Neurocomputing 216 (2016) 735–745

Contents lists available at ScienceDirect

Neurocomputing

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

Latent face model for across-media face recognition

Jiang-Jing Lv, Jia-Shui Huang, Xiang-Dong Zhou*, Xi Zhou, Yong Feng

Intelligent Multimedia Technique Research Center, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, PR China

ARTICLE INFO

Article history:
Received 19 January 2016
Received in revised form
22 June 2016
Accepted 9 August 2016
Communicated by Grana Manuel
Available online 13 August 2016

Keywords:
Across-Media Face Recognition
Latent Face Model
Joint Bayesian model

ABSTRACT

Across-media face recognition refers to recognizing face images from different sources (e.g., face sketch, 3D face model, and low resolution image). In spite of promising processes achieved in face recognition recent years, across-media face recognition is still a challenging problem due to the difficulty of feature matching between different modalities. In this paper, we propose a latent face model that creates mappings from a hidden space to different media space. Images from different media of the same person share the same latent vector in hidden space. A coupled Joint Bayesian model is used to calculate the joint probability of two faces from different media. To verify the effectiveness of our proposed method, extensive experiments conducted on various databases: self-collected low-resolution vs. high-resolution database, sketches vs. photos databases, 3D face model vs. photos on LFW database. Experimental results show that our method boosts the performance of face recognition with images from different sources.

1. Introduction

Since face recognition systems have achieved high performance under constrained environment, face recognition applications progress toward unconstrained environment, such as security systems, intelligent visual surveillance systems, immigration automated clearance system, etc. However, face images come from different modalities in some applications. For instance, for the police mug-shot retrieval system, a sketch drawing based on the recollection of an eyewitness is usually used for searching suspects through a photo database; for immigration automated clearance system, the image in the identity card or e-passport microchip is used to compare with the face of the individual to verify that the holder of the passport is the rightful owner, while the identity card face images are all frontal with low-resolution. These face recognition applications which match images from different sources are called across-media face recognition [1]. Across-media face recognition is more difficult than traditional face recognition task. First of all, different from traditional face recognition which works on constrained scenarios, across-media face recognition works on unconstrained scenarios. In addition, due to the textural discrepancies between images with different modalities [2], across-media face recognition has to establish a connection between face images from different media for feature matching.

At present, there exist three ways of addressing across-media face recognition: synthesis based methods, common space projection based methods and feature descriptor based methods. Image synthesis based methods [3–8] tend to transform images into the same modality to make all data comparable. However, the synthesis process is difficult and the fidelity of the synthesized images cannot be guaranteed. Common space projection based methods [9–14] project face images from different sources into a common space to make them directly comparable. But the projection needs large number of training data and sometimes the projected features may lose the information of discrimination which impact recognition performance. Feature descriptor based methods [15–18] represent face images by local feature descriptors which is robust to different modalities and only work to specific situations (e.g., photo-sketch). Therefore, across-media face recognition is still a challenge problem.

Compared with previous machine learning methods, faces under different modalities can be easier recognized by human eyes due to our brain is able to abstract discriminative feature shared by faces with different modalities. Inspired from this phenomenon, we believe there exist mappings, which are capable of mapping images from different scenarios to the invariant feature space. Therefore, in this paper, we propose a novel latent face model for across-media face recognition. As shown in Fig. 1, it is assumed that the representations of the same person from different media (e.g., low-resolution images, 3D face models, and sketches) can be generated by the same underlying latent vector from a hidden space. Specifically, for different types of face media, we can find

* Corresponding author.
E-mail addresses: lvjiangjing@cigit.ac.cn (J.-J. Lv), huangjiashui@cigit.ac.cn (J.-S. Huang), zhouxiangdong@cigit.ac.cn (X.-D. Zhou), zhouxi@cigit.ac.cn (X. Zhou), fengyong@cigit.ac.cn (Y. Feng).

http://dx.doi.org/10.1016/j.neucom.2016.08.036
0925-2312/

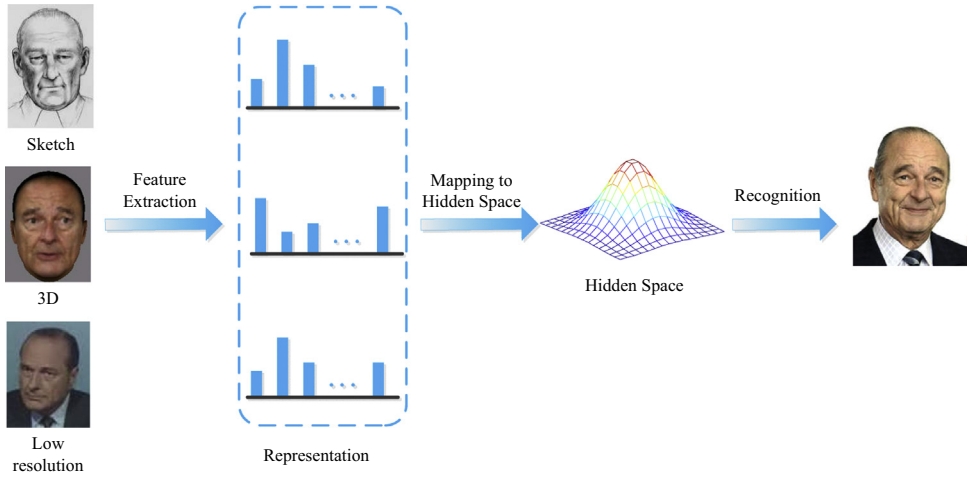


Fig. 1. Procedure of the proposed framework.

the corresponding mappings between the hidden space and different media. After the mapping models between hidden space and media are established, the Bayes' rule and EM algorithm are adopted for solving the parameters. Finally, a Joint Bayesian model is used to calculate joint probability of two faces from different media, meanwhile, inter-personal and intra-personal variations over different media are implicitly learned.

Our method has stronger discrimination ability than Common space projection based methods [9–14] and can be applied to different unconstrained scenarios. The performance of the proposed approach is thoroughly validated on various databases: self-collected low-resolution vs. high-resolution database, sketches vs. photos databases [16,7,19], 3D face model vs. photos on LFW database [20].

The rest of this paper is organized as follows. In Section 2, we briefly review published methods related to across-media face recognition. We detail the proposed latent face model in Section 3. Experimental setup and results are presented in Section 4. We conclude our work in Section 5.

2. Related work

In this section, we briefly review some prior works on across-media face recognition.

Image synthesis was first proposed for across-media recognition, which try to convert one modality to another modality. Tang and Wang proposed eigen-transformation algorithm [5] for holistic mapping, which try to reconstruct face by PCA from training samples using linear mapping. In some cases, holistic mapping cannot be well approximated, therefore nonlinear face mapping between face patches were proposed. In [4], Liu et al. proposed a nonlinear face sketch synthesis and recognition method using local linear embedding between patches. In [3,21], the authors employed embedded hidden Markov model and a selective ensemble strategy to synthesize a sketch. In [7], Wang and Tang proposed a multiscale Markov Random Fields model for face photo-sketch synthesis.

Recently, numerous learning-based approaches [9–14] are used to represent face in different scenes, which aim to minimize the intra-modality difference. In [11], Lin and Tang proposed a common discriminant feature extraction algorithm to map the images from different modality into a common feature space. In [14], Yi et al. employed a canonical correlation analysis correlation mechanism to learn the correlation between NIR and VIS. In [13], Sharma and Jacobs used partial least squares to linearly map heterogeneous faces to a common linear subspace. In [22], Eleftheriadis et al. learned a discriminative manifold shared by multiple views of a facial expression for efficient view-invariant facial expression recognition.

In addition, some feature descriptor based approaches [15–18,23] were also proposed, which are insensitive to changes in modality, but are only suitable to specific situations. In [2], Peng et al. proposed graphical representation heterogeneous face recognition approach, which does not rely on any synthesis or projection procedure but takes spatial information into consideration.

Our approach belongs to learning-based approaches. Contrast to previous methods, which is specific to constrained scenarios with one-versus-one manner, our method is suitable for various unconstrained scenarios. In our model, samples from different media are projected into a discriminant common space, faces with different modalities can be efficiently recognized. Two prior methods [24,25] are similar to our method. In [24], the authors have used factor analysis method to map an idealized identity space to the observed data space where same people with different poses share the same hidden vector, and posterior probabilistic over possible matches is used to calculate the similarity. But it only considered the constrained environment and within-class variations in each scenario that are not efficiently learned. In [25], Chen et al. tried to model the feature representation of face as the sum of inter-personal and intra-personal variations, and the Joint Bayesian was used to calculate joint probability of two faces being the same or different person. However, this method is only suitable to calculate the similarities of faces from the same scenario. In our method, we combine the benefits of these two methods and make some extensions. Firstly, in the generative process, Prince et al. [24] assumed that noise term ε_{ijk} is a zero-mean multivariate Gaussian distribution with an unknown diagonal covariance matrix Σ_k, while in our model ε_{ijk} represents the face variance (e.g., lighting, pose, and expressions) within the same identity and follows Gaussian distribution with a zero-mean and full covariance matrix Σ_k. Secondly, in feature matching process, [25] used images form the same modality to calculate the two joint probabilities, while we extend the Joint Bayesian model for across-media face recognition. Lastly, both prior methods work on constrained scenarios, while our method is suitable for unconstrained scenarios.

3. Latent face model

In this section, we will firstly give a derivation of latent face model, and then illustrate the details about the formulation and optimization, followed by feature matching process.

3.1. Derivation of latent face model

Given a series of a person's images from different sources, we assume that the representations of the same person from different sources can be generated by the same underlying latent vector from a common hidden space. Therefore, the objective is to find the mappings between the hidden space and different sources. Inspired by previous work [24], the mapping model can be defined as:

$$x_{ijk} = F_k h_i + m_k + \varepsilon_{ijk} \tag{1}$$

where x_{ijk} denotes the feature vector of the jth image of individual i in the kth circumstance; h_i denotes the underlying identity variable of individual i; m_k and F_k are the parameters of the offset (average face) and linear function specialized for each circumstance k; and ε_{ijk} represents the face variance (e.g., lighting, pose, and expressions) within the same identity and follows Gaussian distribution with a zero-mean and covariance matrix Σ_k.

The conditional probabilities for the model can be formulated as follows:

$$\Pr(x_{ijk}|h_i) = N_x[F_k h_i + m_k, \Sigma_k] \tag{2}$$

$$\Pr(h_i) = N_h[0, I] \tag{3}$$

where $N_a[b, C]$ denotes a Gaussian in a with mean b and covariance C.

3.2. Model learning

After the model has been defined, we need to find the parameters $\theta = \{F_1, m_1, \Sigma_1, F_2, m_2, \Sigma_2\}$ that maximize the joint likelihood $\Pr(x, h|\theta)$. Unfortunately, identity variables h cannot be observed directly. Therefore, the EM algorithm [26] is adopted, and the objective function can be described as

$$Q(\theta_t, \theta_{t-1}) = \sum_{i=1}^{I} \int \Pr(h_i|x_i, \theta_{t-1}) \times \left[\sum_{j=1}^{J} \sum_{k=1}^{K} \log \Pr(x_{ijk}|h_i, \theta_t) + \log \Pr(h_i) \right] dh_i \tag{4}$$

An iterative strategy is utilized to update h and θ respectively by maximizing (4), where t is the number of iteration. And the first log probability $\log \Pr(x_{ijk}|h_i, \theta_t)$ in the second term can be written as

$$\log \Pr(x_{ijk}|h_i, \theta_t) = \kappa + \frac{1}{2}\left(\log|\Sigma_k^{-1}| - (x_{ijk} - m_k - F_k h_i)^T \Sigma_k^{-1} (x_{ijk} - m_k - F_k h_i) \right) \tag{5}$$

where κ is constant.

The EM algorithm is used for alternately optimization the object function.

E-Step: We fix the parameter set $\theta = \theta_{t-1}$, which are estimated at the previous time $t-1$. The posterior distribution of h_i can be calculated using Bayes's rule

$$\Pr(h_i|x_i, \theta_{t-1}) = \frac{\Pr(x_i|h_i, \theta_{t-1})\Pr(h_i)}{\int \Pr(x_i|h_i, \theta_{t-1})\Pr(h_i) dh_i} \tag{6}$$

For each data from individual i is independent and identically distributed, the first two moments of h_i can be represented as follows:

$$E[h_i|x_i] = \left(I + \sum_{j=1}^{J} \sum_{k=1}^{K} F_k^T \Sigma_k^{-1} F_k \right)^{-1} \cdot \sum_{j=1}^{J} \sum_{k=1}^{K} F_k^T \Sigma_k^{-1} (x_{ijk} - m_k) \tag{7}$$

$$E[h_i h_i^T|x_i] = \left(I + \sum_{j=1}^{J} \sum_{k=1}^{K} F_k^T \Sigma_k^{-1} F_k \right)^{-1} + E[h_i|x_i] E[h_i|x_i]^T \tag{8}$$

M-Step: We fix the identity variables h and optimization θ. We take derivatives of (4) with respect to each F_k and Σ_k^{-1}, then we have:

$$\frac{\partial Q}{\partial F_k} = -\sum_{i=1}^{I} \sum_{j=1}^{J} \int \Pr(h_i|x_i) \cdot \left(\Sigma_k^{-1} (x_{ijk} - m_k - F_k h_i) h_i^T \right) dh_i \tag{9}$$

$$\frac{\partial Q}{\partial \Sigma_k^{-1}} = -\sum_{i=1}^{I} \sum_{j=1}^{J} \int \Pr(h_i|x_i) \times \left(\Sigma_k - (x_{ijk} - m_k - F_k h_i)(x_{ijk} - m_k - F_k h_i)^T \right) dh_i \tag{10}$$

Let $\frac{\partial Q}{\partial F_k} = 0$ and $\frac{\partial Q}{\partial \Sigma_k^{-1}} = 0$, we obtain:

$$F_k = \left(\sum_{i=1}^{I} \sum_{j=1}^{J} ((x_{ijk} - m_k) \cdot E[h_i|x_i]^T) \right) \cdot \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \left((x_{ijk} - m_k) \cdot E[h_i h_i^T|x_i]^T \right) \right)^{-1} \tag{11}$$

$$\Sigma_k = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left[(x_{ijk} - m_k)(x_{ijk} - m_k)^T - F_k E[h_i|x_i] \cdot (x_{ijk} - m_k)^T \right] \tag{12}$$

We plug formula (11) into formula (12). After expanding, we could get another meaningful formula.

$$\underbrace{\Sigma_k}_{S_w} = \underbrace{\frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ijk} - m_k)(x_{ijk} - m_k)^T}_{S_t} - \underbrace{\frac{1}{IJ} X_k C_k^{-1} X_k^T}_{S_b} \tag{13}$$

where

$$X_k = \sum_{i=1}^{I} \sum_{j=1}^{J} \left((x_{ijk} - m_k) \cdot E[h_i|x_i]^T \right)$$

$$C_k = \sum_{i=1}^{I} \sum_{j=1}^{J} \left((x_{ijk} - m_k) \cdot E\left[h_i h_i^T|x_i \right]^T \right)$$

As in our model, we assume Σ_k as intra-personal covariance matrix (within-class scatter matrix) in the kth circumstance. While the first term of formula (13) is covariance matrix (total scatter matrix), we could derive that the second formula (13) is inter-personal covariance matrix (between-class scatter matrix). Therefore, inter-personal covariance is implicitly learned. In addition, our model is trying to learn optimal parameters of Σ_k and F_k, which can capture the distributions of our training data and contribute to discrimination ability of our model. While in [24], Σ_k is only a diagonal covariance matrix, which learns some independent noises in images. Therefore, the method introduced in [24] is only suitable for some constrained environment without large within-class variations.

Sometimes, due to the limited training data set, the covariance Σ_k on the particular circumstance might not be invertible and cannot be appropriate estimated. Therefore, we add a regularization item to formula (12) to guarantee Σ_k as a positive definite matrix:

$$\Sigma'_k = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left[(x_{ijk}-m_k)(x_{ijk}-m_k)^T - F_k E[h_i|x_i]\cdot(x_{ijk}-m_k)^T\right] + \lambda_k \mathbf{I} \tag{14}$$

where $\mathbf{I}$ denotes identity matrix.

By repeating above E-step and M-step until convergence, we can efficiently learn the parameters.

3.3. Feature matching

Given the two vectors x_1 and x_2 from different sources, two alternative hypotheses are tested: H_s that both x_1 and x_2 share the same identity latent variable h_i, or H_d that the vectors were generated from different identity variables h_i and $h_j (i \neq j)$. The verification score can be computed as the log-likelihood ratio for the two hypothesis:

$$score = \log \frac{\Pr(x_1, x_2|H_s)}{\Pr(x_1, x_2|H_d)} \tag{15}$$

The evidence for H_s and H_d are

$$\Pr(x_1, x_2|H_s) = \int \Pr(x_1, x_2|h_i)\Pr(h_i)dh_i \tag{16}$$

$$\Pr(x_1, x_2|H_d) = \Pr(x_1|H_d)\Pr(x_2|H_d) = \int \Pr(x_1|h_i)\Pr(h_i)dh_i \int \Pr(x_2|h_j)\Pr(h_j)dh_j \tag{17}$$

Under H_s hypothesis, the identity x_1, x_2 of the pair are the same. In order to calculate integral in H_s, we reformulate the generative equation as a standard factor analyzer.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} h_i + \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \tag{18}$$

where F_1 and F_2 are linear transformation depend on the scenarios of x_1 and x_2.

This equation now takes the form of a standard factor analyzer, and the solution to the integral is

$$N_{x'}\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} F_1F_1^T + \Sigma_1 & F_1F_2^T \\ F_2F_1^T & F_2F_2^T + \Sigma_2 \end{bmatrix}\right)$$

where $x' = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, Σ_1 and Σ_2 are intra-personal covariance matrixes depend on the scenarios of x_1 and x_2.

Under H_d hypothesis, both the identities and intra-person variations are independent. The likelihood can be expressed in similar format:

$$N_{x'}\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} F_1F_1^T + \Sigma_1 & 0 \\ 0 & F_2F_2^T + \Sigma_2 \end{bmatrix}\right)$$

With the above two conditional joint probabilities, the log-likelihood ratio can be easily computed in closed-form solution.

$$score = \log N_{x'}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{t1} & \Sigma_{b12}^T \\ \Sigma_{b12} & \Sigma_{t2} \end{bmatrix}\right) - \log N_{x'}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{t1} & 0 \\ 0 & \Sigma_{t2} \end{bmatrix}\right) \tag{19}$$

where $\Sigma_{t1} = F_1F_1^T + \Sigma_1$, $\Sigma_{t2} = F_2F_2^T + \Sigma_2$, $\Sigma_{b12} = F_2F_1^T$. Moreover, x_1 can x_2 can be normalized by subtract corresponded mean vector, so by setting m_1 and m_2 to 0 and expanding, we get

$$score = x_1^T Q_1 x_1 + x_2^T Q_2 x_2 + 2x_1^T P x_2 + const \tag{20}$$

where

$$Q_1 = -\Sigma_{t1}^{-1}\Sigma_{b12}^T\left(\Sigma_{t2} - \Sigma_{b12}\Sigma_{t1}^{-1}\Sigma_{b12}^T\right)^{-1}\Sigma_{b12}\Sigma_{t1}^{-1}$$

$$Q_2 = \Sigma_{t2}^{-1} - \left(\Sigma_{t2} - \Sigma_{b12}\Sigma_{t1}^{-1}\Sigma_{b12}^T\right)^{-1}$$

$$P = \Sigma_{t1}^{-1}\Sigma_{b12}^T\left(\Sigma_{t2} - \Sigma_{b12}\Sigma_{t1}^{-1}\Sigma_{b12}^T\right)^{-1}$$

The final formula (20) that does not depend on an estimate of h is used to calculate the similarity score. While in [25], it can only calculate similarities when images form the same modality. On that condition, x_1 can x_2 share the same linear transformation matrix F, which is an inter-personal covariance matrix, and intra-personal covariance matrix Σ.

As for face identification task, each probe image is compared with gallery images to find the most similar one. Therefore, formula (20) is also suit for face identification.

4. Experiments

In this section, we evaluated the performance of the proposed approach on the three different scenarios: low-resolution vs. high-resolution, sketches vs. photos, 3D face model vs. photos. We first evaluated the effectiveness of the proposed method on self-collected low-resolution vs. high-resolution database and investigated the effect of different factors (e.g., data augmentation, training database size, regularization coefficients) on the verification performance. Then we tested our method on the widely studied task of face sketch vs. photo recognition. Finally we verified the generalization ability of our approach on 3D face model vs. photos on LFW.

In our experiments, without explicit declaration, the hidden dimension is default set as 300 and regularization coefficients are default set as 0.001. In addition, our model converges quickly within 20 EM iterations and we set the maximum iteration to 100.

4.1. Low-resolution vs. high-resolution

In this section, we present the performance of the proposed method on self-collected database. The database was divided into a training set and a test set. The training set consists of 2045 subjects and each subject has one low-resolution ID card images and almost 80 high-resolution photo images. It was collected from different schools and factories. The test set consists of 218 subjects and each one has one ID card images and two photo images. The test set was collected from Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences (CIGIT).

The pipeline of image collecting and pre-processing is shown in Fig. 2. In the image collecting phase, ID card images were collected through ID card reader and person photo images were captured by camera. In the image pre-processing phase, face regions were detected by a Viola-Jones [27] based face detector. Then, the SDM algorithm [28] was adopted to detect the facial landmark points. Finally, affine transformation is used to get face image normalized. Examples are shown in Fig. 3. ID card face images are all frontal with low-resolution, while person face images have different poses and occlusions with relatively high-resolution. Moreover, ID card image may be taken several years ago, which looks younger.

During test, each photo image is compared with each ID card image, so there are totally 95,048 image pairs, including about 436 genuine pairs and 94,612 impostor pairs. For each image in the experiments, the face is cropped, aligned and warped to 128*128 pixels by affine transformation. By taking the advantages of recent develop of deep learning, we adopt the widely use Deep Convolutional Neural Network architecture, GoogLeNet [29], to extract

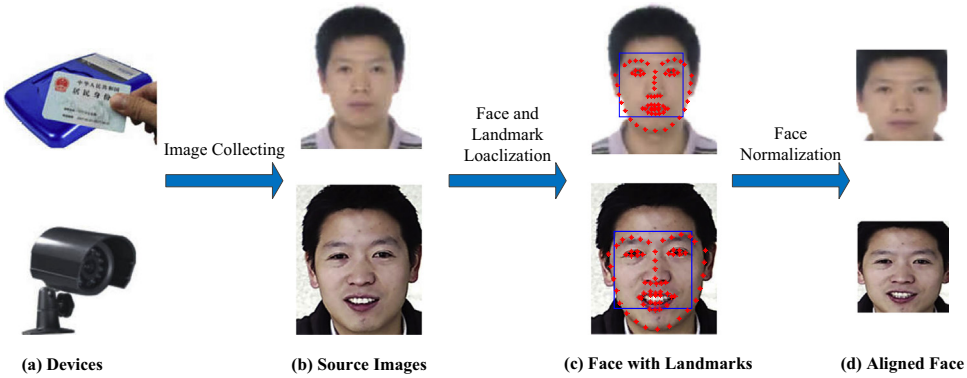


Fig. 2. The pipeline of image collecting and pre-processing.

face image feature. The GoogLeNet was trained on self-collected database containing 6,631 subjects of 54,4514 images.

In order to efficient evaluate the performance, the Receiving Operating Characteristic Equal Error Rate (ROC-EER) is adopted, which is the accuracy at the ROC operation point where the false positive and false negative rates are equal.

4.1.1. Experiments on data augmentation

Firstly, we test the effectiveness of data augmentation for latent face model training. Due to each subject has only one ID card image and the quality of images are various to different subject as shown in Fig. 3(a), so data augmentation is used to generate more ID card images to simulate different quality images in practice. As shown in Fig. 4, the first line and the second line show images generated by different degree of blur and contrast respectively.

In this experiment, face verification rates without data augmentation are used for comparison. Besides, we also test the influence of the number of person used for training and dimension of hidden space to our approach. The experimental results are shown in Fig. 5. We note that, with limited training subjects, data augmentation methods seems to have little improvement to the performance. While increasing the number of training subjects and dimension of hidden space, data augmentation methods significantly improve the verification accuracy. We argue that with small training dataset (e.g., the number of people is 100) low hidden dimension (e.g., 40) is sufficient to learn the intra-personal variations and mapping matrixes in this training set. However, the learned variations and mappings have limited generalization ability, which get poor verification rates (Fig. 5(a)). When increasing the number of training subject and enrich the training data with augmentation methods, face images include different kinds of noises and higher hidden dimension with more parameters is required to learn these variations. As shown in Fig. 5(e), augmentation methods help enrich the training dataset and we achieved best performance with hidden dimension of 600.

4.1.2. Experiments on number of training database size

Secondly, we evaluate our method on various training database size. The training database is divided into six subsets which have the number of subject with {100, 300, 500, 1000, 1500, 2000} respectively. On each subset, we test the performance of different picture number of each subject. Hidden dimension is default set as 100 in this experiment.

Results are shown in Fig. 6. As seen, by slightly increasing of people number, the performance is increased remarkably at beginning. We get best performance with each people has 30 photos. When each people's photo number larger than 30, there is little improvement to the performance. For training subsets, with larger subject number we get much better performance. We argue that for larger subject number each person has different kinds of images which help to better learn mapping matrixes. In addition, each people with 30 photos is sufficient for learning intra-personal variations.

4.1.3. Experiments on regularization parameters

We conducted more studies on λ_k, which is an additional item to guarantee the positive definite of Σ_k. In addition, λ_k also shows its sparse regularization to the parameters of Σ_k and accelerating convergence.

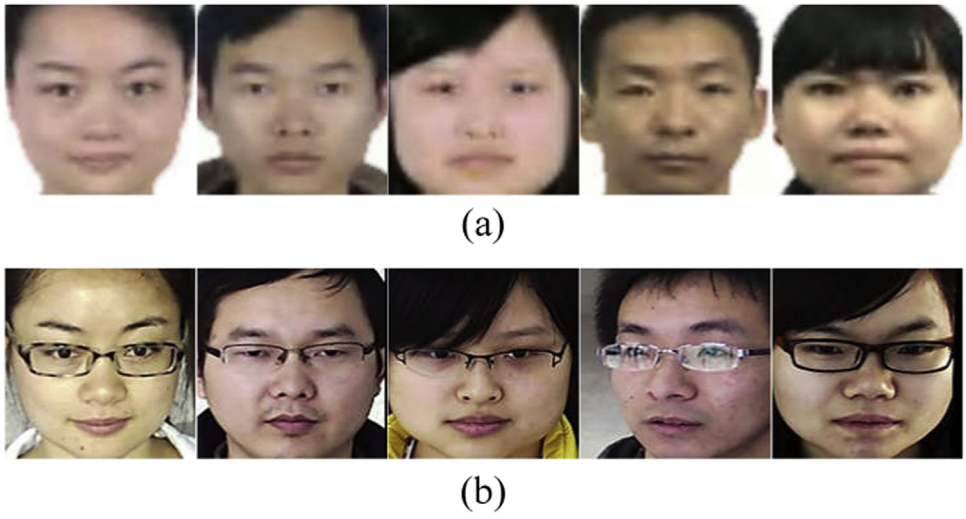


Fig. 3. Examples of Low-resolution and high-resolution images: (a) the low-resolution ID card images, (b) the corresponding high-resolution photo images.

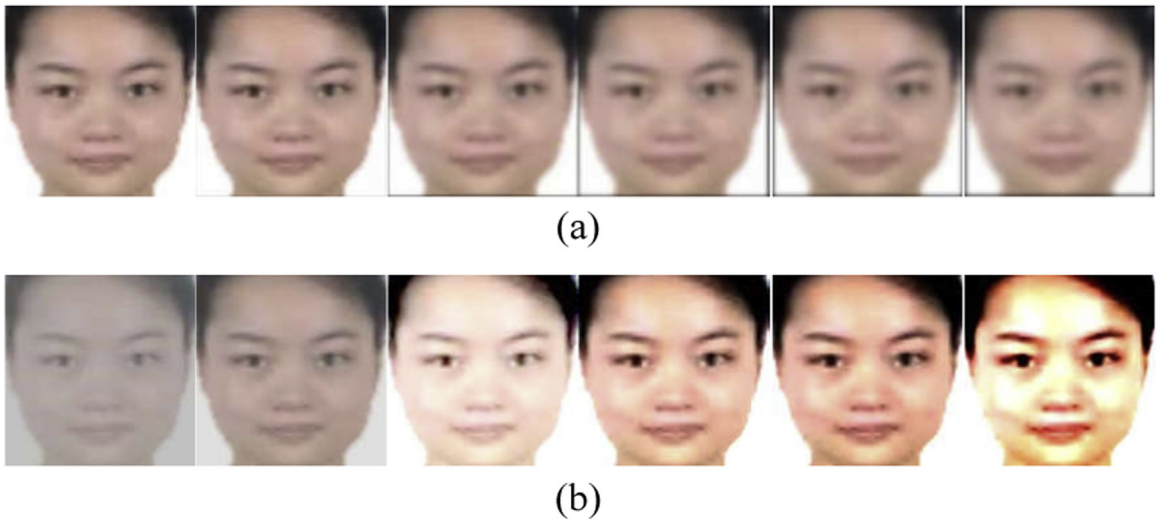


Fig. 4. Examples of data augmentation for ID card image: (a) images of different degree of blur, (b) images of different degree of contrast.

On the training set, we randomly select 30 photo images for each person and ID card images with data augmentation. Fig. 7 reports the face verification rates with different values of λ_k, where λ_1 is correlated with model of ID card images and λ_2 for photo images. From Fig. 7(a), we note that a better performance can be achieved with small regularization coefficients. When regularization coefficients were set to be 0, the verification rate get slightly degraded which is clearly observed in Fig. 7(b). We conclude that best performance can be achieved with small regularization coefficients and λ_1 is greater than λ_2. We argue that each person has a few ID card images without large variations and larger regularization coefficient is used to control the sparseness of Σ_1.

4.1.4. Compare with other metric learning methods

To further evaluate the performance of the proposed method, we compare our face verification method with other prevalent matric learning methods, including Joint Bayesian [25], LMNN (large margin nearest neighbor classification) [30], LDML (Logistic Discriminant Metric Learning) [31], KISSME (keep it simple and straightforward metric) [32] and Euclidean. Joint Bayesian method was implemented by ourselves according to [25]. The other

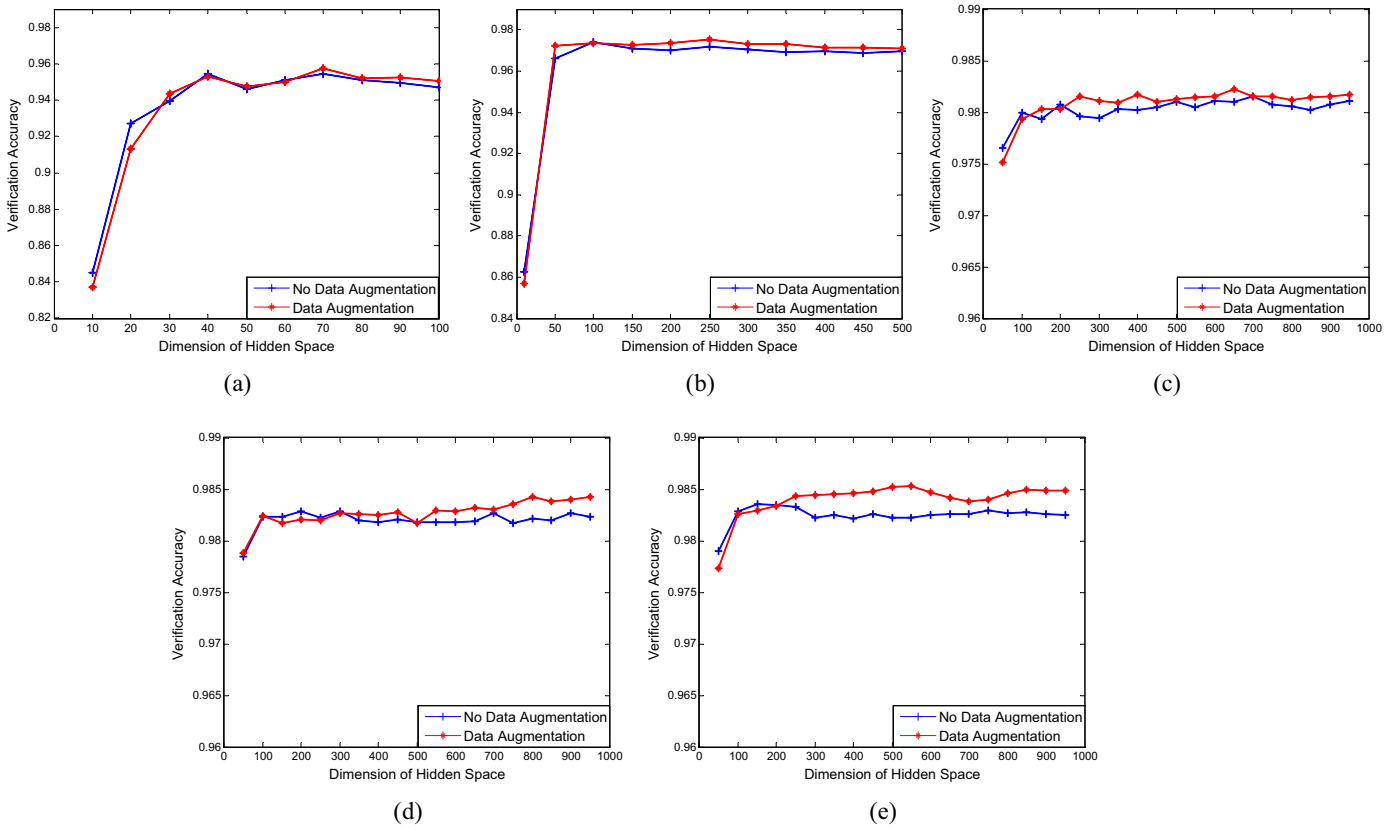


Fig. 5. Face verification with using data augmentation or not. (a) the number of training subject is 100; (b) the number of training subject is 500; (c) the number of training subject is 1000; (d) the number of training subject is 1500; (e) the number of training subject is 2000.

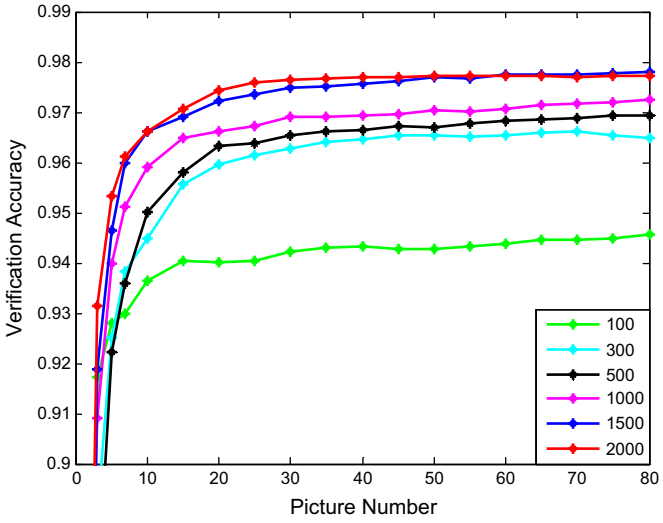


Fig. 6. Face verification rates with different training database size.

methods except Euclidean were download from each author's homepage and the default settings were adopted in our experiments.

In our experiment, the feature dimension was firstly reduced by PCA to 300. Then, different metric learning models were trained respectively. Experimental results are shown in Fig. 8 and Table 1. We note that the proposed method achieves higher verification accuracy than the other algorithms. When compared with Joint Bayesian, it reduces the error up to 24%. Different from other metric learning method, our approach learns mappings of different modalities to the same hidden space rather than directly learning the discriminative information without consider different modalities. The benefit from this property is that we achieve much better performance.

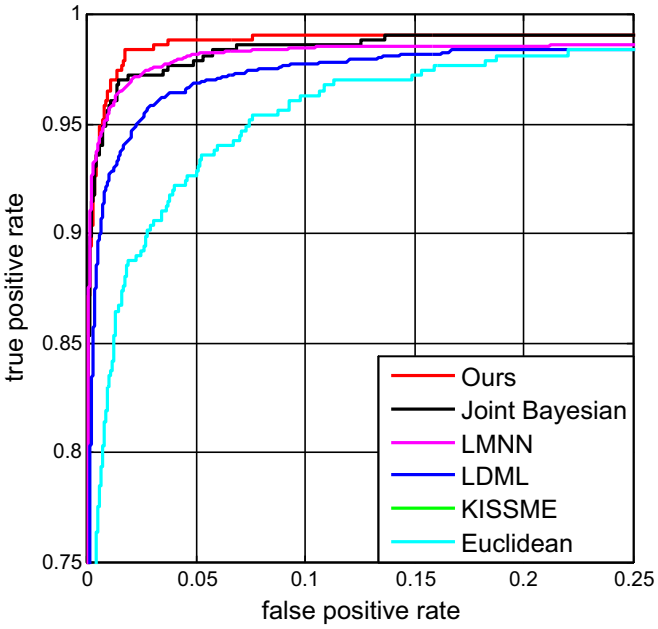


Fig. 8. ROC curves of various metric learning methods.

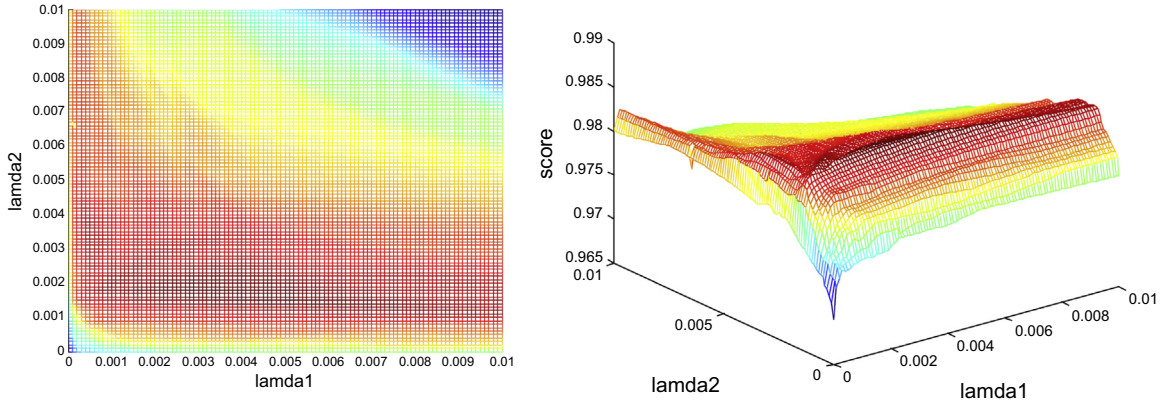
Table 1
Accuracy comparison with various metric learning methods.

Method	Accuracy (%)
Ours	98.34
Joint Bayesian [25]	97.81
LMNN [30]	97.71
LDML [31]	96.38
KISSME [32]	96.51
Euclidean	94.04

4.2. Sketch vs. photo

We also test face recognition performance of our method on two widely used sketch databases: the CUHK Face Sketch FERET Database (CUFSF) [16] and the CUHK Face Sketch Database (CUFS) [7]. CUFSF contains 1194 sketch-photo pairs with photos collected from the FERET database [19]. In this database, photos are taken under various lighting conditions and each sketch has shape exaggeration corresponding its photo. CUFS consists of 606 sketch-photo pairs with photos collected from the CUHK student database, the AR database [33] and the XM2VTS database [34]. All the photos in CUFS are taken under a normal illumination condition, and the sketches are drawn without exaggeration. Some examples are shown in Fig. 9.

In this experiment, Hierarchical Gaussianization encoding method [35] is used to extract each image's feature, and each database is randomly split in two sets which used for training and testing. On the CUFSF database, 500 persons are randomly selected as set for training, while the remaining 694 persons from the testing set. On the CUFS database, 306 persons are used for training and the remaining 300 are used for testing. The average



(a) (b)

Fig. 7. Face verification rates with different regularization coefficients.

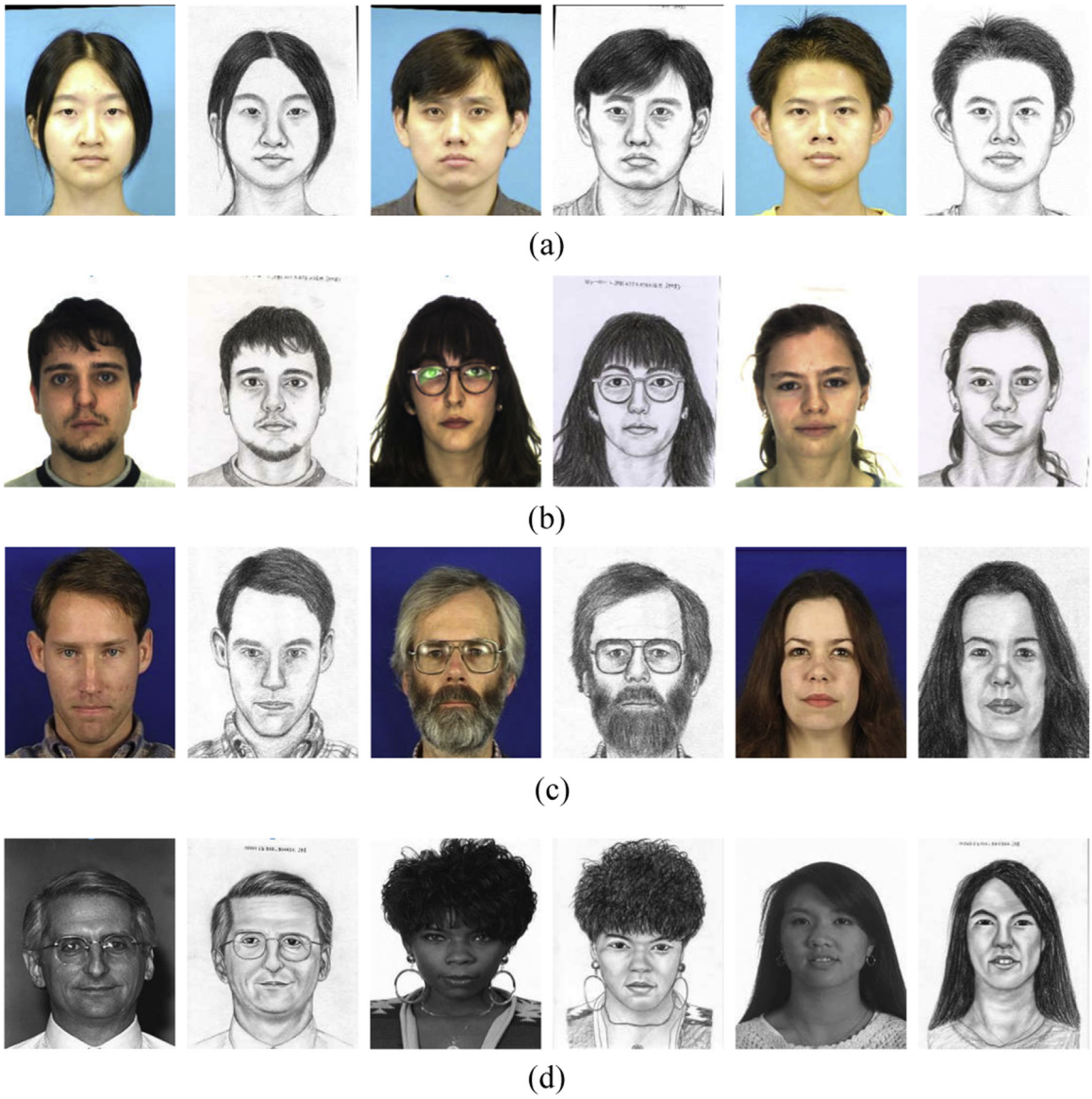


Fig. 9. Examples of face photos and sketches from (a) the CUHK student database, (b) the AR database, (c) the XM2VTS database, and (d) the FERET database.

Table 2
Recognition rates (%) of sketch-photo.

Databases	Methods					
	PLS [13]	TFSPS [6]	LDoGBP [18]	LRBP [17]	G-HFR [2]	Ours
CUFS	93.60	97.70	96.53	99.51	99.90	99.27
CUFSF	51	72.62	91.04	91.12	96.04	99.29

accuracies are reported over 10 cross validation by 10 random partitions of the training set and testing set.

The results of experiment are illustrated in Table 2. As shown, our method achieves high performance over both databases. On the CUFSF database, even with the large shape degradation of CUFSF, our method significantly outperforms the other methods on the CUFSF database. On the CUFS database, we achieve competitive result with others. Fig. 10 shows some recognition errors of our method. By analyzing these error pairs, we find that prediction and ground truth on CUFS database have similar contour and hairstyle, which confuse our model for correct recognition. As errors on CUFSF, some sketch images have serious exaggeration, especially line 2–3 images in Fig. 10(b), which look different from the ground truth and lead to difficult to recognition.

4.3. 3D face model vs. photo

We further conduct experiments with 3D-photo face recognition on Labeled Faces in the Wild database (LFW) [20]. We follow the protocol of [1] for closed set identification experiments, where the gallery set consists of 4249 frontal images and the probe set contains 3143 images. In this experiment, the GoogLeNet [29] is adopted to train the feature extraction models of 3D aligned face images and normal aligned face images respectively on CASIA-WebFace database [36].

We use the 3D face modeling method introduced in [37] to get 3D aligned face images. Firstly, face landmarks are detected for each face. Then, according to facial landmarks, the pose can be estimated by POSIT [38]. Thirdly, 3D shape model is constructed and the image texture is projected to the 3D geometry. Finally, according to the estimated pose, the 3D face model is used to get

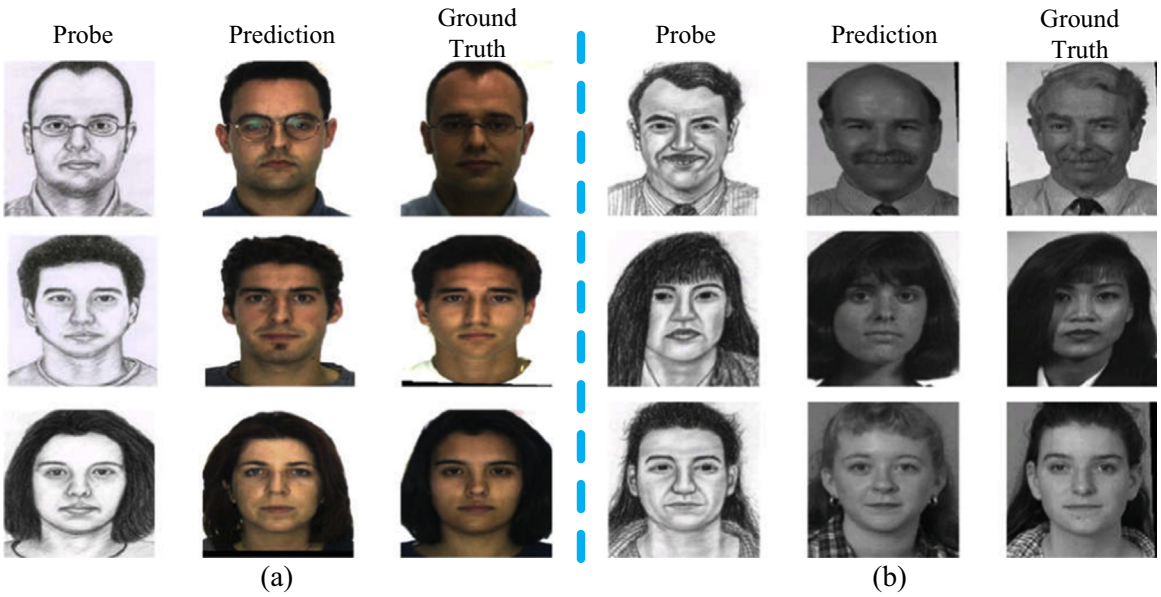


Fig. 10. Face recognition errors sketch vs. photo databases. (a) errors on CUFS database, (b) errors on CUFSF database.

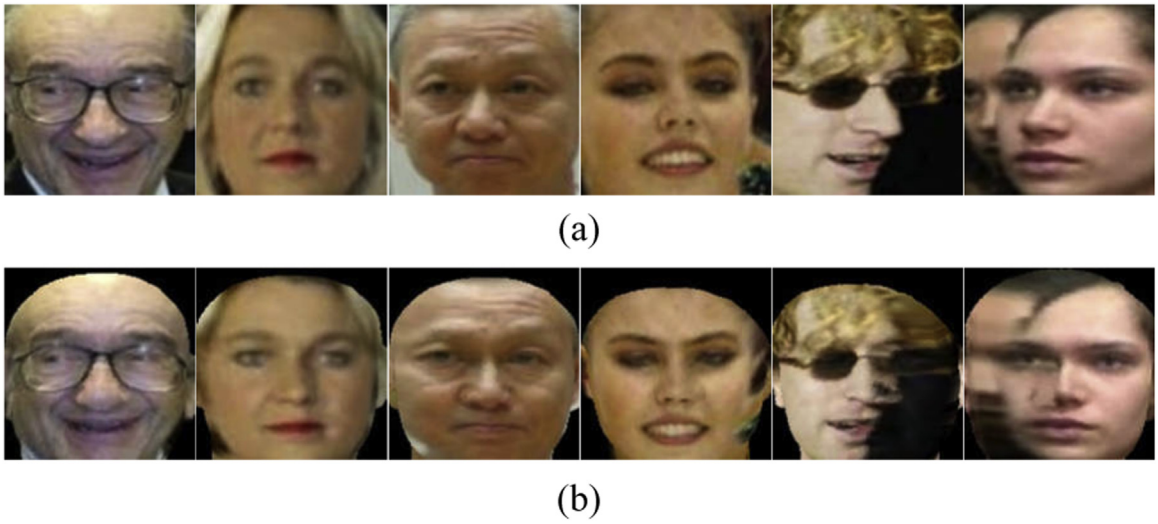


Fig. 11. Examples of 3D aligned face images and photo images from (a) images aligned by affine transformation, (b) images aligned by 3D face model.

Table 3
Closed set identification accuracy (%) for gallery face images using 3D aligned images and probe face images using normal aligned images.

Method	R-1	R-20	R-100	R-200
s_2 [1]	57.7	77.6	86.0	89.9
Ours	70.9	94.0	98.1	99.1

Table 4
Closed set identification accuracy (%) for gallery face images using normal aligned imagess and probe face images using 3D aligned images.

Method	R-1	R-20	R-100	R-200
s_3 [1]	63.9	83.4	90.7	93.6
Ours	67.5	94.0	97.8	98.6

the frontal image and projected to 2D space. As for photo image, affine transformation is used for normal face alignment. In Fig. 11, we show some examples of these two types face alignment.

Similar to [1], we experiment on two scenarios: (1) 3D aligned face images as gallery and normal aligned face images as probe. (2) Normal aligned face images as gallery and 3D aligned face images as probe.

We train our model on FaceScrub data set [39] which contains 530 different individuals with an average of approximately 200 images per person. The results are listed in Tables 3 and 4, where accuracies of different top ranked list retrieved form the gallery are reported. Compared with previous results [1] which used commercial off the shelf face recognition systems, we find that our method significantly improves the performance by large margin.

5. Conclusion

In this paper, we presented a latent face model for across-media face recognition. Our method first learns the mappings between a common hidden space and different media (e.g., low-resolution images, 3D face models, and sketches) and within-class variations in each medium. Then an improved Joint Bayesian method is used to calculate the similarities of faces from different

media. These two steps contribute to our method to have better generalization and discrimination ability than previous algorithms. Experiments on three different databases for across-media face recognition verify that our method achieves superior performance in term of face recognition accuracy and has strong generalization ability.

Acknowledgements

This work was supported by the National Natural Science Foundations of China (Grant nos. 61472386 and 61502444) and Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA 06040103).

References

[1] L. Best-Rowden, H. Han, C. Otto, B.F. Klare, A.K. Jain, Unconstrained face recognition: identifying a person of interest from a media collection, IEEE Trans. Inf. Forensics Security 9 (12) (2014) 2144–2157.
[2] C. Peng, X. Gao, N. Wang, J. Li, Graphical representation for heterogeneous face recognition, arXiv preprint arxiv:1503.00488.
[3] X. Gao, J. Zhong, J. Li, C. Tian, Face sketch synthesis algorithm based on e-hmm and selective ensemble, IEEE Trans. Circuits Syst. Video Technol. 18 (4) (2008) 487–496.
[4] Q. Liu, X. Tang, H. Jin, H. Lu, S. Ma, A nonlinear approach for face sketch synthesis and recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE, New York, NY, USA, 2005, pp. 1005–1010.
[5] X. Tang, X. Wang, Face sketch synthesis and recognition, in: IEEE International Conference on Computer Vision, IEEE, 2003, pp. 687–694.
[6] N. Wang, D. Tao, X. Gao, X. Li, J. Li, Transductive face sketch-photo synthesis, IEEE Trans. Neural Netw. Learn. Syst. 24 (9) (2013) 1364–1376.
[7] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1955–1967.
[8] L. Zhao, X. Gao, Y. Yuan, D. Tao, Sparse frontal face image synthesis from an arbitrary profile image, Neurocomputing 128 (2014) 466–475.
[9] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, in: European Conference on Computer Vision, Springer, Berlin Heidelberg, 2012, pp. 808–821.
[10] Z. Lei, S. Z. Li, Coupled spectral regression for matching heterogeneous faces, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, New York, NY, USA, 2009, pp. 1123–1128.
[11] D. Lin, X. Tang, Inter-modality face recognition, in: European Conference on Computer Vision, Springer, Berlin Heidelberg, 2006, pp. 13–26.
[12] A. Mignon, F. Jurie, Cmml: A new metric learning approach for cross modal matching, in: Asian Conference on Computer Vision, 2012, 14pp.
[13] A. Sharma, D. W. Jacobs, Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, New York, NY, USA, 2011, pp. 593–600.
[14] D. Yi, R. Liu, R. Chu, Z. Lei, S. Z. Li, Face matching between near infrared and visible light images, in: Advances in Biometrics, Springer, Berlin Heidelberg, 2007, pp. 523–530.
[15] B.F. Klare, Z. Li, A.K. Jain, Matching forensic sketches to mug shot photos, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 639–646.
[16] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 513–520.
[17] H.K. Galoogahi, T. Sim, Face sketch recognition by local radon binary pattern: Lrbp, in: IEEE International Conference on Image Processing, IEEE, New York, NY, USA, 2012, pp. 1837–1840.
[18] A.T. Alex, V.K. Asari, A. Mathew, Local difference of gaussian binary pattern: Robust features for face sketch recognition, in: IEEE International Conference on Systems, Man, and Cybernetics, IEEE, New York, NY, USA, 2013, pp. 1211–1216.
[19] P.J. Phillips, H. Moon, S. Rizvi, P.J. Rauss, et al., The feret evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[20] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report 07-49, University of Massachusetts, Amherst, 2007.
[21] J. Zhong, X. Gao, C. Tian, Face sketch synthesis using e-hmm and selective ensemble, in: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, IEEE, 2007, pp. I–485.
[22] S. Eleftheriadis, O. Rudovic, M. Pantic, Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition, IEEE Trans. Image Process. 24 (1) (2015) 189–204.
[23] M.R. Faraji, X. Qi, Face recognition under varying illuminations using logarithmic fractal dimension-based complete eight local directional patterns, Neurocomputing 199 (2016) 16–30.
[24] S.J. Prince, J. Warrell, J.H. Elder, F.M. Felisberti, Tied factor analysis for face recognition across large pose differences, IEEE Trans. Pattern Anal. Mach. Intell. 30 (6) (2008) 970–984.
[25] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: A joint formulation, in: European Conference on Computer Vision, Springer, Berlin Heidelberg, 2012, pp. 566–579.
[26] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. 5 (1977) 1–38.
[27] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. I–511.
[28] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[30] K. Weinberger, L. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[31] M. Guillaumin, J. Verbeek, C. Schmid, Multiple instance metric learning from automatically labeled bags of faces, in: European Conference on Computer Vision, 2010.
[32] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, New York, NY, USA, 2012, pp. 2288–2295.
[33] A.M. Martinez, The ar face database, CVC Technical Report 24.
[34] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, Xm2vtsdb: The extended m2vts database, in: International Conference on Audio and Video-Based Biometric Person Authentication, Vol. 964, Citeseer, 1999, pp. 965–966.
[35] X. Zhou, N. Cui, Z. Li, F. Liang, T. S. Huang, Hierarchical gaussianization for image classification, in: International Conference on Computer Vision, IEEE, New York, NY, USA, 2009, pp. 1971–1977.
[36] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, arXiv preprint arxiv:1411.7923.
[37] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, W. Gao, Efficient 3d reconstruction for face recognition, Pattern Recognit. 38 (6) (2005) 787–798.
[38] D.F. Dementhon, L.S. Davis, Model-based object pose in 25 lines of code, Int. J. Comput. Vis. 15 (1–2) (1995) 123–141.
[39] H.-W. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, in: IEEE International Conference on Image Processing, IEEE, New York, NY, USA, 2014, pp. 343–347.

Jiang-Jing Lv received the B.S. degree in information and computing science from University of Science and Technology of Hunan, Hunan, China, in 2012. He is currently pursuing a Ph.D. degree in pattern recognition at Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China. His research interests include face recognition and deep learning.

Jia-Shui Huang received the M.S. and Ph.D. degrees in Computer Science from the Zhejiang University, Zhejiang, China, in 2006 and 2010 respectively. He is currently an associate professor at Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. His research interests include computer vision and machine learning, with focus on face recognition and deep learning.

Xiang-Dong Zhou is an associate professor at the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. He received the B.S. degree in Applied Mathematics, the M.S. degree in Management Science and Engineering from National University of Defense Technology, Changsha, China, the Ph.D. degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1998, 2003 and 2009, respectively. He was a postdoctoral fellow at Tokyo University of Agriculture and Technology from March 2009 to March 2011. From May 2011 to October 2013, he was a research assistant and later an associate professor at the Institute of Software, Chinese Academy of Sciences. His research interests include machine learning and pattern recognition.

Xi Zhou received the B.S. and M.S. degrees in electronic science and technology from University of Science and Technology of China, Hefei, China, and the Ph.D. degree in electrical and computer engineering from University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2010. He is a Professor with Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China, and the Founding Lead of the Intelligent Multimedia Research Center. He has authored or co-authored more than 40 technical papers with Google Scholar Citation more than 600 times. His research interests include pattern recognition, machine learning, and computer vision and multimedia. Dr. Zhou received the Best Paper Award from the International Conference on Image Processing in 2007, the Best Student Paper Award from International Conference on Pattern Recognition in 2008, and the Best Paper Award from ACM Multimedia in 2013.

Yong Feng is a professor at the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. His research interests include symbolic-numeric computation and finite element method.