Pattern Recognition 46 (2013) 2497–2509

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Adaptive discriminant learning for face recognition

Meina Kan [a], Shiguang Shan [a,*], Yu Su [c], Dong Xu [b], Xilin Chen [a]

[a] *Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China*
[b] *School of Computer Engineering, Nanyang Technological University, Singapore*
[c] *Department of Computer Science, University of Caen, France*

ARTICLE INFO

Article history:
Received 31 December 2011
Received in revised form
26 October 2012
Accepted 30 January 2013
Available online 11 February 2013

Keywords:
Adaptive discriminant analysis
Face recognition
Single sample per person

ABSTRACT

Face recognition from Single Sample per Person (SSPP) is extremely challenging because only one sample is available for each person. While many discriminant analysis methods, such as Fisherfaces and its numerous variants, have achieved great success in face recognition, these methods cannot work in this scenario, because more than one sample per person are needed to calculate the within-class scatter matrix. To address this problem, we propose Adaptive Discriminant Analysis (ADA) in which the within-class scatter matrix of each enrolled subject is inferred using his/her single sample, by leveraging a generic set with multiple samples per person. Our method is motivated from the assumption that subjects who look alike to each other generally share similar within-class variations. In ADA, a limited number of neighbors for each single sample are first determined from the generic set by using kNN regression or Lasso regression. Then, the within-class scatter matrix of this single sample is inferred as the weighted average of the within-class scatter matrices of these neighbors based on the arithmetic mean or Riemannian mean. Finally, the optimal ADA projection directions can be computed analytically by using the inferred within-class scatter matrices and the actual between-class scatter matrix. The proposed method is evaluated on three databases including FERET database, FRGC database and a large real-world passport-like face database. The extensive results demonstrate the effectiveness of our ADA when compared with the existing solutions to the SSPP problem.

1. Introduction

Within the past decades, face recognition has received increasing attentions owing to its wide range of potential applications, e.g., identity authentication, homeland security, surveillance, human–computer interface and so on. Meanwhile, great challenges have been confronted in current recognition systems due to the large appearance variations in terms of illumination, expression, pose and so on. Numerous methods have been proposed to improve face recognition accuracy [1]. These methods can be roughly divided into two categories: geometry based methods and appearance-based methods [2]. The former describe a face using the relationship of facial components, e.g., the relative position of eyes and nose, while the latter represent the face using holistic appearance. In recent years, appearance-based methods have become the dominant approaches for face recognition.

Most appearance-based methods employ statistical learning technology, in which many samples are assumed for each person. So, the performance of these methods may be heavily affected by the number of samples from each person. More specifically, intra-personal and inter-personal variations may not be correctly estimated when only few samples from each person are available. Although a few methods are proposed to learn a robust and effective model [3–5], they are not applicable in the worst case that only single sample per person is available. In this case, the performance of many methods, e.g., the most popular Fisherfaces [6] will degrade seriously, or even fail to work. However, such Single Sample per Person (SSPP) problem [7] exists in many real-world applications, e.g., e-passport, watch list screening, because it is generally difficult to collect more than one sample per person in these scenarios. Hereafter, we call the dataset with only single sample per person as *single sample set*.

Recently, many methods have been developed [8] to address the SSPP problem, which can be roughly divided into three categories according to the information used for learning the recognition model as reviewed briefly in the following.

In the first category the single sample set was exploited as the only training set for the model learning. Most of them attempt to extract discriminative features for face recognition from SSPP. Some early methods [9,10] exploited the feature from facial-points of the single image. The typical methods employed the feature from the local region [11–17]. In some of them, each single image was partitioned into blocks which were further combined for the final classification by using different methods, e.g., HMM [12] and SOM

* Corresponding author. Tel./fax: +86 10 6260 0558.
E-mail address: sgshan@ict.ac.cn (S. Shan).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.037

[13]. Other methods followed the holistic-feature based scheme. For example, the Eigenface [18] estimated the total class scatter matrix using only the single sample set which actually degenerated to the between-class scatter matrix. Moreover, the various extension of PCA can be also used for the SSPP problem, e.g., 2DPCA [19], (PC)^{2}A [20,21], and Kernel PCA [22]. In [23], the optical flow between images was used to define an unequal feature weight based metric measure. In [24], an LBP feature representation was proposed. In [25], an approach using multiple representations for each image was employed. In [26], an image was decomposed into two parts to estimate the within-class and between-class scatter matrix respectively. In [27,28], the neighboring information was employed to obtain more discriminative low dimensional feature representation. While the above methods have addressed the SSPP problem, most of them are unsupervised methods which did not consider the intra-class variations.

In the second category, multiple virtual images were generated from each single sample such that the existing machine learning technologies can be applied. In order to extract the intra-class variations for each single sample, some researchers proposed to synthesize virtual samples or partition a single image into sub-images. After that, the single sample set was augmented to *multiple samples set* with multiple samples per person, thus many discriminant analysis methods can be applied. In [29,30], new virtual images were obtained by using the learned information. In [31–35], the virtual images were generated via affine transformation, photometric changes, noise perturbation, shifting and varying degrees of edge information. In [36], a component-based method was proposed by moving each face component along four directions to generate virtual face images. In [37], Chen et al. directly partitioned one face image into several sub-images with the same dimension and treated these sub-images as multiple samples of each person. Other researchers tried to generate the virtual images from new pose, illumination and expression by rendering the recovered 3D face model [38,39]. Overall speaking, all above methods need prior knowledge to guide the generation of virtual images that implies the estimation of 'virtual' intra-personal variations. However, how to guarantee the quality and reality of the virtual images is still an open problem for these methods.

In the third category, an auxiliary set containing multiple samples per person from other subjects was exploited to assist in learning the classification model, called *generic set* hereafter. Intuitively, the faces of all human beings look alike, which implies that different persons can have similar intra-personal variations. Therefore, the intra-personal variations of subjects in the single sample set can be approximated by using a generic set containing multiple samples per person [7,40–46]. In [41], the linear discriminant analysis model including both within-class and between-class scatter matrices was learned based on the images in the generic set and then applied directly to the single sample set for feature extraction. However, the variation distribution of the generic set is often quite different from that of the single sample set. Therefore, the discriminant model learned from the generic set is more suitable to distinguish the persons in the generic set but not those in the single sample set.

To address this problem, we previously proposed an Adaptive Generic Learning (AGL) method [7] to estimate the within-class scatter matrix of the single sample, based on the property that the variance of the sum of independent random variables equals to the sum of variance of each random variable. If images from different persons are independent which means the cross covariance matrix is zero, AGL can estimate an accurate within-class scatter matrix for the single sample through least square regression on the samples from all subjects in the generic set. But if the images from different persons are relevant, e.g., the images from the subjects that look like each other, the cross covariance matrix from different subjects cannot be ignored which inevitably leads to degeneration of the estimation from AGL. To handle this problem, we further proposed the Adaptive Discriminant Analysis (ADA) [46] based on the fact that the persons whose appearances are similar to each other also have similar intra-personal variations. So the within-class scatter matrix of the subject with single sample can be better estimated by using only the scatter matrices of the several look-alikes (called *neighbors*) in the generic set who are most similar to the single sample. In ADA, the estimation is obtained by linearly combining the within-class scatter matrices of the neighbors, called *arithmetic mean*, with the neighbors determined by kNN or Lasso regression. However, in case of the neighbors spreading a large region, a gap between the arithmetic mean and the ground truth will appear. To deal with this, we further propose a new nonlinear estimation of the within-class scatter matrix for the single sample called *Riemannian mean* in this work.

Overall, this work is a combination and extension of our previous methods AGL [7] and ADA [46]. The differences between this work and the conference papers are as follows: (1) this paper combines AGL and ADA into a unified framework (called Adaptive Discriminant Learning) and gives a detailed theoretical analysis (see Section 3); (2) this paper proposes a new nonlinear method to estimate the within-class scatter matrix for single sample under the proposed framework; (3) in the experiments of this paper, the accuracy of estimation for the individual and total within-class scatter matrix is additionally evaluated in terms of the similarity between the estimated matrix and the ground-truth; (4) In addition to the FERET and passport-like datasets used in [7,46], the large scale FRGC v2.0 dataset is also used in the experiments of this paper.

The remainder of this paper is organized as follows. Section 2 describes the Adaptive Discriminative Learning framework for dealing with SSPP problem. Sections 3 and 4 present the inference of the within-class scatter matrix for the single sample by exploiting the arithmetic mean and Riemannian mean respectively. Section 5 summarizes the whole algorithm of ADA. Section 6 evaluated the ADA on three face databases. Finally, conclusion is given in the last section.

2. Adaptive discriminant learning for SSPP face recognition

In the case of SSPP, many discriminant methods, e.g., the Fisher Linear Discriminant Analysis, fail to work. In this section, we begin with a brief introduction of the Fisher Linear Discriminant Analysis and show why it fails in the case of SSPP scenario. Then the proposed Adaptive Discriminant Learning (ADL) framework is described briefly. After that, a key step in the framework, i.e., inferring the within-class scatter matrix from a single sample, is introduced. Some deep discussion on the motivation and principle behind our method is given in the last subsection.

2.1. Fisher linear discriminant analysis

Fisher Linear Discriminant (FLD) Analysis is an efficient discriminant model for face recognition [1,6]. It aims to find a set of most discriminative linear projections by maximizing the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix:

$$W_{opt} = \arg\max_{W} \frac{\|W^T S_B W\|}{\|W^T S_W W\|}. \tag{1}$$

The within-class scatter matrix S_W and between-class scatter matrix S_B are respectively defined as

$$S_W = \sum_{i=1}^{c} \sum_{x \in X_i} (x - m_i)(x - m_i)^T, \tag{2}$$

$$S_B = \sum_{i=1}^{c} N_i(m_i - m)(m_i - m)^T, \quad (3)$$

where c is the number of classes in the training set, N_i is the number of samples from class i, m_i is the mean of all samples in class i, and m is the mean of all the samples in the training set.

From (2), we can see that, more than one sample is needed to calculate the S_W. So in the case of SSPP, S_W degenerates to 0, then any W can maximize (1) to be infinite which means FLD fails to work.

2.2. Adaptive discriminant learning framework

Since FLD cannot directly calculate the within-class scatter matrix with only one sample per person, the ADL framework is proposed to make the FLD-like methods applicable to the SSPP scenario by inferring the within-class scatter matrix for each single sample as shown in Fig. 1. First, the within-class scatter matrix is inferred from the single sample by leveraging an auxiliary generic set; then the between-class scatter matrix is directly calculated from the single sample set; and finally the FLD model can be achieved by applying the singular value decomposition on the inferred within-class scatter matrix and the actual between-class scatter matrix. Here the generic set is an auxiliary dataset with multiple samples per person, which does not necessarily contain any person in the single sample set.

2.3. Estimation of the within-class scatter matrix of single sample

Clearly, the critical step in ADL framework is inferring the within-class scatter matrix given a single sample. In AGL, the images from different persons are assumed to be independent, so the cross covariance matrix from different persons should be zeros, and thus the within-class scatter matrix of the single sample can be constructed from those of other persons by using least square regression.

However, in the real world, the images from different persons may be relevant which means the cross covariance matrix is not zero, e.g., images from similar persons, and in this case AGL cannot estimate an accurate within-class scatter matrix for the single sample. To handle this problem, we further propose ADA to infer the within-class scatter matrix of each single sample by using only a limited number of neighbors who are most similar to the single sample.

Briefly speaking, ADA is based on the assumption that persons who look alike each other are inclined to have similar intra-personal variations. So intuitively, we can approximate the intra-personal variations of any person by using a limited number of persons who are similar enough to this person as shown in Fig. 2. Specifically, given a single sample, a limited number of neighbors are first determined from the generic set by using kNN regression or Lasso regression. Then the within-class scatter matrix for this single sample is inferred by combining the within-class matrices of these neighbors.

In this work, two combination methods are exploited, i.e., arithmetic mean (detailed in Section 3) and Riemannian mean (detailed in Section 4). If the neighbors lie in a smaller region of the single sample which means their within-class scatter matrices can be considered to lie in an almost linear subspace, the within-class scatter matrix of the single sample can be well inferred by linearly combining those of the neighbors. This inference is actually the weighted arithmetic mean of the neighboring within-class scatter matrices. In contrast, if the neighbors spread in a larger region, the corresponding within-class scatter matrices should lie on a Riemannian manifold and cannot be considered to be in a subspace anymore. In this case, the weighted

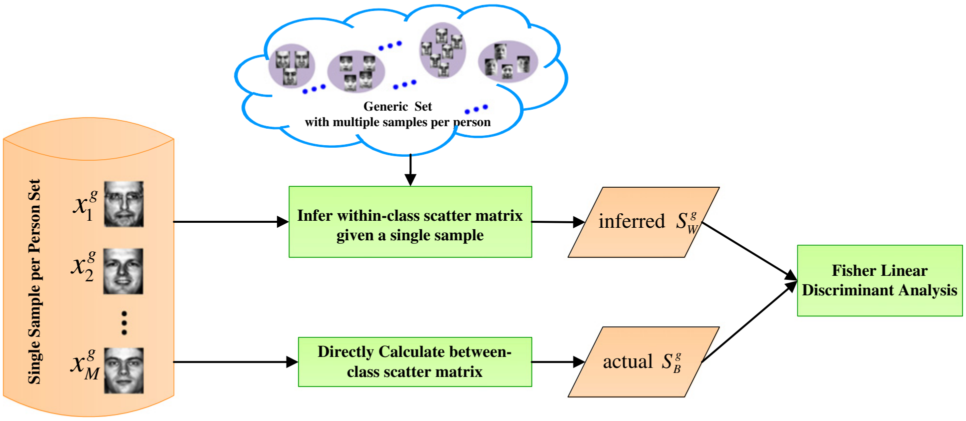
Generic Set
with multiple samples per person
Single Sample per Person Set
x_1^g
x_2^g
x_M^g
Infer within-class scatter matrix given a single sample
inferred S_W^g
Fisher Linear Discriminant Analysis
Directly Calculate between-class scatter matrix
actual S_B^g

Fig. 1. The framework of the proposed Adaptive Discriminant Learning.

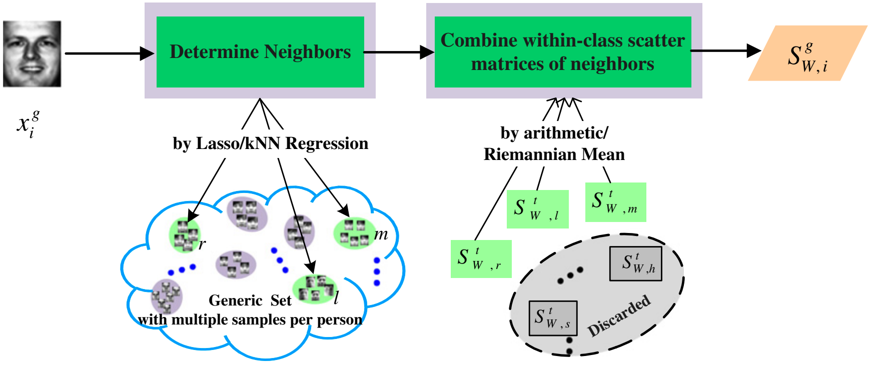

Fig. 2. The inference of the within-class scatter matrix given a single sample.

Riemannian mean becomes a natural way for inferring the within-class scatter matrix of the single sample.

2.4. *Motivation of the adaptation*

As is well known, the classic FLD assumes that all classes share the same homoscedastic distribution. However, in real applications, it may not be reasonable to assume that the subjects from different groups share the same within-class scatter matrix. Intuitively speaking, the difference between subjects from different groups (e.g., the Asian and the American) can be very large. Therefore, it is more reasonable to assume they are heteroscedastic. It also explains why cross-database FLD (i.e., the generic FLD method tested in this paper) performs very poor (see Table 2). On the contrary, the subjects from the same group can be assumed statistically homoscedastic, and in this case homoscedastic FLD becomes applicable. The above analysis explains our motivation to transfer the within-class model of the generic set to the specific single sample set.

In this work, we assume the generic set is a large dataset with images collected from different sources and even probably from different people groups, therefore they are essentially heteroscedastic. We also reasonably assume the subjects in the single sample set are from the same people group, so they are statistically homoscedastic. Therefore, we aim to estimate the total within-class scatter matrix of the single sample set by re-sampling the generic set. Specifically, we first estimate the within-class scatter matrix for each subject in the single sample set and then we average all the matrices to approximate the total within-class scatter for all the subjects in the single sample set.

In short, we are actually estimating the homoscedastic within-class distribution of the single sample set by re-sampling the heteroscedastic within-class distribution of the generic set.

3. Inferring the within-class scatter matrix by arithmetic mean

Face images can be considered lying on a manifold and the corresponding within-class scatter matrices lie on a manifold as well. If the single sample and its neighbors lie in a small region, the corresponding two local regions on these two manifolds can be both considered as linear subspace as shown in Fig. 3(a). As a result, the neighborhood relationship on the manifold of within-class scatter matrices is same as that on the manifold of face images. Therefore, the neighborhood relationship determined on the face image manifold can be directly used for estimating the within-class scatter matrix of a single sample on the within-class scatter matrix manifold. In this work, the neighbors and the combining weights are obtained by two methods, kNN and Lasso regression, as described below.

3.1. *What can similar persons having similar intra-class variations tell us?*

We observe that the persons who look alike have similar expression, aging, and pose variations with a high probability. That is, similar persons usually have similar intra-personal variations. This indicates that the within-class scatter matrices change slowly between similar persons, and should lie on a manifold, in fact on a Riemannian manifold [47]. Meanwhile, the face images also lie on a manifold. Intuitively these two manifolds should have similar structure, which means if person i is a neighbor of person j, then the within-class scatter matrix of person i is also a neighbor of the within-class scatter matrix of person j, because they both capture the relationship between person i and j just by using different statistical measures, i.e., the raw feature or the covariance of feature.

Since the within-class scatter matrices lie on a Riemannian manifold, the ones in a local region can be considered to lie in a linear subspace. Based on this observation, the within-class scatter matrix of a single sample can be estimated by using a linear combination of its near neighbors. The question is how to know the neighbors without knowing the exact within-class scatter matrix? Fortunately, as mentioned above, the manifold of within-class scatter matrices and the manifold of face images have similar structures, that is, their neighborhood relationships are almost the same as each other (as in Fig. 3(a)). Hence for each single sample, its neighbors and their corresponding weights determined from the face images manifold can be used on the manifold of within-class scatter matrices.

Formally, we first denote the random variable of the image set from any person as X, and $f(X) = S_W = XX^T$ as the corresponding

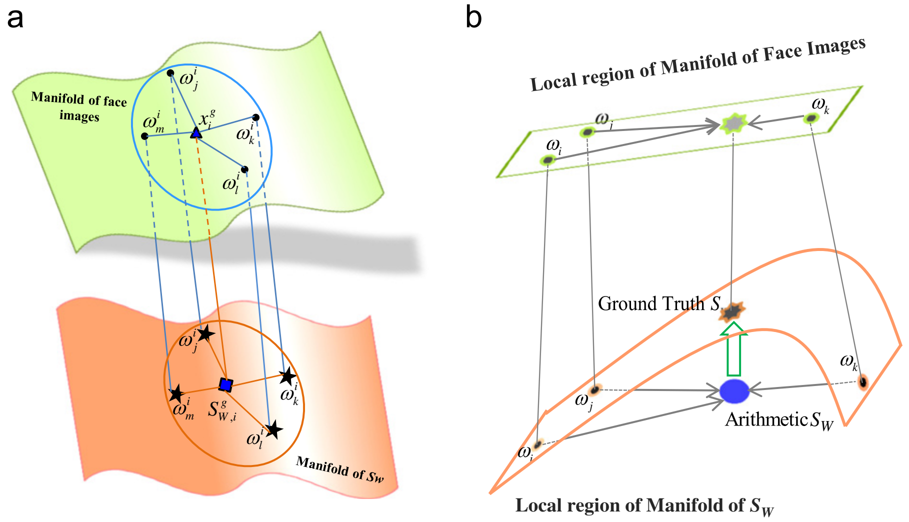

Fig. 3. Face image manifold and the corresponding within-class scatter matrix manifold: (a) The local region of face image manifold and the corresponding S_W manifold have similar structure in case of neighbors lying in a nearly linear region and (b) a gap between the estimation using the arithmetic mean and the ground truth appears in case of neighbors spreading in a non-linear region.

within-class scatter matrix. Here X is the variable after centralization, i.e., the subtraction of the mean. As illustrated in Fig. 3(a), the face image manifold is locally linear in a small region, and the corresponding S_W manifold is also locally linear in the corresponding small region. Then we can reveal that the mapping function $f(X)$ is almost linear in the small region. A detailed analysis of the neighborhood relationship between the S_W manifold and the face image manifold is given below.

Lemma 1. *The within-class scatter matrix $f(X)$ can be considered as a linear function of X in a local region.*

Proof. According to Taylor series, we have

$$f(X)=XX^T=X_0X_0^T+X_0(X-X_0)^T+(X-X_0)X_0^T+(X-X_0)(X-X_0)^T, \tag{4}$$

where X_0 is the mean of the neighbors within the small region around X. Therefore, the high order terms are small and can be ignored in (4), and then we have

$$f(X)\approx X_0X_0^T+X_0(X-X_0)^T+(X-X_0)X_0^T=XX_0^T+X_0X^T-X_0X_0^T\triangleq\hat{f}(X). \tag{5}$$

From (5), we observe that $f(X)$ can be considered linear in this local region where X is close enough to X_0. With this observation, we can further obtain that the neighborhood relationship on two manifolds related to X and S_W is preserved.

Lemma 2. *The neighborhood relationship on face image manifold is preserved on the manifold of the within-class scatter matrices.*

Proof. In a small region, when the neighbors are close enough to the single sample, X can be linearly reconstructed by k neighbors $X_{m1},X_{m2},\ldots,X_{mk}$ with the reconstructed coefficients denoted as $\omega_{m1},\omega_{m2},\ldots,\omega_{mk}$ and $\sum_{i=1}^{k}\omega_{mi}=1$, namely,

$$X=\omega_{m1}X_{m1}+\omega_{m2}X_{m2}+\cdots+\omega_{mk}X_{mk}. \tag{6}$$

Following Lemma 1, we can arrive at

$$\begin{aligned}f(X)&\approx XX_0^T+X_0X^T-X_0X_0^T\\&=(\omega_{m1}X_{m1}+\omega_{m2}X_{m2}+\cdots+\omega_{mk}X_{mk})X_0^T\\&\quad+X_0(\omega_{m1}X_{m1}+\omega_{m2}X_{m2}+\cdots+\omega_{mk}X_{mk})^T-\sum_{i=1}^{k}\omega_{mi}X_0X_0^T\\&=(\omega_{m1}X_{m1}X_0^T+\omega_{m1}X_0X_{m1}^T-\omega_{m1}X_0X_0^T)+\cdots\\&\quad+(\omega_{mk}X_{mk}X_0^T+\omega_{mk}X_0X_{mk}^T-\omega_{mk}X_0X_0^T)\\&=\omega_{m1}\hat{f}(X_{m1})+\omega_{m2}\hat{f}(X_{m2})+\cdots+\omega_{mk}\hat{f}(X_{mk}).\end{aligned} \tag{7}$$

Following (5), we can further have

$$f(X)\approx\omega_{m1}f(X_{m1})+\omega_{m2}f(X_{m2})+\cdots+\omega_{mk}f(X_{mk}). \qquad \square \tag{8}$$

From the above two lemmas, we can conclude that the within-class scatter matrix of a single sample can be estimated by using the linear combination of the within-class scatter matrices of its neighbors who are determined in a small region on the face image manifold.

3.2. How to find the neighbors and corresponding weights?

Having (8), if we can obtain the neighbors and corresponding weights, we can infer the within-class scatter matrix of the single sample. According to Lemma 2, we can determine the neighbors and the corresponding coefficients on the face image manifold. In our study, two methods, kNN regression and Lasso regression, are employed to determine the neighbors and corresponding weights.

Formally, we denote the single sample set as

$$G=[x_1^g,x_2^g,\ldots,x_M^g]\in R^{d\times M}, \tag{9}$$

where x_i^g is the sample from the i-th person in the single sample set, d is the feature dimension, and M is the number of persons (i.e., samples). We also denote the generic set as

$$A=[X_1^t,X_2^t,\ldots,X_c^t]\in R^{d\times N}, \tag{10}$$

where X_i^t is the samples from the i-th person, c is the number of the persons in the generic set and N is the total number of images from all persons. Note that the persons in this set can have no overlap with those in the single sample set, and may be captured under different conditions from the single sample set.

kNN regression: kNN is the most common method to find neighbors. Given x_i^g (the sample of the i-th person in the single sample set), we first simply find its k *nearest neighboring persons* in the generic set. Then, we assign a weight ω_j^i to each neighbor in the generic set as

$$\omega_j^i=\begin{cases}s(x_i^g,m_j^t), & m_j^t\in \text{ kNN of } x_i^g,\\ 0 & \text{otherwise},\end{cases} \tag{11}$$

where m_j^t is the sample mean of the j-th person in the generic set computed as

$$m_j^t=\frac{1}{N_j}\sum_{l=1}^{N_j}x_{j,l}^t, \tag{12}$$

$x_{j,l}^t$ is the l-th sample of j-th person in generic set, N_j is the number of samples from j-th person and $s(.)$ measures the similarity of two samples. kNN is simple yet without guarantee of complementation among the neighbors. So, Lasso regression is adopted alternatively.

Lasso regression represents the single sample by sparsely linear combination of other samples. To keep the sparsity of the regression coefficients, the samples in the generic set that are more similar to the single sample will have larger coefficients, and coefficients of most samples will be zero. So Lasso regression can determine the neighbors and coefficients simultaneously.

Formally, the weights for all the c persons in the generic set are simultaneously optimized by the following minimizing procedure [48]:

$$(\omega_1^i,\omega_2^i,\ldots,\omega_c^i)=\arg\min_{\omega_1,\omega_2,\ldots,\omega_c}\frac{1}{2}\left\|x_i^g-\sum_{j=1}^{c}\omega_j\cdot m_j^t\right\|_2^2+\lambda|\omega|_1, \tag{13}$$

where $\omega=(\omega_1,\omega_2,\ldots,\omega_c)$. Here, it is worth pointing out that, we do not give the exact definition of neighbors in Lasso regression. Nevertheless, in the cost function, the L_1 constraint term can lead to many zero weights. In other words, only a few of persons will be assigned non-zero weights. All these persons with non-zero weights can be regarded as 'neighbors'.

3.3. Inferring the within-class scatter matrix of the single sample

After obtaining the neighbors and their corresponding weights by (11) or (13), we can estimate the within-class scatter matrix $S_{W,i}^g$ of the i-th person with single sample x_i^g, by linearly combining of the within-class scatter matrices of the neighbors determined from the generic set according to (8) as follows:

$$S_{W,i}^g=\sum_{j=1}^{c}\omega_j^iS_{W,j}^t, \tag{14}$$

where $S_{W,j}^t$ is the within-class scatter matrix of the j-th person in the generic set. In fact (14) is the weighted arithmetic mean of the within-class scatter matrices of the neighbors.

4. Inferring the within-class scatter matrix by Riemannian mean

If the neighbors lie in a small region around the single sample, the within-class scatter matrix of this single sample can be well estimated by using the arithmetic mean in (14). Then what will happen if the neighbors spread in a large region?

Actually a gap between the estimated one and the ground truth will appear as shown in Fig. 3(b). To bridge this gap, we propose to employ the Riemannian mean to infer the within-class scatter matrix of a single sample and give a solution for the Riemannian mean to avoid the singular problem for matrix logarithm.

4.1. Gap between the arithmetic mean and Riemannian mean

The within-class scatter matrix is quadratic of the images, so the manifold of within-class scatter matrices may have a higher dimension than face image manifold. For this reason, the neighbors may lie in a large region, and cannot be considered as a linear subspace anymore. In this case, the within-class scatter matrix estimated by (14) should lie in the subspace spanned by the within-class scatter matrices of the neighbors denoted as 'Arithmetic S_W' shown in Fig. 3(b).

However, as illustrated in Fig. 3(b), the ground truth within-class scatter matrix of the single sample should lie on the same Riemannian manifold as the neighbors, like the 'Ground Truth S_W' in Fig. 3(b). So a gap between the estimated within-class scatter matrix by (14) and the ground truth appears. And the larger the region that the neighbors spread in, the bigger the gap.

This gap appears because we still exploit the arithmetic mean in a large region which cannot be considered as a linear subspace anymore. Thus a natural way to bridge this gap is to calculate the mean on the manifold, i.e., Riemannian mean.

4.2. Inferring the within-class scatter matrix with Riemannian mean

As in [49], the Riemannian mean can be obtained as follows:

$$S_{W,i}^{g} = \arg\min_{S} \sum_{j=1}^{c} d^2(S, S_{W,j}^{t}), \tag{15}$$

where the $S_{W,j}^t$ is the within-class scatter matrix of j-th person in the generic set. The estimated $S_{W,i}^g$ is expected to be as close to the ground truth $S_{W,i}^{g*}$ as possible. Inspired by the LLE [50], the estimated within-class scatter matrix should capture the intrinsic neighborhood geometry same as the ground truth. So we further exploit the weighted mean as follows:

$$S_{W,i}^{g} = \arg\min_{S} \sum_{j=1}^{c} \omega_j^i d^2(S, S_{W,j}^{t}). \tag{16}$$

Here ω_j^i is used to characterize the local geometry of the i-th single sample x_i^g, so only the neighbors should have non-zero values. According to [49], (16) can be efficiently solved as

$$S_{W,i}^{g} = \exp\left(\frac{1}{\sum_{j=1}^{c} \omega_j^i} \sum_{j=1}^{c} \omega_j^i \log(S_{W,j}^{t}) \right). \tag{17}$$

In the case of SSPP, there are two problems in (17). One is how to set the weights ω_j^i to preserve the local geometry and the other is how to calculate the logarithm for singular matrix $S_{W,j}^t$.

4.2.1. Setting the weights

In (16), the weights ω_j^i are used to capture the local geometry, i.e., the local neighborhood relationship, and the similarity between the ground truth within-class scatter matrix and $S_{W,j}^t$ is a good choice:

$$\omega_j^i = s(S_{W,i}^{g*}, S_{W,j}^{t}). \tag{18}$$

However, we do not know the ground truth, so it is impossible to obtain the exact similarity. Fortunately, we can approximate it by using the similarity obtained from X and therefore ω_j^i can be calculated same as in (11) or (13).

4.2.2. Solution for the Riemannian mean

In practice, most persons in the generic set may only have a limited number of samples, which is generally much smaller than the feature dimension. That is, $S_{W,j}^t$ is generally singular. Since the matrix logarithm operation is only defined for nonsingular matrix, (17) fails to work in this case. But we observe that the matrix logarithm operation can be well estimated by its first-order approximation as below.

For nonsingular symmetric matrix, the logarithm operation can be formulated as

$$\log(S) = \log(U\Sigma U^T) = U\log(\Sigma)U^T = U\log(\Lambda^2)U^T = 2U\log(\Lambda)U^T, \tag{19}$$

where U is the eigenvectors of S, Σ is the diagonal eigenvalue matrix and Λ is the element-wise square root of Σ. Note that $\log(\Lambda)$ is operated on each diagonal element of Λ independently.

For clarity, we define $g(x) = \log(x)$ and expand it at $x_0 = ne$, then we have

$$\begin{aligned} g(x) &= g(x_0) + g'(x_0)(x - x_0) + h(x) \approx g(x_0) + g'(x_0)(x - x_0) \\ &= \log(x_0) + \frac{1}{x_0}(x - x_0) = \log(n) + \frac{1}{ne}x, \end{aligned} \tag{20}$$

where $h(x)$ is the high-order term and $(1/e)x$ is the first-order term of $g(x)$. In this work, n is set to 1 for simplicity. Correspondingly, (19) can be estimated by using the first-order approximation as

$$\log(S) = \log(U\Sigma U^T) \approx \frac{2}{e} U\Lambda U^T. \tag{21}$$

Though the first-order approximation may lead to some information loss, it has a great advantage that it can still work for the singular covariance matrix. With the first-order approximation, the estimated within-class scatter matrix of a single sample using (17) can be calculated as

$$\begin{aligned} S_{W,i}^{g} &= \exp\left(\frac{1}{\sum_{j=1}^{c} \omega_j^i} \sum_{j=1}^{c} \omega_j^i \log(S_{W,j}^{t}) \right) \\ &\approx \exp\left(\frac{2}{e \cdot \sum_{j=1}^{c} \omega_j^i} \sum_{j=1}^{c} \omega_j^i U_{W,j}^{t} \Lambda_{W,j}^{t} (U_{W,j}^{t})^T \right), \end{aligned} \tag{22}$$

where $U_{W,j}^t$ and $\Lambda_{W,j}^t$ are the eigenvectors and diagonal matrix of square root of eigenvalues of $S_{W,j}^t$.

5. Algorithm of adaptive discriminant analysis

As mentioned above, in practice, the samples in the single sample set can be assumed be collected from similar sources, therefore their within-class scatter matrices are roughly homoscedastic. So, the total within-class scatter matrix of the given single sample set can be approximated by averaging that of each subject as

$$S_W^g = \sum_{i=1}^{M} S_{W,i}^{g}. \tag{23}$$

Meanwhile, the total between-class scatter matrix can also be directly calculated by using the samples in the single sample set:

$$S_B^g = \sum_{i=1}^{M} (x_i^g - m_g)(x_i^g - m_g)^T,$$

$$m_g = \frac{1}{M}\sum_{i=1}^{M} x_i^g. \qquad (24)$$

Now, we can learn the ADA model with the estimated within-class scatter matrix and the actual between-class scatter matrix as below:

$$W_{opt} = \arg\max_{W} \frac{\|W^T S_B^g W\|}{\|W^T S_W^g W\|}, \qquad (25)$$

which can be solved by using the generalized eigenvalue decomposition. The proposed Adaptive Discriminant Analysis algorithm is summarized in Table 1.

It seems that ADA is a little more complex than directly applying FLD on the generic set. However, in terms of computation complexity, it takes almost the same time as the FLD. Specifically, the AGL and ADA with arithmetic mean take roughly equal time as the FLD. The ADA with Riemannian mean takes a little more time, but it can estimate the within-class scatter matrix more accurately, thus leading to better recognition performance.

6. Experiments

In this section, we evaluate our ADA method for the SSPP problem on three large databases: FERET [51], a real-world passport-like face database and FRGC database [52]. We first evaluate the influence of the main parameters in ADA; then evaluate how closely the estimated within-class scatter matrices can approach the ground truth; finally compare with the existing methods for SSPP problem.

As mentioned, our ADA exploits two methods, kNN and Lasso regression, to determine the neighbors and also two methods, arithmetic mean and Riemannian mean, to combine the within-class scatter matrices for estimation. For short, we denote the ADA using kNN to find neighbors and arithmetic mean for estimation as *ADA-AM-kNN*, ADA using Lasso regression to find neighbors and arithmetic mean for estimation as *ADA-AM-Lasso*, ADA using kNN to find neighbors and Riemannian mean for estimation as *ADA-RM-kNN*, ADA using Lasso regression to find neighbors and Riemannian mean for estimation as *ADA-RM-Lasso* respectively.

6.1. Databases for evaluation

In our experiment, three databases are involved for evaluation. The first one is the FERET database [51] (some examples are shown in Fig. 4(a)), whose gallery contains 1196 images, one image per person. According to the FERET evaluation protocol, there are four probe sets: fafb, fafc, dupI, and dupII. The images in fafb and fafc sets are with expression variations and lighting variations respectively, and the images in dupI and dupII sets are collected at different dates.

The second test database contains the real world passport-like images collected by ourselves. The gallery consists of 3000 persons with a single image per person, and the probe set has 4190 images. Some example faces are shown in Fig. 4(b). Here, it is worth pointing out that, the images for the same person in this database were acquired with the interval of a few years and with various image acquiring devices. Therefore, this datasets forms a quite challenging SSPP scenario.

The third one is FRGC version 2.0 database [52]. FRGC database consists of about 50,000 recordings divided into training and validation partitions. It has six experiments, among which experiments 1 and 4 study the performance of face recognition in controlled and uncontrolled conditions respectively. Experiments 1 and 4 share the same target set consisting of 16,028 images collected in controlled condition. For experiment 1, the query set consists of 16,028 controlled images, while for experiment 4 the query set consists of 8014 uncontrolled images. All images in each target or query set correspond to 466 persons, and some examples are shown in Fig. 5(a). We also evaluate our method on this database. But, different from the typical protocol for face verification, here we perform face identification tasks since this paper attempts to address the SSPP problem. Because the full target set has multiple images for each person, one image per person is randomly selected to form the single sample set to simulate the SSPP problem. So the single sample set for FRGC database

Table 1
Algorithm of adaptive discriminant analysis.

Input:	A single sample set G and a generic set A.
Step 1:	Compute m_j^t and $S_{W,j}^t$ for each class in the generic set.
Step 2:	for each x_i^g in the single sample set: (a) Find its neighbors and corresponding weights by kNN (11) or Lasso (13) regression. (b) Estimate $S_{W,i}^g$ according to (14) or (22).
Step 3:	Compute the total within-class and between-class scatter matrices. (a) Compute the total within-class scatter matrix according to (23) (b) Compute the total between-class scatter matrix according to (24)
Step 4:	Learn ADA model by (25).

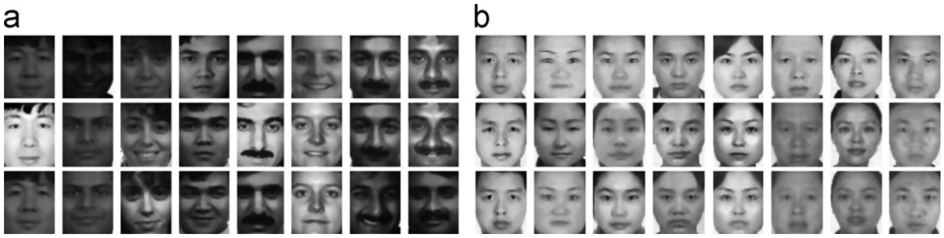

Fig. 4. Examples from (a) FERET and (b) the real world passport-like databases. Images in each column are from one person.

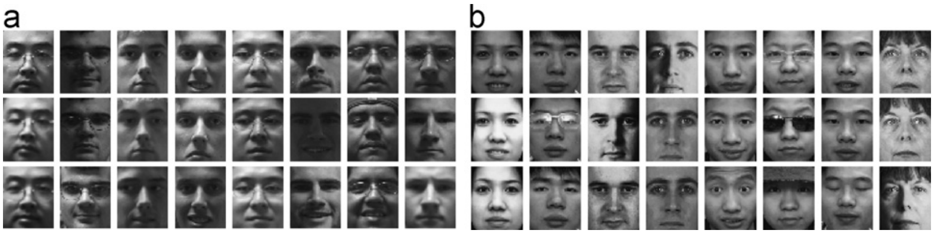

Fig. 5. Examples from (a) FRGC database and (b) the generic set. Images in each column are from one person.

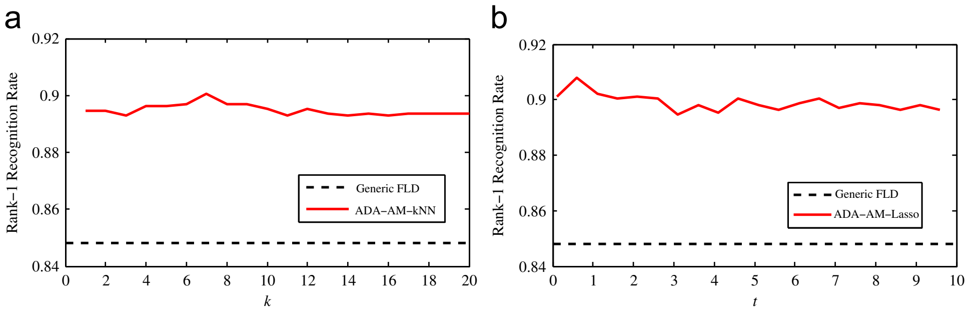

Fig. 6. Rank-1 face recognition rate with different parameters in ADA: (a) results with different k for ADA-AM-kNN and (b) results with different t for ADA-AM-Lasso.

contains 466 images, one image per person. The two query sets are used directly as the probe sets containing 8014 and 16,028 images respectively.

For these three test databases, we use the same generic set, which consists of the images from two databases: XM2VTS [53] and the training set of CAS-PEAL [54]. The XM2VTS database contains 3440 images of 295 persons taken over a period of four months with slight head pose and illumination variations. The CAS-PEAL training set contains 1200 images of 300 persons. The images cover large variations mainly due to expression and lighting. So, we obtain a generic set with 4640 images of 595 persons by merging the XM2VTS and CAS-PEAL databases. Some examples images in the generic set are shown in Fig. 5(b).

In our experiments, all face images are aligned according to manually labeled eye locations, normalized to 40∗50 pixels and preprocessed by histogram equalization.

6.2. Influence of parameters

In our method, the main parameter is k and λ in kNN regression in (11) and Lasso regression in (13). So we evaluate how performance changes with these parameters on fafb which is the largest probe set of FERET. We take the arithmetic mean based ADA as an example for evaluating the influence of k and λ as shown in Fig. 6, however, we have the similar observations for Riemannian mean based ADA as shown in Figs. 10 and 11. Actually (13) is solved in its equivalent form:

$$(\omega_1^i,\omega_2^i,\ldots,\omega_c^i) = \arg\min_{\omega_1,\omega_2,\ldots,\omega_c} \frac{1}{2}\left\| x_i^g - \sum_{j=1}^{c} \omega_j \cdot m_j^t \right\|_2^2,$$
$$\text{s.t.} \sum_{j=1}^{c} |\omega_j|_1 < t. \tag{26}$$

So, here we evaluate the performance of ADA over different t instead of λ.

Fig. 6 shows the performance of different k in ADA-AM-kNN and different t in ADA-AM-Lasso respectively. Note that, the Generic FLD stands for FLD method trained on the generic set and tested on the probe set.

From Fig. 6, it is clear that, no matter how to set the parameter, the proposed method always outperforms generic FLD. It can be also seen that the recognition rates of our method change slowly with the parameters k and t which forms an advantage of our method. Based on these results, k can be safely set to be smaller than 10, while t can be set to a value smaller than 1.0.

To further investigate the insensitivity of ADA to parameters, we also evaluate its performance when setting its parameters by cross-validation. The evaluation results on the FERET and passport-like databases are shown in Table 2, with the comparison to the results under the optimal parameters. The so-called optimal parameters are determined by searching in a range, i.e., [1 20] for k in kNN and [0.0001 2] for t in Lasso in our method. In contrast, for the cross-validation setting, the parameters are determined by validating on an independent validation set. Specifically, for the evaluation on FERET, the standard training set of FERET database is used as the validation set, and for the passport-like database, 3230 passport-like images of another 1500 subjects are collected as the validation set. From the comparisons, we can find that the performance

Table 2
Comparison of rank-1 face recognition rates of our ADA under two different parameter settings: the so-called optimal parameters and the parameters set by cross-validation (CV).

Methods	Parameter settings	FERET				Passport-like database
		fafb	fafc	dupI	dupII	
ADA-AM-kNN	Optimal	0.901	0.748	0.525	0.368	0.520
	CV	0.895	0.748	0.519	0.368	0.511
ADA-AM-Lasso	Optimal	0.912	0.758	0.519	0.372	0.508
	CV	0.905	0.758	0.514	0.372	0.507
ADA-RM kNN	Optimal	0.926	0.778	0.526	0.402	0.535
	CV	0.921	0.773	0.513	0.390	0.533
ADA-RM-Lasso	Optimal	0.916	0.768	0.532	0.393	0.514
	CV	0.913	0.753	0.532	0.385	0.505

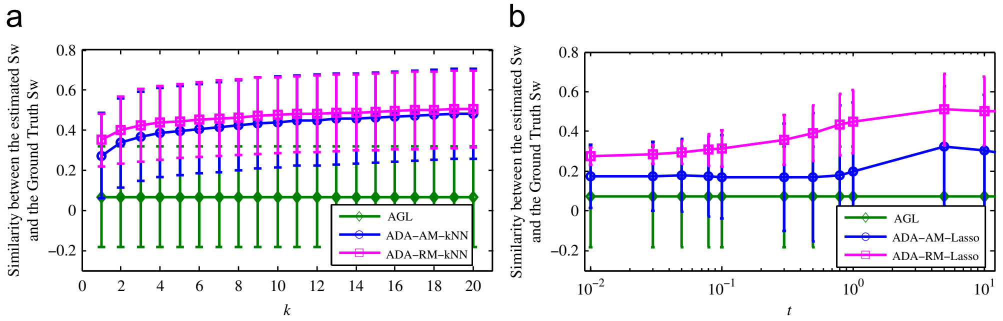

Fig. 7. Estimation accuracy in terms of cosine similarity between the estimated and the ground truth S_W. The horizontal axis represents the k in kNN-ADA or t in Lasso-ADA. The vertical axis represents the mean estimation accuracy with the standard variance bar: (a) estimation accuracy from ADA-kNN and (b) estimation accuracy from ADA-Lasso.

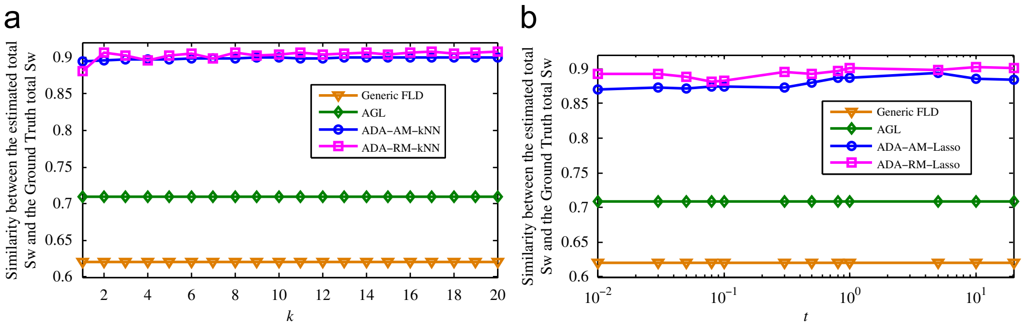

Fig. 8. The cosine similarity between the estimated total S_W and the ground truth total S_W. The horizontal axis represents the k in kNN based ADA or t in Lasso based ADA: (a) estimation accuracy from ADA-kNN and (b) estimation accuracy from ADA-Lasso.

under cross-validation setting only decreases slightly, compared with the best tuned ones, i.e., the optimal ones. These comparisons further illustrate the insensitivity of our method to its free parameters.

6.3. Accuracy of the estimation for within-class scatter matrix of single sample

To evaluate how well the estimated S_W from ADA and AGL can approach the ground truth S_W, we calculate their similarity as the measurement for evaluating the accuracy of the estimation.

Since the subjects in FRGC target set have multiple images per person, ground truth S_W of each subject is available. So, the accuracy of estimation is evaluated on this database. The full FRGC target set contains 16,028 images corresponding to 466 individuals with maximum 88, and minimum four images per person. One image of each person in the target set is randomly selected to form the single sample set.

The ground truth S_W of each subject is calculated by using all his/her images in the target set. Then S_W for each sample in the single sample set is estimated by using ADA with (14) and (22). Finally, the similarity of the estimated and ground truth S_W for each person is calculated through cosine function. The mean and standard variance of these similarities is shown in Fig. 7.

As seen, the estimation accuracy of AGL is 0.07. For kNN based ADA, the estimation accuracy of ADA with arithmetic mean is between 0.27 and 0.48, and that of ADA with Riemannian mean is between 0.35 and 0.51. In Lasso based ADA, the estimation accuracy of the ADA with arithmetic mean is between 0.17 and 0.30, and that of the ADA with Riemannian mean is between 0.27 and 0.50.

From these results, we can see that (1) ADA can indeed infer a more accurate within-class scatter matrix for the single sample than AGL. (2) ADA with Riemannian mean can further obtain a better estimation than ADA with arithmetic mean. (3) kNN based ADA and Lasso regression based ADA perform similarly well.

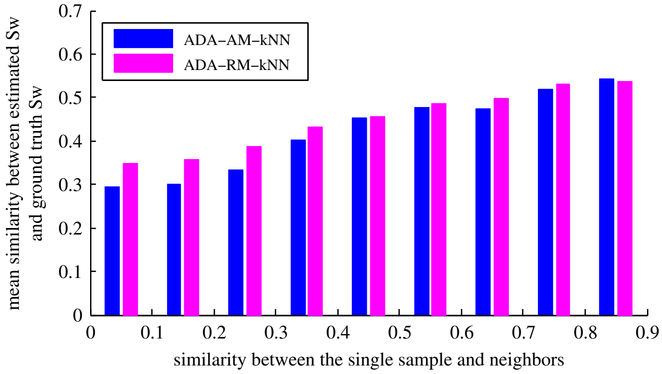

Fig. 9. The estimation accuracy with neighbors lying in different distance to the single sample. The horizontal axis represents the similarity between the single sample and its neighbors. The vertical axis represents the mean similarity of estimated and ground truth S_W.

Additionally, the similarity between the estimated total S_W and the ground truth total S_W are evaluated as shown in Fig. 8, from which the same conclusion can be drawn.

As presented previously, our method is based on the fact that the neighbors should share similar intra-class variations with the single sample. Intuitively, this should imply that, the estimation will be more accurate if the neighbors are closer to the single sample. To validate this guess, we explore how the estimation accuracy is affected by the distance of the neighbors to the single sample. The results are shown in Fig. 9, which plots the relationship between the estimation accuracy and the distance of the neighbors to the single sample. For example, the first bar represents the mean similarity of the ground truth and the estimated S_W with all neighbors similar with the single sample between 0 and 0.1.

From Fig. 9, we can see (1) the closer the neighbors are to the single sample, the more accurate the estimation is; (2) the closer the neighbors are to the single sample, the smaller the accuracy gap between the arithmetic mean and the Riemannian mean. In other words, if the neighbors are closer to the single sample, the within-class scatter matrices of the neighbors and the single sample are more likely to lie in a linear subspace and the gap between the ADA with arithmetic mean and with Riemannian mean becomes smaller. And even if the similarities between the neighbors and the single sample are large enough, for example larger than 0.8, the arithmetic mean works better than Riemannian mean. This might be attributed to information loss caused by the first-order approximation in (22).

6.4. Comparison with the existing methods

In this subsection, we compare our method with other typical methods dealing with SSPP problem. Among them, PCA, (PC)^{2}A, LBP, SVD-based FLD and Block FLD only use information from the single sample set; LPOE and FLD-VirtualImg employ the virtually generated images; Generic FLD only uses the generic set without adaptation. To make a deep comparison, the Generic FLD with the S_B directly calculated from the single sample set is also evaluated. Our AGL and ADA utilizes the information from the generic set to benefit the model for the single sample set. The basic information about these methods is briefly described in the following.

(1) *PCA* [18]: PCA is trained directly on the single sample set with 500 dimensions preserving about 95% energy. (2) *(PC)^{2}A* [20]: The weight of projection-combined face images is set to 0.3 according to [20]. *(PC)^{2}A* is also trained on the single sample set. (3) *LBP* [24]: Since the spatial partition of the face image is important to LBP, we tried different numbers of image blocks, and report the best result achieved with 80 blocks. (4) *SVD-FLD* [26]: Since the width and height of image are not equal in our experiment, so the within-class scatter matrix degenerates to $S_W=(1/c)\sum_{k=1}^{c}[(A_k-\overline{A}_k)^T(A_k-\overline{A}_k)]$. (5) *Block FLD* [37]: For Block FLD, the key parameter is the size of image blocks; we tested four different sizes, and report the best result with 10*25 size. (6) *LPOE* [11]: Generate 100 virtual images according to [11]. (7) *FLD-VirtualImg*: FLD [6] is trained on the virtual images from LPOE. (8) *Generic FLD* [41]: Fisherfaces are trained on the generic set. The peak result is presented by ransacking all dimensions. (9) *Generic FLD-Sb(SSS)*: The S_W is calculated from the generic set, while S_B from the single sample set as in (24). (10) *AGL* [7]: The PCA dimension is set the same as in [7]. (11) *ADA-AM-kNN*: ADA uses kNN to find neighbors and arithmetic mean for estimation. (12) *ADA-AM-Lasso*: ADA uses Lasso regression to find neighbors and arithmetic mean for estimation. (13) *ADA-RM-kNN*: ADA uses kNN to find neighbors and Riemannian mean for S_W estimation. (14) *ADA-RM-Lasso*: ADA uses Lasso regression to find neighbors and Riemannian mean for estimation. (15) *FERET-FLD* [6]: Fisherfaces are trained on the training set of FERET. In should be noted that, the training set of FERET contains images collected in the same condition as the test set of FERET, but no training images are in test set. Models obtained by this method are employed to just simulate the ground truth model and show how well the proposed ADA can approach the ground truth, so it should not be compared to other methods.

Table 3
Rank-1 face recognition rates on FERET and the real world passport-like databases.

Methods	**FERET**				**Passport-like database**
	fafb	**fafc**	**dupI**	**dupII**	
PCA [18]	0.896	0.134	0.399	0.150	0.168
PC^{2}A [20]	0.896	0.144	0.404	0.150	0.170
LBP [24]	0.976	0.557	0.575	0.329	0.335
Block FLD [37]	0.783	0.485	0.432	0.321	0.393
SVD-FLD [26]	0.833	0.253	0.314	0.120	0.202
LPOE [11]	0.891	0.134	0.421	0.158	0.165
FLD-VirtualImg	0.853	0.093	0.430	0.210	0.182
Generic FLD [41]	0.841	0.675	0.475	0.235	0.263
Generic FLD-Sb(SSS)	0.792	0.665	0.469	0.222	0.267
AGL [7]	**0.890**	**0.720**	**0.515**	**0.350**	**0.455**
ADA-AM-kNN	**0.901**	**0.748**	**0.525**	**0.368**	**0.520**
ADA-AM-Lasso	**0.912**	**0.758**	**0.519**	**0.372**	**0.508**
ADA-RM-kNN	**0.926**	**0.778**	**0.526**	**0.402**	**0.535**
ADA-RM-Lasso	**0.916**	**0.768**	**0.532**	**0.393**	**0.514**
FERET-FLD	0.980	0.711	0.616	0.316	–

Table 3 gives the comparison results of above methods on FERET and the real-world passport-like databases based on gray intensity feature. As seen, we can have that (1) methods only using the single sample set, virtual images, or only the generic set, e.g., (PC)^{2}A, LPOE and Generic FLD, perform not very well on most test sets; In contrast, the proposed AGL and ADA with an adaptation perform much better, even up to 6% and 14% on the more challenging real-world passport-like database. (2) ADA with Riemannian mean performs better than ADA with arithmetic mean up to 3.4%. (3) kNN and Lasso regression based ADA perform similarly well. (4) Compared to the FERET-FLD which is employed to stand for the 'ground truth model', ADA performs worse on fafb and dupI, but surprisingly performs better on fafc and dupII. This illustrates that our ADA can estimate a within-class scatter matrix for the single sample comparable to the 'ground truth'.

We also evaluate the proposed ADA on the FRGC v2 database. Since the above experiments have shown that AGL and ADA are much better than other methods that exploited only the single sample set or the virtual images, here we only evaluate the ADA-AM-kNN, ADA-AM-Lasso, ADA-RM-kNN and ADA-RM-Lasso, AGL, Generic FLD and Generic FLD-Sb(SSS). Besides, the results of ground truth FLD trained on the full FRGC target set is also given denoted as 'Ground Truth FLD'.

As mentioned in Section 6.1, we form two face identification tasks based on the datasets of FRGC experiments 1 and 4 respectively. Among them, experiment 1 explores the face identification for controlled test images. In Fig. 10, the kNN based ADA and Lasso based ADA are compared with existing methods on experiment 1.

We can see that AGL outperforms Generic FLD, Generic FLD-Sb(SSS) and ADA can further perform better than AGL. As expected, the ADA with Riemannian mean works better than ADA with arithmetic mean, which demonstrates that the ADA framework can achieve a better estimation of the recognition model. Another clear observation from Fig. 10 is that, the Ground Truth FLD outperforms all other methods including ADA significantly. However, this is reasonable and does not depreciate the proposed ADA method, since the testing images have similar distribution with the ground truth training data.

Compared with experiment 1, experiment 4 is much challenging since it deals with face images captured in uncontrolled environment. Fig. 11 gives the comparison of the kNN base ADA and Lasso based ADA with existing methods on experiment 4. From these two figures, we can reach similar conclusion as the previous experiment in spite of all methods reporting much lower recognition rates. Nevertheless, in this experiment, the proposed ADA methods achieve comparable accuracies to the

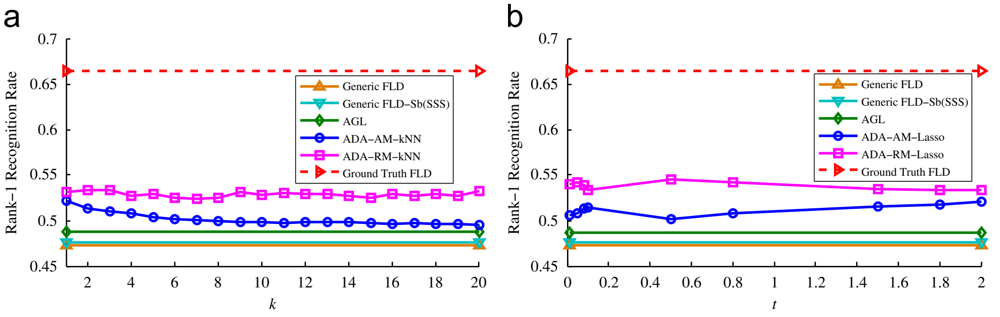

Fig. 10. Performance variations w.r.t. different (a) *k* in kNN based ADA and (b) *t* in lasso based ADA on FRGC experiment 1.

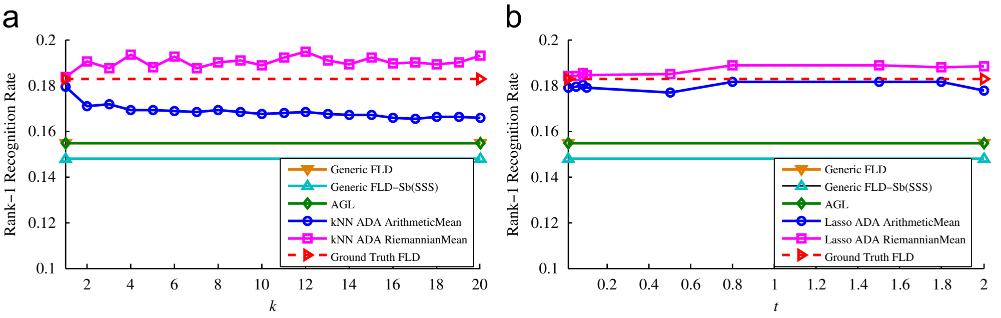

Fig. 11. Performance variations w.r.t. different (a) *k* in kNN based ADA and (b) *t* in Lasso based ADA on FRGC experiment 4.

Ground Truth FLD. The ADA with Riemannian mean even surprisingly outperforms the Ground Truth FLD. This is interesting but can be well interpreted by the following fact: the images in the target set of FRGC experiment 4 are all collected in controlled condition, so the Ground Truth FLD trained on these images does not model the uncontrolled variations in the testing images, and thus degenerates the performance abruptly. On the contrary, our generic set contains some images with uncontrolled variations, which is exploited by our ADA and thus improves the recognition accuracy.

7. Conclusion and future works

To deal with the SSPP problem, we propose Adaptive Discriminant Learning framework to estimate the within-class scatter matrix for each single sample and then exploit the Fisher Linear Discriminant Analysis to learn an adaptive discriminant model with the estimated within-class scatter matrix and the actual between-class scatter matrix of the single sample set. Thus, our ADL not only exploits the between-class discriminative information among samples in the single sample set, but also adaptively borrows the within-class variations from the generic set.

Under the ADL framework, the within-class scatter matrix of single sample is inferred by combining that of the subjects from the generic set. AGL linearly combines the within-class scatter matrices of all subjects from the generic set, while ADA only combines that of a limited number of neighbors with linear arithmetic mean or non-linear Riemannian mean. As evaluated on several large scale databases, ADL can estimated a more accurate model for the single sample set, and especially ADA with Riemannian mean achieved an impressive performance.

Although a great improvement has been obtained, however, there is still large room for the future progress as shown in Figs. 7 and 8. Besides the SSPP problem, our work could also be applicable for the small sample size scenario, which will be explored in future. Furthermore, as a general method adapting FLD-like method to SSPP problem, our method can be similarly further extended to other variants of FLD as long as they need to estimate the within-class scatter matrices from the single sample set.

Conflict of interest statement

None declared.

Acknowledgments

The work is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; Natural Science Foundation of China under contracts nos. 61025010, 61222211, and 61173065. The authors also thank the editor and the reviewers for their informative comments and suggestions.

References

[1] W.Y. Zhao, R. Chellappa, P.J. Phillips, A.P. Rosenfeld, Face recognition: a literature survey, ACM Computing Surveys 35 (4) (2003) 399–458.
[2] R. Brunelli, T. Poggio, Face recognition: features versus templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (10) (1993) 1042–1052.
[3] J.H. Friedman, Regularized discriminant analysis, Journal of the American Statistical Association 84 (405) (1989) 165–175.
[4] M. Loog, R. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise Fisher criteria, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (7) (2001) 762–766.
[5] O.C. Hamsici, A.M. Martinez, Bayes optimality in linear discriminant analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (4) (2008) 647–657.
[6] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[7] Y. Su, S. Shan, X. Chen, W. Gao, Adaptive generic learning for face recognition from a single sample per person, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3699–2706.
[8] X. Tan, S. Chen, Z.-H. Zhou, F. Zhang, Face recognition from a single image per person: a survey, Pattern Recognition 39 (9) (2006) 1725–1745.

[9] K.-M. Lam, H. Yan, An analytic-to-holistic approach for face recognition based on a single frontal view, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (7) (1998) 673–686.

[10] Y. Gao, Y. Qi, Robust visual similarity retrieval in single model face databases, Pattern Recognition 38 (7) (2004) 1009–1020.

[11] A.M. Martinez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (6) (2002) 748–763.

[12] H.-S. Le, H. Li, Recognizing frontal face images using hidden Markov models with one training image per person, in: International Conference on Pattern Recognition, vol. 1, 2004, pp. 318–321.

[13] X. Tan, S. Chen, Z.-H. Zhou, F. Zhang, Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble, IEEE Transactions on Neural Networks 16 (4) (2005) 875-886.

[14] H.R. Kanan, K. Faez, Y. Gao, Face recognition using adaptively weighted patch PZM array from a single exemplar image per person, Pattern Recognition 41 (12) (2008) 3799–3812.

[15] H.R. Kanan, Y. Gao, Recognition of expression variant faces from one sample image per enrolled subject, in: IEEE International Conference on Image Processing 2009, pp. 3309–3312.

[16] H.R. Kanan, M.S. Moin, Face recognition using entropy weighted patch pca array under variation of lighting conditions from a single sample image per person, in: Information, Communications and Signal Processing, 2009, pp. 1–5.

[17] H.R. Kanan, K. Faez, Recognizing faces using adaptively weighted sub-Gabor array from a single sample image per enrolled subject, Image and Vision Computing 28 (3) (2010) 438–448.

[18] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1991) 71–86.

[19] J. Yang, D. Zhang, A.F. Frangi, J.-y. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (1) (2004) 131–137.

[20] J. Wu, Z.-H. Zhou, Face recognition with one training image per person, Pattern Recognition Letters 23 (14) (2002) 1711–1719.

[21] S. Chen, D. Zhang, Z.-H. Zhou, Enhanced $(PC)^2A$ for face recognition with one training image per person, Pattern Recognition Letters 25 (10) (2004) 1173–1181.

[22] B. Scholkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neurocomputing 10 (5) (1998) 1299–1319.

[23] A.M. Martinez, Recognizing expression variant faces from a single sample image per class, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2003, pp. I–353– I–358.

[24] T. Ahonen, A. Hadid, M. Pietikainen, Face recognition with local binary patterns, in: European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 3021, 2004, pp. 469–481.

[25] F.D.l. Torre, R. Gross, S. Baker, B.V. Kumar, Representational oriented component analysis (ROCA) for face recognition with one sample image per training class, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 266–273.

[26] Q.-x. Gao, L. Zhang, D. Zhang, Face recognition using FLDA with single training image per person, Applied Mathematics and Computation 205 (2) (2008) 726–734. (Special Issue on Advanced Intelligent Computing Theory and Methodology).

[27] L. Qiao, S. Chen, X. Tan, Sparsity preserving discriminant analysis for single training image face recognition, Pattern Recognition Letters 31 (5) (2010) 422–429.

[28] W. Deng, J. Hu, J. Guo, W. Cai, D. Feng, Robust accurate and efficient face recognition from a single training image: a uniform pursuit approach, Pattern Recognition 43 (5) (2010) 1748–1762.

[29] D. Beymer, T. Poggio, Face recognition from one example view, in: International Conference on Computer Vision, 1995, pp. 500–507.

[30] A. Sharma, A. Dubey, P. Tripathi, V. Kumar, Pose invariant virtual classifiers from single training image using novel hybrid-eigenfaces, Neurocomputing 73 (10–12) (2010) 1868–1880.

[31] H.-C. Jung, B.-W. Hwang, S.-W. Lee, Authenticating corrupted face image based on noise model, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 272–277.

[32] D. Zhang, S. Chen, Z.-H. Zhou, A new face recognition method based on SVD perturbation for single example image per person, Applied Mathematics and Computation 163 (2) (2005) 895–907.

[33] J. Liu, S. Chen, Z.-H. Zhou, X. Tan, Single image subspace for face recognition, in: Analysis and Modeling of Faces and Gestures, Lecture Notes in Computer Science, vol. 4778, 2007, pp. 205–219.

[34] A. Majumdar, R.K. Ward, Single image per person face recognition with images synthesized by non-linear approximation, in: International Conference on Image Processing, 2008, pp. 2740–2743.

[35] S. Shan, B. Cao, W. Gao, D. Zhao, Extended Fisherface for face recognition from a single example image per person, in: IEEE International Symposium on Circuits and Systems, vol. 2, 2002, pp. II-81–II-84.

[36] J. Huang, P.C. Yuen, W.-S. Chen, J.H. Lai, Component-based LDA method for face recognition with one training sample, in: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 120–126.

[37] S. Chen, J. Liu, Z.-H. Zhou, Making FLDA applicable to face recognition with one sample per person, Pattern Recognition 37 (7) (2004) 1553–1555.

[38] T. Vetter, Synthesis of novel views from a single face image, International Journal of Computer Vision 28 (2) (1998) 103–116.

[39] P. Niyogi, F. Girosi, T. Poggio, Incorporating prior information in machine learning by creating virtual examples, Proceedings of the IEEE 86 (11) (1998) 2196–2209.

[40] J. Wang, K. Plataniotis, A. Venetsanopoulos, Selecting discriminant Eigenfaces for face recognition, Pattern Recognition Letters 26 (10) (2005) 1470–1482.

[41] J. Wang, K. Plataniotis, J. Lu, A. Venetsanopoulos, On solving the face recognition problem with one training sample per subject, Pattern Recognition (2006) 1746–1762.

[42] L. Zhang, D. Samaras, Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (3) (2006) 351–363.

[43] A. Majumdar, R.K. Ward, Pseudo-Fisherface method for single image per person face recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 989–992.

[44] L. Zhu, Y. Jiang, L. Li, Making discriminative common vectors applicable to face recognition with one training image per person, in: IEEE Conference on Cybernetics and Intelligent Systems, 2008, pp. 385–387.

[45] S. Chen, C. Sanderson, S. Sun, B.C. Lovell, Representative feature chain for single gallery image face recognition, in: International Conference on Pattern Recognition, 2008, pp. 1–4.

[46] M. Kan, S. Shan, Y. Su, X. Chen, W. Gao, Adaptive discriminant analysis for face recognition from single sample per person, in: International Conference on Automatic Face and Gesture Recognition 2011.

[47] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on Riemannian manifolds, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (10) (2008) 1713–1727.

[48] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2001.

[49] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Geometric means in novel vector space structure on symmetric positive-definite matrices, SIAM Journal on Matrix Analysis and Applications 29 (1) (2007) 328–347.

[50] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 232–2326.

[51] P.J. Phillips, H. Wechsler, J. Huang, P.J. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, Image and Vision Computing 16 (5) (1998) 295–306.

[52] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 947–954.

[53] K. Messer, J. Matas, J. Kittler, J. Lttin, G. Maitre, XM2VTSDB: The extended m2vts database, in: Second International Conference on Audio and Video-based Biometric Person Authentication, 1999, pp. 72–77.

[54] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, D. Zhao, The CAS-PEAL large-scale chinese face database and baseline evaluations, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 38 (1) (2008) 149–161.

Meina Kan received the B.S. from Shandong University. Now she is pursuing the Ph.D. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. Her research interests include pattern recognition, computer vision, especially face recognition.

Shiguang Shan received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences Beijing, in 2004. He has been with ICT, CAS since 2002 and has been a Professor since 2010. He is also the Vice Director of the Key Lab of Intelligent Information Processing of CAS. His research interests cover image analysis, pattern recognition, and computer vision. He is focusing especially on face recognition related research topics. He received the China's State Scientific and Technological Progress Awards in 2005 for his work on face recognition technologies.

Yu Su received the B.S., M.S. and Ph.D degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2003, 2005, 2009 respectively. His research interests include pattern recognition, computer vision, and especially focus on image classification and face recognition. From November of 2009, he worked as a postdoc researcher at University of Caen, with Prof. Frederic Jurie.

Dong Xu received the B.Eng. and Ph.D. degrees from the Electronic Engineering and Information Science Department, University of Science and Technology of China, in 2001 and 2005 respectively. He is currently an associate professor at Nanyang Technological University, Singapore. During his Ph.D. studies, he worked with Microsoft Research Asia and The Chinese University of Hong Kong. He also spent one year at Columbia University, New York, as a postdoctoral research scientist. His research interests include computer vision, pattern recognition, statistical learning and multimedia content analysis. He is a member of IEEE.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1988, 1991, and 1994 respectively. He was a Professor with the HIT from 1999 to 2005 and was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2004. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, since August 2004. His research interests include image processing, pattern recognition, computer vision, and multimodal interface. He has received several awards, including the China's State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work.