the International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 2872–2879.

[29] D. Gong, Z. Li, D. Tao, J. Liu, X. Li, A maximum entropy feature descriptor for age invariant face recognition, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5289–5297.

[30] B.-C. Chen, C.-S. Chen, W.H. Hsu, Cross-age reference coding for age-invariant face recognition and retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2014, pp. 768–783.

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

[32] D. Bouchaffra, Nonlinear topological component analysis: application to age-invariant face recognition, IEEE Trans. Neural Netw. Learn. Syst. 26 (7) (2015) 1375–1387.

[33] Z. Li, D. Gong, X. Li, D. Tao, Aging face recognition: a hierarchical learning model based on local patterns selection, IEEE Trans. Image Process. 25 (5) (2016) 2146–2154.

[34] C. Xu, Q. Liu, M. Ye, Age invariant face recognition and retrieval by coupled auto-encoder networks, Neurocomputing (2016) 1–10.

[35] Y. Wen, Z. Li, Y. Qiao, Latent factor guided convolutional neural networks for age-invariant face recognition, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4893–4901.

[36] E.P. Xing, M.I. Jordan, S. Russell, A.Y. Ng, Distance metric learning with application to clustering with side-information, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2002, pp. 505–512.

[37] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.

[38] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of the International Conference on Machine Learning (ICML), ACM, 2007, pp. 209–216.

[39] G.-J. Qi, J. Tang, Z.-J. Zha, T.-S. Chua, H.-J. Zhang, An efficient sparse metric learning in high-dimensional space via l 1-penalized log-determinant regularization, in: Proceedings of the International Conference on Machine Learning (ICML), ACM, 2009, pp. 841–848.

[40] G. Kunapuli, J. Shavlik, Mirror descent for metric learning: A unified approach, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2012, pp. 859–874.

[41] J. Wang, H.T. Do, A. Woznica, A. Kalousis, Metric learning with multiple kernels, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2011, pp. 1170–1178.

[42] D. Kedem, S. Tyree, F. Sha, G.R. Lanckriet, K.Q. Weinberger, Non-linear metric learning, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2012, pp. 2573–2581.

[43] K. Ricanek, T. Tesafaye, MORPH: a longitudinal image database of normal adult age-progression, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR), IEEE, 2006, pp. 341–345.

[44] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 3476–3483.

[45] K. He, J. Sun, Convolutional neural networks at constrained time cost, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5353–5360.

[46] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.

[47] I. Kemelmacher-Shlizerman, S.M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4873–4882.

Ya Li is a lecturer in School of Computer Science and Educational Software, Guangzhou University, Guangzhou, China. She received the B.E. degree from Zhengzhou University, Zhengzhou, China, in 2002, M.E. degree from Southwest Jiaotong University, Chengdu, China, in 2006 and Ph.D. degree from Sun Yat-sen University, Guangzhou, in 2015. Her current research focuses on computer vision and machine learning.

Guangrun Wang received the B.E. degree from the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China, in 2013. He is currently pursuing the Ph.D. degree in the School of Data and Computer Science, Sun Yat-sen University. His research interests include computer vision and machine learning.

Lin Nie received the B.E. degree from Beijing Institute of Technology, Beijing, China, in 2004, M.E. degree from University of California, Los Angeles UCLA, in 2010. Her research interests include computer vision and machine learning.

Qing Wang is an associate professor in School of Data and Computer Science, Sun Yat-sen University. He received the Ph.D. degree from Sun Yat-sen University, Guangzhou, in 2010. His research interests include human computer interaction, computer vision and machine learning. He received the Best Paper Honorable Mention Award in ACM CHI 2010, Google Research/Education Award in 2010, and Google Faculty Award in 2011.

Wenwei Tan is a senior research engineer in Hisilicon Technologies co., LTD. He received the B.S. and M.D. degrees from Guangdong University of Technology, Guangzhou, China, in 2005 and 2008, respectively. His currently research interests include deep learning chip and artificial intelligence.