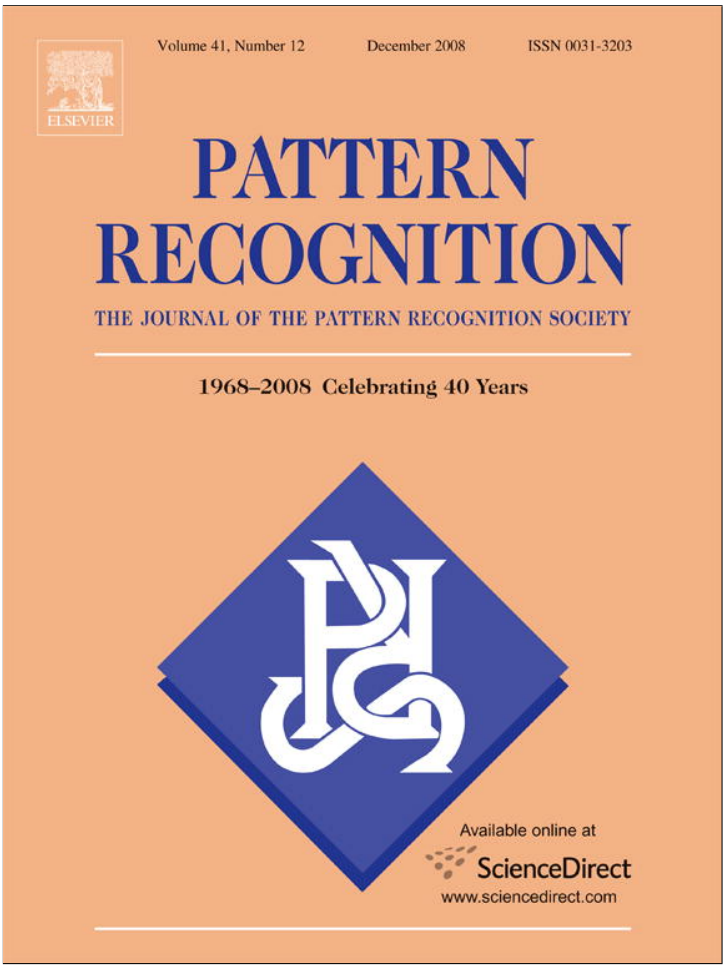

Pattern Recognition 41 (2008) 3813 – 3821

Contents lists available at ScienceDirect

Pattern Recognition

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

Locally linear discriminant embedding: An efficient method for face recognition

Bo Li[a,b], Chun-Hou Zheng[c], De-Shuang Huang[a,*]

[a]Intelligent Computing Lab, Institute of Intelligent Machine, Chinese Academy of Science, P.O. Box 1130, Hefei, Anhui 230031, China
[b]Department of Automation, University of Science and Technology of China, Hefei, Anhui 230027, China
[c]College of Information and Communication Technology, Qufu Normal University, Rizhao, Shandong 276826, China

ARTICLE INFO

Article history:
Received 22 September 2007
Received in revised form 12 April 2008
Accepted 27 May 2008

Keywords:
Feature extraction
Dimensionality reduction
Manifold learning
Locally linear embedding
Face recognition

ABSTRACT

In this paper an efficient feature extraction method named as locally linear discriminant embedding (LLDE) is proposed for face recognition. It is well known that a point can be linearly reconstructed by its neighbors and the reconstruction weights are under the sum-to-one constraint in the classical locally linear embedding (LLE). So the constrained weights obey an important symmetry: for any particular data point, they are invariant to rotations, rescalings and translations. The latter two are introduced to the proposed method to strengthen the classification ability of the original LLE. The data with different class labels are translated by the corresponding vectors and those belonging to the same class are translated by the same vector. In order to cluster the data with the same label closer, they are also rescaled to some extent. So after translation and rescaling, the discriminability of the data will be improved significantly. The proposed method is compared with some related feature extraction methods such as maximum margin criterion (MMC), as well as other supervised manifold learning-based approaches, for example ensemble unified LLE and linear discriminant analysis (En-ULLELDA), locally linear discriminant analysis (LLDA). Experimental results on Yale and CMU PIE face databases convince us that the proposed method provides a better representation of the class information and obtains much higher recognition accuracies.

1. Introduction

Over the past few decades, face recognition has received a lot of attention since the wide applications in many fields, such as video coding, surveillance and human–computer interface, and many face recognition techniques have been developed. Among them, appearance-based methods are well studied. Two issues are central to appearance-based face recognition: one is feature extraction for face representation; the other is classification of a new face image on the basis of the chosen features. When carrying out these methods, a face image of size $n \times m$ is represented as a vector in the image space $\mathbb{R}^{n \times m}$. However, the $n \times m$-dimensional space is too large to perform fast face recognition. A widely used way to attempt to resolve the problem is feature extraction. Being the key step in appearance-based face recognition, feature extraction aims to project the input data into a feature space that reflects the inherent structure of the original data and holds the useful information as much as possible. Thus the low dimensional representations of the faces can be obtained. Based on these representations, the new face images can be easily projected to the low dimensional space. At last a suitable classifier is adopted to predict the labels of these new face images. In this study, we shall focus on the topics of feature extraction for face recognition.

Currently, researchers have developed many feature extraction techniques for face recognition. These methods can be categorized into two classes based on either or not taking the class information into account: supervised or unsupervised. They are also broadly partitioned into linear methods and nonlinear ones. Linear feature extraction seeks a meaningful low dimensional subspace in a high dimensional input space by linear transformation. The subspace can provide a compact representation of the input data when the structure of data embedded in the input space is linear. Among all the linear feature extraction methods, the most well known are principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2].

PCA projects the original data into a low dimensional space, which is spanned by the eigenvectors associated with the largest eigenvalues of the covariance matrix of all the sample points, where PCA is the optimal representation of the input data in the sense of minimizing mean squared error (MSE) [1]. However, PCA is completely unsupervised because of not taking the class information of the input data into account, which may probably discard much useful information and weaken the recognition accuracy, especially when the number of sample points is very large [3].

* Corresponding author. Tel.: +86 0551 5592751.
E-mail address: dshuang@iim.ac.cn (D.-S. Huang).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.05.027

However, by considering the labels of the input data, LDA can produce an optimally discriminant projection by a linear transformation. The transformation matrix consists of the eigenvectors whose corresponding eigenvalues can maximize the ratio of the trace of the between-class scatter to the trace of the within-class scatter. Unlike PCA, LDA takes consideration of the labels of the input data. It is generally believed that the class information can improve the recognition ability. But there still exist some drawbacks in LDA. When the sample number is smaller than the class number of the input data, the optimal subspace cannot be found by carrying out LDA because the corresponding within-class scatter matrix is not inversed, which is named as the small sample size (SSS) problem. So far many effective and efficient methods [4–14] have been explored to solve the problem.

Both PCA and LDA have been successfully applied to some linear data. However, they fail to explore the essential structure of the data with nonlinear distribution. In order to overcome the problem, many nonlinear feature extraction methods have to be developed. Kernel-based approaches and manifold learning-based ones are promising for nonlinear feature extraction. Kernel-based technique is implicitly mapping the observed patterns into potentially much high dimensional feature space by a kernel trick, such as Gaussian kernel. It is made possible that the nonlinear structure data will be linearly separable in the kernel space. The widely used kernel techniques are kernel principal component analysis (KPCA) [15] and kernel Fisher discriminant analysis (KFDA) [16], which can be viewed as the kernel versions of PCA and LDA. KPCA and KFDA have been proved to be effective in some real world applications. However, the kernel-based methods can improve the linear discriminability at the cost of increasing dimensions and therefore high computational cost. Furthermore, due to introducing the kernel trick, how to select different kernels and how to assign the optimal parameters in them remain unclear. In most of the cases, experience still plays an important role.

Unlike kernel-based methods, manifold learning-based methods are straightforward in finding the inherent nonlinear structure hidden in the observe space. In the past few years many manifold learning-based algorithms have been presented. Among them, isometric feature mapping (ISOMAP) [17], locally linear embedding (LLE) [18,19] and Laplacian eigenmap (LE) [20,21] are widely used. They have yielded impressive results on artificial and real world data sets.

LLE is a representative local linear manifold learning method. Based on the assumption of the local linearity, LLE first constitutes local coordinates with the least constructed cost and then maps them to a global one. Experiments have proven that LLE is an effective method for visualization. However, some limitations are exposed when LLE is applied to pattern recognition.

One limitation is the out-of-sample problem. Because the weighted matrix of LLE is constructed on the training data, when a new data point is coming, how to generalize the results of training samples to the coming data attracts a lot of attention. Saul et al. [19] provided a non-parametric model and a parametric model to solve the problem. In the non-parametric model, what should be done first is to find the k nearest neighbors of the new data point, and then compute the linear weights that can best reconstruct the new data point with their k nearest neighbors under sum-to-one constraint, and in the end obtain the output linearly reconstructed by the corresponding k nearest neighbors and their corresponding weights in a low dimensional space. The work of non-parametric model amounts to a local linear representation. In the parametric model, the probability distribution and a hidden variable parameter are introduced. The output can be calculated based on some prior information of the probability distribution. In addition, Bengio et al. [22] and DeCoste [43] proposed a kernel method to embed the new data points because of the generalization ability of Mercer kernel. Recently, Kokiopoulou et al. [23] plugged a linear transformation to the original LLE. The embedding results of the coming points can be successfully attained with the linear transformation. Among the approaches mentioned above, the method using linear transformation can surmount the out-of-sample problem with the cheapest computational cost.

Another limitation lies in that the classical LLE neglects the class information, which will impair the recognition accuracy. Recently, many modified LLE algorithms have been put forward to make use of the label information. Some supervised versions of LLE [24–27] were introduced to deal with data sets labeled with class information. The intuition of these algorithms is to obtain disjoint embedding for the individual classes. The local neighbors of a sample point are selected by the following steps: firstly, the k nearest neighbors are found; secondly, the local neighborhoods of the sample point should be composed of the points among the k nearest neighbors with the same label. This can be achieved by artificially increasing the distances between samples belonging to different classes, but keeping them unchanged if samples are from the same class [25]. Those supervised LLE methods have achieved good classification results on some data sets. At the same time, they divide all the sample points into disconnected parts instead of an entire graph in original LLE. So these supervised LLE algorithms also bring a problem about how to apply the LLE to disconnected components. Saul et al. [19] suggested calculating each disconnected component using the original LLE, respectively. Similar to Saul et al., Polito et al. [28] also advanced a group method to overcome the problem. However, both techniques make the algorithms more complicated.

Recently, some other supervised LLE algorithms combined with LDA are becoming popular. Zhang et al. presented a unified framework of LLE and LDA [29,30]. This framework essentially equals to LLE + LDA. There are still some weaknesses in this proposed algorithm. Firstly, the original LLE is carried out and then the embedding results can be obtained. Secondly, LDA is adopted to find the discriminant features from these embedding results. In the whole process the dimensionality has been reduced two times. First of all, the dimensions must be reduced to be smaller than the number of the classes to avoid the SSS problem, thus some useful embedding information may be probably thrown away. Pang et al. [31] also brought forward a model that is linearly constructed by the objective function of LLE and LDA with some constraints. The model can be either pure LLE or pure LDA when the coefficient is one or zero, respectively. They have proven that the model outperforms some linear and kernel-based methods in face recognition [31]. But it remains unclear how to select an optimal coefficient to obtain high recognition rate on other data sets. A local Fisher embedding was put forward by de Ridder et al. [32]. When applying LLE to the sample points, the class information is contained in the matrix M, for which the embedding results can be obtained by performing eigen-decomposition. However, how to find the optimal weighted coefficients still needs further demonstration.

In this paper, following the intuition that the naturally occurring face may be sampled from the data with a probability distribution on a sub-manifold of ambient space, a supervised version of LLE, namely locally linear discriminant embedding (LLDE), is proposed for face recognition. In the proposed algorithm, we construct a vector translation and distance rescaling model to enhance the recognition ability of the original LLE from two aspects. One is the property that the embedding cost function is invariant to translations and rescalings [18,19], and the other is that the transformation to maximize the modified maximizing margin criterion (MMMC) is introduced. Based on the first property, the embeddings can be translated to any places without changing the reconstruction error. And then the optimal translated vectors for classification can be determined by maximizing MMMC, which is linearly composed of the between-class scatter and the within-class scatter of the input points. Thus the SSS problem can be successfully avoided because the inverse of the

within-class scatter will be not taken into account in MMMC. At the same time, the class information is not used to find the neighbors with the same label but to explore the optimal translated vectors, so it does not need to obtain disjoint embedding for the individual classes. Furthermore, it is the translations and rescalings that make the data with different class labels separated farther and data belonging to the same class clustered closer. So the proposed algorithm will improve the discriminability of the data significantly.

This paper is organized as follows. Section 2 describes classical LLE algorithm. Section 3 presents the proposed LLDE algorithm and the corresponding theoretical analysis. Experimental results and simulations on Yale and CMU PIE data sets are given in Section 4. Finally, discussions and conclusions are presented in Section 5.

2. Locally linear embedding

Let $X=[X_1,X_2,\ldots,X_n]\in R^{D\times n}$ denote n points in a high D dimensional space. The data points are well sampled from a nonlinear manifold, of which the intrinsic dimensionality is $d(d\ll D)$. The goal of LLE is to map the high dimensional data into a low dimensional manifold space. Let us denote the corresponding set of n points in the embedding space as $Y=[Y_1,Y_2,\ldots,Y_n]\in R^{d\times n}$. The outline of LLE can be summarized as follows:

Step 1: For each data point X_i, identify its k nearest neighbors by *kNN* criterion or ε-ball criterion.

Step 2: Compute the optimal reconstruction weights that can minimize the error of linearly reconstructing X_i by its k nearest neighbors.

Step 3: Compute the low dimensional embedding Y for X that best preserves the local geometry represented by the reconstruction weights.

Step 1 is typically done by using Euclidean distance to define neighborhood, although more sophisticated criteria may also be used, such as Euclidean distance in kernel space or cosine distance.

After identifying the k nearest neighbors of points X_i, Step 2 seeks the best reconstruction weights. Optimality is achieved by minimizing the local reconstruction error of X_i:

$$\varepsilon_i(W)=\arg\min\left\|X_i-\sum_{j=1}^{k}W_{ij}X_j\right\|^2 \tag{1}$$

where the weights are subject to following constraints:

$$\begin{cases}\sum_{j=1}^{k}W_{ij}=1 & \text{if } X_j\in N_i(X_i)\\ W_{ij}=0 & \text{if } X_j\notin N_i(X_i)\end{cases} \tag{2}$$

where $N_i(X_i)$ denotes the k nearest neighbors of point X_i.

Clearly, minimizing ε_i subject to the above constraints is a constrained least squares problem. After repeating Steps 1 and 2 for all the n data points, the reconstruction weights consist of a weight matrix $W=[W_{ij}]_{n\times n}$.

Step 3 in the LLE algorithm computes the optimal low dimensional embedding Y based on the weight matrix W obtained from Step 2. This means solving Eq. (3) under the constraints of $Y_{d\times n}e_n=0_d$ and $Y_{d\times n}Y_{d\times n}^{\mathrm{T}}=nI_{d\times d}$. These constraints are appended to remove the translational degree of freedom and the rotational degree of freedom, respectively.

$$\begin{aligned}\varepsilon(Y)&=\arg\min\sum_i\left\|Y_i-\sum_{j=1}^{k}W_{ij}Y_j\right\|^2\\&=\arg\min\operatorname{tr}\left\{\sum_{ij}Y_j(\delta_{ij}-W_{ij})(\delta_{ij}-W_{ij})^{\mathrm{T}}Y_i^{\mathrm{T}}\right\}\end{aligned} \tag{3}$$

So based on the weighted matrix W, a sparse, symmetric and positive semi-definite matrix M can be defined as follows:

$$M=(I-W)^{\mathrm{T}}(I-W) \tag{4}$$

Thus, Eq. (3) can be expressed in a quadratic form $\varepsilon(Y)=\operatorname{tr}\{\sum_{ij}M_{ij}Y_i^{\mathrm{T}}Y_j\}=\operatorname{tr}\{YMY^{\mathrm{T}}\}$, where $M=[M_{ij}]_{n\times n}$. By the Rayleigh–Ritz theorem, minimizing Eq. (3) can be performed by finding the eigenvectors with the smallest (nonzero) eigenvalues of the sparse matrix M.

3. Locally linear discriminant embedding

3.1. The goal of LLDE

For visualization, the goal of dimensionality reduction methods is to map the original data set into a (2-D or 3-D) space that preserves the intrinsic structure as well as possible. But for classification, it aims to project the data into a feature space in which the members from different classes could be clearly separated. LLE is an effective dimensionality reduction approach to visualize the high dimensional data in a 2-D space. However, little classification ability can be displayed by implementing the original LLE. Fig. 1 shows the 2-D visualization results by carrying out the classical LLE to a synthetic data. Each point is clearly located but the features extracted by LLE cannot be automatically distinguished from their class information.

Based on the fact mentioned above, in this paper, we propose a supervised LLE algorithm named as LLDE. The goal of LLDE is to take full advantage of the class information to improve the classification ability of the original LLE. It is well known that the reconstructing weights are invariant to translation under sum-to-one constraint in the original LLE, which can be confirmed by

$$\Phi(Y)=\sum_i\left\|Y_i-\sum_j W_{ij}Y_j\right\|^2=\sum_i\left\|(Y_i-T_i)-\sum_j W_{ij}(Y_j-T_i)\right\|^2 \tag{5}$$

where T_i is a translation vector corresponding to class i. Thus each point with the same class can be translated by the same vector T_i and so does the points belonging to different labels with the corresponding vectors.

Fig. 2 shows the visualization results after translation, where it can be seen that the discriminability is improved evidently. Moreover, a rescaling coefficient is introduced to the proposed algorithm and the discriminability is also improved, which can be found in Fig. 3.

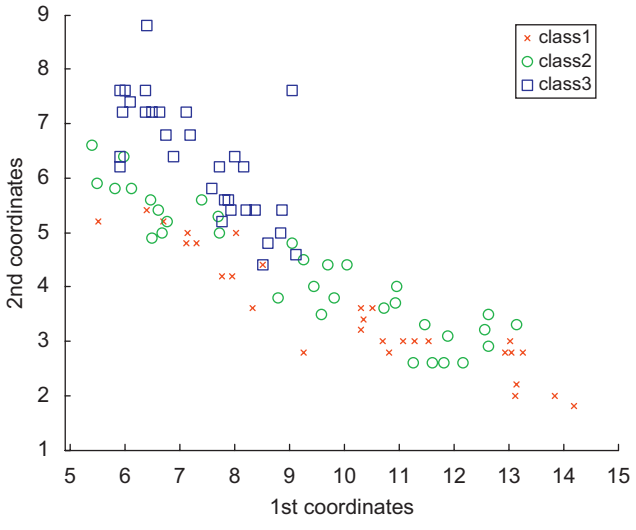

Fig. 1. Data visualization in 2-D space.

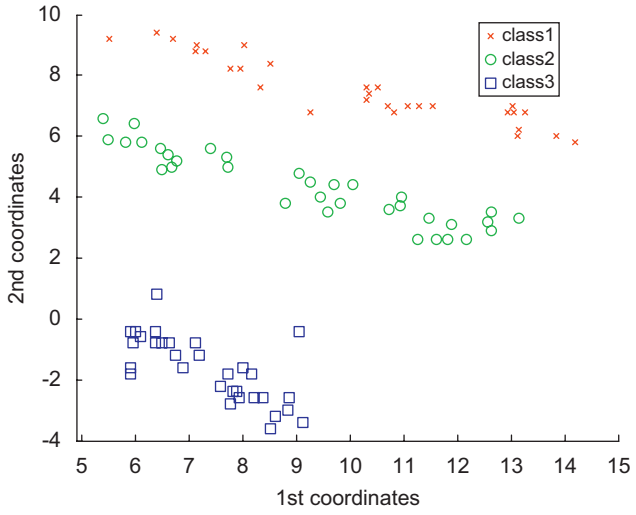

Fig. 2. Data visualization in 2-D space after translation.

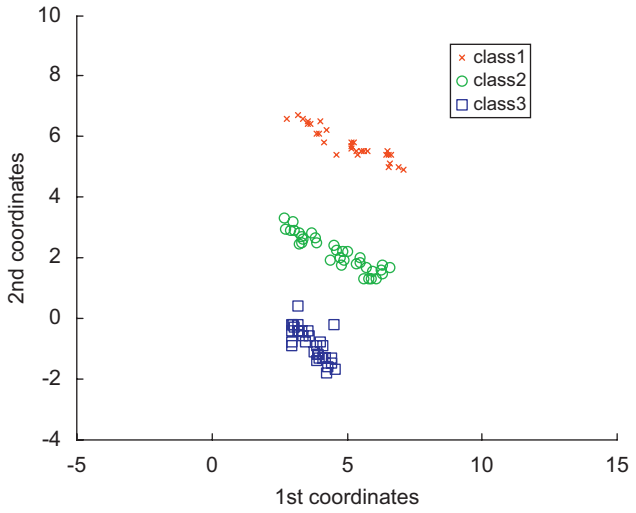

Fig. 3. Data visualization in 2-D space after translation and rescaling.

It should be noted that translations and rescalings can improve the recognition accuracy of the classical LLE significantly. However, when applying the proposed algorithm to real world data, how to explore the optimal translated vectors and rescaling coefficient is still an open problem. So an MMMC is proposed to find the optimal translated vectors and rescaling coefficient, which will be explored by transformation. Furthermore, we define the transformation to be a linear one. Thus the out-of-sample problem can be successfully avoided and the computational cost will be reduced. In the following subsection, a theoretical analysis will be made to the proposed LLDE.

3.2. Analysis to LLDE

3.2.1. A linear approximation to the original LLE

In order to overcome the out-of-sample problem, a linear transformation, i.e. $Y=A^{\mathrm{T}}X$, is plugged. Thus the objective function of the original LLE can be changed into the following form:

$$J_1(A)=\min \operatorname{tr}\{YMY^{\mathrm{T}}\}=\min \operatorname{tr}\{A^{\mathrm{T}}XMX^{\mathrm{T}}A\} \tag{6}$$

Some studies have found that a linear version of LLE shows better recognition ability than the original LLE [23,33]. However, the linear transformation is not always the optimal one that the proposed LLDE pursues. That is to say, LLDE needs a criterion that can be used to automatically find an optimal linear transformation for classification.

3.2.2. Modified maximizing margin criterion

Recently, an MMC was proposed to determine the optimized subspace, which can also successfully conquer the SSS problem [34,8]. The objective function of MMC is listed below:

$$J_2=\max\left\{\sum_{ij}p_ip_j(d(m_i,m_j)-s(m_i)-s(m_j))\right\} \tag{7}$$

where p_i and p_j are the prior probability of class i and class j, m_i and m_j are the centroids of class i and class j. $d(m_i,m_j)$, $s(m_i)$ and $s(m_j)$ have the following definitions:

$$d(m_i,m_j)=\|m_i-m_j\| \tag{8}$$

$$s(m_i)=\operatorname{tr}(S_i) \tag{9}$$

$$s(m_j)=\operatorname{tr}(S_j) \tag{10}$$

Thus the optimized function can be derived as follows:

$$J_2=2\max \operatorname{tr}(S_b-S_w) \tag{11}$$

In order to take advantage of the property of the weighted matrix's invariance to rescaling, the rescaling parameter μ is introduced. Here, we first translated the data to suitable places, and then rescaled the data with the same label to their centroids and all the centroids were kept unchanged. So S_b was still preserved and S_w was rescaled. The derivations are stated below:

$$S'_b=\sum_{i=1}^{c}n_i(m'_i-m')(m'_i-m')^{\mathrm{T}}=\sum_{i=1}^{c}n_i(m_i-m)(m_i-m)^{\mathrm{T}}=S_b \tag{12}$$

$$S'_w=\sum_{i=1}^{n_i}(X'_i-m'_i)(X'_i-m'_i)^{\mathrm{T}}=\sum_{i=1}^{c}\mu(X_i-m_i)(X_i-m_i)^{\mathrm{T}}=\mu S_w \tag{13}$$

where μ denotes the rescaling coefficient and $\mu>0$.

Then the MMMC can be obtained and rewritten in the following form:

$$J_3=\max \operatorname{tr}(S_b-\mu S_w) \tag{14}$$

If a linear transformation $Y=U^{\mathrm{T}}X$ can maximize Eq. (14), an optimal subspace for classification will be explored. This is because the linear transformation aims to project a pattern closer to those with the same class but farther from patterns in different labels, which is just the goal for classification. Under such circumstance, the distance between different centroids will be larger and the within-class scatters will be smaller. In other words, $d(m_i,m_j)$ will be maximized and $s(m_i)$ will be minimized simultaneously in this subspace. Thus $\sum_{ij}p_ip_j(d(m_i,m_j)-s(m_i)-s(m_j))$ and $\operatorname{tr}(S_b-\mu S_w)$ will be maximized accordingly. That is to say, to find an optimal linear subspace for classification means to maximize the following optimized function:

$$J_2=\operatorname{tr}\{U^{\mathrm{T}}(S_b-\mu S_w)U\} \tag{15}$$

3.2.3. Discriminant feature extraction

Let $T_{d\times n}$ denote the translation matrix; thus the discriminant component after performing LLDE can be represented as $[Y-T]_{d\times n}$, which can be also represented by a linear transformation, i.e. $Y-T=V^{\mathrm{T}}X$.

Based on the analysis mentioned above, it can be found that the linear approximation to the original LLE explores a linear subspace with the least reconstructed error. Moreover, under the sum-to-one

constraint, this linear transformation can translate the points to random places, which impacts on the recognition rate of the data greatly. In other words, the linear approximation to LLE can improve the discriminability of the data. However, the projection cannot be ensured to be optimal. At the same time, the MMMC presented above can map the data into an optimal subspace for classification. That is to say, if the linear transformation obtained by linearized LLE can satisfy Eq. (15) simultaneously, the discriminability of the data will be improved greatly. Thus the problem can be represented as the following multi-object optimized problem:

$$\begin{cases} \min \operatorname{tr}\{V^{\mathrm{T}} X M X^{\mathrm{T}} V\} \\ \max \operatorname{tr}\{V^{\mathrm{T}}(S_b - \mu S_w)V\} \end{cases} \tag{16}$$

Moreover, there are two constraints in LLE, that is

$$V^{\mathrm{T}} X X^{\mathrm{T}} V = nI \tag{17}$$

$$Y_{d\times n} e_n = 0_d \tag{18}$$

In the proposed algorithm, we delete the last constraint, which will remove translational degree of freedom. The reason lies in that in the proposed algorithm, the translations are adopted to improve the classification ability of the original LLE. So Eq. (16) can be deduced to solve the following constrained optimized problem:

$$\begin{cases} \min \quad \operatorname{tr}\{V^{\mathrm{T}} X M X^{\mathrm{T}} V\} \\ \max \quad \operatorname{tr}\{V^{\mathrm{T}}(S_b - \mu S_w)V\} \\ \text{s.t.} \quad V^{\mathrm{T}} X X^{\mathrm{T}} V = nI \end{cases} \tag{19}$$

The constrained multi-object optimized function is intent on minimizing the reconstructed error and maximizing the margin between difference classes simultaneously. So it can be changed into the following constrained problem:

$$\begin{aligned} &\min \quad \operatorname{tr}\{V^{\mathrm{T}}(XMX^{\mathrm{T}} - (S_b - \mu S_w))V\} \\ &\text{s.t.} \quad V^{\mathrm{T}} X X^{\mathrm{T}} V = nI \end{aligned} \tag{20}$$

To solve the above optimization problem, we use the Lagrangian multiplier:

$$\frac{\partial}{\partial V} \operatorname{tr}\{V^{\mathrm{T}}(XMX^{\mathrm{T}} - (S_b - \mu S_w))V - \lambda(V^{\mathrm{T}} X X^{\mathrm{T}} V - nI)\} = 0 \tag{21}$$

Thus we can get

$$(XMX^{\mathrm{T}} - (S_b - \mu S_w))V = \lambda X X^{\mathrm{T}} V \tag{22}$$

where λ_i is the generalized eigenvalue of $(XMX^{\mathrm{T}} - (S_b - \mu S_w))$ and XX^{T}, V_i is the corresponding eigenvector. Therefore, the objective function is minimized when V is composed of the first d smallest eigenvectors of the above generalized eigen-decomposition.

3.2.4. The outline of LLDE

The LLDE algorithm can be summarized as follows:

Step 1: For each data point X_i, identify its $k$nearest neighbors by *kNN* algorithm or ε-*ball* algorithm.

Step 2: Compute the reconstruction weights of each point X_i to minimize the error of linearly reconstructing X_i with its k nearest neighbors based on $\varepsilon_i(W) = \arg\min \|X_i - \sum_{j=1}^{k} W_{ij} X_j\|^2$.

Step 3: Repeat Step 3 for all the points and obtain the weighted matrix $W = [W_{ij}]_{n\times n}$.

Step 4: Construct matrix M based on Eq. (4).

Step 5: Construct matrix XMX^{T}.

Step 6: Compute the between-class scatter S_b and within-class scatter S_w and their weighted difference $S_b - \mu S_w$, respectively.

Step 7: Compute the d bottom generalized eigenvalues and the corresponding eigenvectors matrix V of $(XMX^{\mathrm{T}} - (S_b - \mu S_w), XX^{\mathrm{T}})$, and obtain d dimensional embedding $Y = V^{\mathrm{T}} X$.

Step 8: Adopt a suitable classifier to classify the embedding results.

4. Experiments

In this section, the performance of LLDE is evaluated on two different data sets and compared with the performances of MMC, En-ULLELDA and LLDA. LLDA is a newly proposed local linear manifold learning method [35]. Firstly, it clusters the points by k-means method. Secondly, local LDA is applied to each cluster and the corresponding local between-class scatter and within-class scatter are achieved. At last, a linear transformation is attained based on some optimized function constructed by those local between-class scatters and within-class scatters. In the experiments, the first data set is the Yale face database and the second one is CMU PIE face data. In all the experiments, preprocessing is performed to crop the original images. For face data, the original images were normalized such that the two eyes were aligned at the same position, then the facial areas were cropped into the final images for matching. The size of each cropped image in the first experiment is 64×64 pixels and 32×32 pixels in the second experiment with 256 gray levels per pixel. The number of nearest neighbors for constructing the nearest neighbor graph in the proposed algorithm was set from 3 to 7 according to the size of the training set. After MMC, En-ULLELDA, LLDA and the proposed algorithm have been applied to extract features, different pattern classifiers can be adopted for recognition, including k-NN [36], Bayesian [37], support vector machine (SVM) [38], etc. In this study, we apply the 1-NN classifier for its simplicity. The Euclidean metric is used as our distance measure. However, there might be some more sophisticated and better distance metric, e.g. variance normalized distance and cosine distance, which may be used to improve the recognition performance. Then the experiments can be carried out to test the effectiveness of the proposed algorithm.

4.1. Experiment using the Yale face database

The Yale face database [39] was constructed at the Yale Center for Computation Vision and Control. There are 165 images of about 15 individuals in Yale face data sets, where each person has 11 images. The images demonstrate variations in lighting condition (left-light, center-light, right-light), facial expression (normal, happy, sad, sleepy, surprised and wink), and with or without glasses. Shown in Fig. 4 is one cropped object from Yale database.

Firstly, we randomly select the six images as training sets and the rest five images as test sets for each class. Fig. 5 shows the best mean recognition rates for 20 times. It can be found that our proposed method outperforms the other techniques. The recognition rate approaches the maximal average results at 96.82(±1.82)%, 89.33(±2.12)%, 90.74(±2.38)% and 92.46(±2.53)% for the proposed algorithm, MMC, En-ULLELDA and LLDA, respectively.

Fig. 4. Sample images of one person in Yale database.

Secondly, we selected different dimensions after performing MMC, En-ULLELDA, LLDA and LLDE. Fig. 6 shows the recognition rate curves corresponding to different feature extraction methods. At the beginning, with the increase in dimensions, the recognition rates also improved. However, the trend is not maintained for all the dimensions. When they attain theirs tops at 22, 24, 16 and 14 dimensions for MMC, En-ULLELDA, LLDA and LLDE, respectively, all recognition rate curves begin to decrease with the increase in the dimensions.

Thirdly, the experiments are conducted to examine the effect of the training number on the performance. For each feature extraction method, the training sample number is set to 3, 4, 5 and 6, respectively. Accordingly, the rest is test samples. Moreover, all the experimental results are obtained across 20 runs on the Yale database. Table 1 shows the maximal average recognition accuracy and the corresponding standard deviations and the reduced dimensions for MMC, En-ULLELDA, LLDA and LLDE.

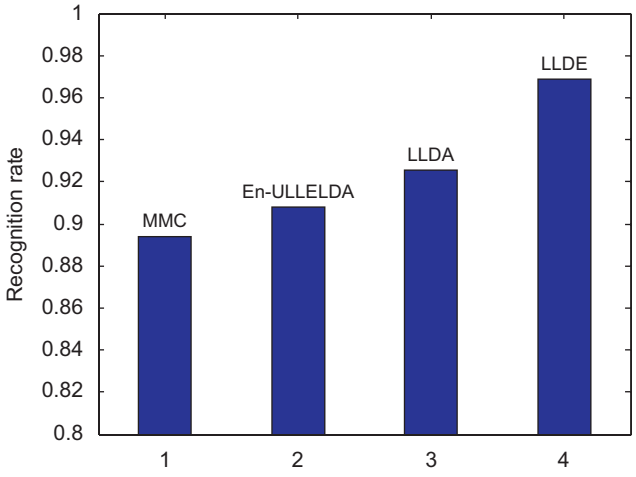

Fig. 5. Performance comparison of recognition rates using MMC, En-ULLELDA, LLDA and LLDE.

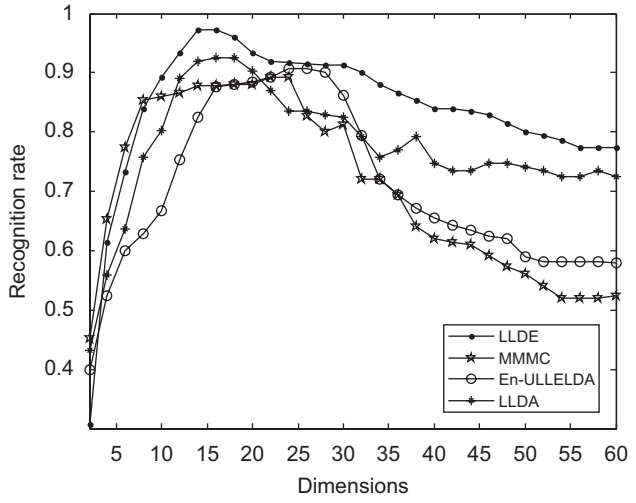

Fig. 6. Performance comparison of recognition rates using MMC, En-ULLELDA, LLDA and LLDE by varying the dimensions.

At last, we test the impact of rescaling coefficient on the recognition rate. The coefficient, i.e. μ, is set to 0.01, 0.1, 1, 10 and 100, respectively. The maximal average recognition rates for different coefficient are stated in Table 2. The optimal recognition rates can be obtained with different coefficients and the corresponding dimensions, for example, when coefficient is 0.01, the recognition rate is 96.82% at 14 dimensions. However, the recognition rate reaches 96.82% at 12 dimensions with μ equaling 100. It can be found that the rescaling coefficient shows few effects on the recognition rate on Yale face database.

4.2. Experiment using the CMU PIE face database

The CMU PIE face database includes 68 subjects with 41,368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination, and expression. We used 170 face images for each individual in our experiment, random 90 images for training and the left 80 face images as the test samples for each person. Fig. 7 displays some samples of one person from CMU PIE database.

The recognition results are shown in Table 3. It is also found that the recognition rate by performing our proposed algorithm outperforms those by applying MMC, En-ULLELDA and LLDA. We investigate the maximal average recognition accuracy at 64, 80, 100 and 84 dimensions for MMC, En-ULLELDA, LLDA and our proposed algorithm, respectively. The best mean recognition rates of MMC, En-ULLELDA, LLDA and the proposed algorithm are 92.66%, 94.63%, 94.98% and 97.13%, and the standard deviations are 1.49%, 1.87%, 2.11% and 1.24%, respectively. The corresponding face subspaces obtained by carrying out the methods mentioned above are called optimal face subspace for each method. In addition, it was also found that there is no significant improvement if more dimensions are used. Fig. 8 shows the plots of recognition rate versus dimensions.

Moreover, the effect of the training sample number is also tested in the following experiment. We randomly selected 60, 70, 80 and 90 training samples and then the rest 110, 100, 90 and 80 samples as test ones. The best mean recognition rates are computed by repeating the experiments 20 times for the corresponding training and test samples, which are shown in Fig. 9.

Table 2
Dimensions versus recognition rate by varying rescaling coefficient in the proposed algorithm

Rescaling coefficient	0.01	0.1	1	10	100
The best average recognitionrate (%)	96.82	96.82	96.82	96.82	96.82
Dimensions	14	14	14	14	12

Table 1
The maximal average recognition rate and the corresponding standard deviations (percent) with the reduced dimensions for MMC, En-ULLELDA, LLDA and LLDE on Yale database

Method	3 Train	4 Train	5 Train	6 Train
MMC	72.23 ± 1.93 (18)	76.19 ± 2.43 (20)	81.13 ± 1.75 (22)	89.33 ± 2.12 (22)
En-ULLELDA	75.64 ± 2.17 (24)	79.59 ± 1.94 (22)	83.33 ± 2.05 (24)	90.74 ± 2.38 (24)
LLDA	74.34 ± 1.69 (16)	81.24 ± 2.56 (16)	85.56 ± 1.86 (16)	92.46 ± 2.53 (16)
LLDE	76.67 ± 2.36 (16)	83.76 ± 2.17 (14)	88.89 ± 1.52 (14)	96.82 ± 1.82 (14)

Fig. 7. The cropped sample face images of one person from CMU PIE database.

Table 3
Performance comparison and the corresponding standard deviations with the reduced dimensions for MMC, En-ULLELDA, LLDA and LLDE on CMU PIE face

Approach	Dimensions	Recognition rate (%)
MMC	64	92.66 ± 1.49
En-ULLELDA	80	94.63 ± 1.87
LLDA	100	94.98 ± 2.11
LLDE	84	97.13 ± 1.24

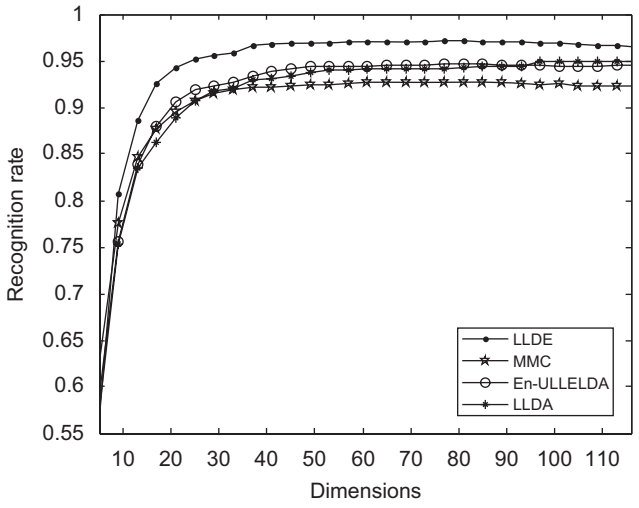

Fig. 8. Performance comparison of recognition rates with different dimensions using MMC, En-ULLELDA, LLDA and LLDE on CMU PIE face.

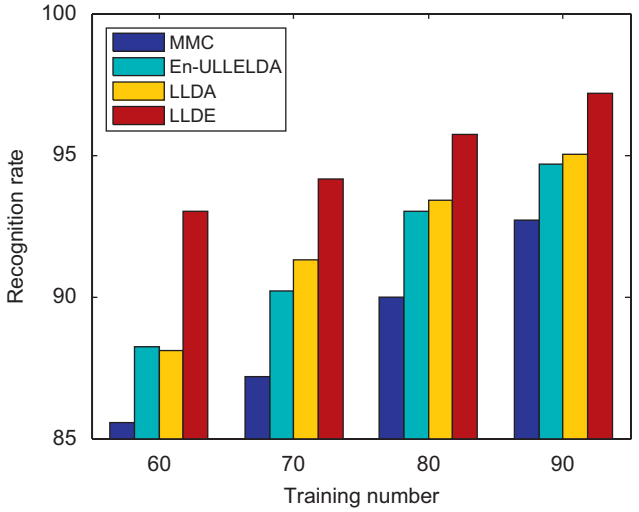

Fig. 9. Performance comparison of recognition rates with different training sample number using MMC, En-ULLELDA, LLDA and LLDE on CMU PIE face.

5. Discussions and conclusions

5.1. Discussions

From the experimental results mentioned above, we can find some interesting points as follows:

(1) Compared to some feature extraction methods, the proposed one can gain better recognition rate. LLDE is a local linear manifold learning-based method. When applied to the data lying on a manifold, the proposed method can extract features efficiently. However, when using the linear methods to extract features, the global nonlinear structure of nonlinear data will be destroyed so that the recognition rate is reduced. The data sets used in this study are Yale and CMU PIE face; the images for each person are varied from pose, illumination to facial expression. Some research efforts have also shown that various conditions such as lighting, pose, expression and so on are essential features for a sub-manifold [17–19,40–42]. That is to say, manifold learning-based methods can successfully explore these essential features in a high dimensional face space. Thus manifold learning methods are superior to some linear feature extraction methods. Moreover, compared to other supervised manifold learning techniques, on the one hand, the proposed LLDE takes full advantages of the property that the original LLE is invariant to translations and rescalings; on the other hand, the translations and rescalings can be automatically determined by an MMMC instead of being randomly set. The proposed MMMC aims to separate data with different labels farther and cluster data with the same label closer. Thus the proposed algorithm can gain better recognition rate.

(2) Rescaling and translation are contained in the proposed algorithm. The between-scatter matrix S'_b has been changed although the within-scatter matrix S'_w has still been kept after translation, which can be found from the following derivations:

$$S'_b = \sum_{i=1}^{c} n_i((m_i - T_i) - (m - t))((m_i - T_i) - (m - t))^{\mathrm{T}}$$

$$S'_w = \sum_{i=1}^{n_i}((X_i - T_i) - (m_i - T_i))((X_i - T_i) - (m_i - T_i))^{\mathrm{T}}$$

$$= \sum_{i=1}^{n_i}(X_i - m_i)(X_i - m_i)^{\mathrm{T}} = S_w$$

where

$$t = \frac{1}{n}\sum_{i=1}^{c} n_i T_i.$$

If we rescale the data set, the within-scatter matrix will be changed, which can be found from Eq. (14). Compared to changing the within-scatter matrix, the contribution for improving the discriminability will be bigger by changing the between-class scatter

matrix. This is because rescaling cannot change the distances between centroids of different classes, i.e. S_b. In order to map the data belonging to different labels farther, the translations are taken into the proposed algorithm, which is a key to enhancing the discriminability of the data. Moreover, the rescalings are also adopted to cluster the data closer, which helps the data to be recognized.

6. Conclusions

In appearance-based face recognition, feature extraction techniques are widely employed to reduce the dimensions and to enhance the discriminability of the original data. In this paper, a discriminant method based on the classical LLE is presented. The proposed approach can effectively extract the most discriminant features. Compared to other feature extraction algorithms, the new technique does not suffer from the SSS problem, the problem of dimensionality reduction two times and the disconnected component problems. The experimental results show that the new method is effective.

Acknowledgments

This work was supported by the grants of the National Science Foundation of China, nos. 60705007 and 30700161, the grant from the National Basic Research Program of China (973 Program), no. 2007CB311002, the grants from the National High Technology Research and Development Program of China (863 Program), nos. 2007AA01Z167 and 2006AA02Z309, the grant of the Guide Project of Innovative Base of Chinese Academy of Sciences (CAS), no. KSCX1-YW-R-30, and the grant of Oversea Outstanding Scholars Fund of CAS, no. 2005-1-18.

References

[1] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[2] K. Etemad, R. Chellapa, Discriminant analysis for recognition of human face images, J. Opt. Am. A 14 (8) (1997) 1724–1733.
[3] A.M. MartõÂnez, A.C. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233.
[4] P. Howland, J. Wang, H. Park, Solving the small sample size problem in face recognition using generalized discriminant analysis, Pattern Recognition 39 (2006) 277–287.
[5] W. Zheng, L. Zhao, C. Zou, Foley–Sammon optimal discriminant vectors using kernel approach, IEEE Trans. Neural Networks 16 (1) (2005) 1–9.
[6] W. Zheng, L. Zhao, C. Zou, An efficient algorithm to solve the small sample size problem for LDA, Pattern Recognition 37 (2004) 1077–1079.
[7] J.H. Friedman, Regularized discriminant analysis, J. Am. Statist. Assoc. 84 (405) (1989) 165–175.
[8] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[9] P.N. Belhumeour, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[10] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.
[11] P. Howland, M. Jeon, H. Park, Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition, SIAM J. Matrix Anal. Appl. 25 (1) (2003) 165–179.
[12] J. Ye, R. Janardan, C.H. Park, H. Park, An optimization criterion for generalized discriminant analysis on undersampled problems, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8) (2004) 982–994.
[13] J. Ye, Q. Li, A two-stage linear discriminant analysis via QR-decomposition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 929–941.
[14] L.-F. Chen, X. Hong-Yuan, M. Liao, M.-T. Ko, J.-C. Lin, G.-J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (2000) 1713–1726.
[15] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[16] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, K.-R. Muller, Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (5) (2003) 623–628.
[17] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[18] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[19] L.K. Saul, S.T. Roweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, J. Mach. Learn. Res. (4) (2003) 119–155.
[20] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, vol. 14, MIT Press, Cambridge, MA, USA, 2002, pp. 585–591.
[21] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[22] Y. Bengio, J.-F. Paiement, P. Vincent, Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering, Technical Report 1238, D'epartement d'Informatique et Recherche Op'erationnelle, July 25, 2003.
[23] E. Kokiopoulou, Y. Saad, Orthogonal neighborhood preserving projections, in: Proceedings of the Fifth IEEE International Conference on Data Mining, 2005, pp. 1–7.
[24] D. de Ridder, R.P.W. Duin, Locally linear embedding for classification, Technical Report PH-2002-01, Pattern Recognition Group, Department of Imaging Science and Technology, Delft University of Technology, Delft, The Netherlands, 2002.
[25] D. de Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, R.P.W. Duin, Supervised locally linear embedding, artificial neural networks and neural information processing, in: ICANN/ICONIP 2003 Proceedings, Lecture Notes in Computer Science, vol. 2714, Springer, Berlin, 2003, pp. 333–341.
[26] X. Bai, B. Yin, Q. Shi, Y. Sun, Face recognition based on supervised locally linear embedding method, J. Inf. Comput. Sci. (4) (2005) 641–646.
[27] O. Kouropteva, O. Okun, M. Pietikainen, Supervised locally linear embedding algorithm for pattern recognition, IbPRIA 2003, Lecture Notes in Computer Science, vol. 2652, Springer, Berlin, 2003, pp. 386–394.
[28] M. Polito, P. Perona, Grouping and dimensionality reduction by locally linear embedding, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, vol. 14, MIT Press, Cambridge, MA, USA, 2002, pp. 1255–1262.
[29] J. Zhang, H. Shen, Z.-H. Zhou, Unified Locally Linear Embedding and Linear Discriminant Analysis Algorithm for Face Recognition, Lecture Notes in Computer Science, Springer, Berlin, 2004.
[30] J. Zhang, H. Shen, Z.-H. Zhou, Ensemble-based discriminant manifold learning for face recognition, ICNC 2006, Part I, Lecture Notes in Computer Science, vol. 4221, Springer, Berlin, 2006, pp. 29–38.
[31] Y. Pang, Z. Liu, N. Yu, A new nonlinear feature extraction method for face recognition, Neurocomputing 69 (2006) 949–953.
[32] D. de Ridder, M. Loog, M.J.T. Reinders, Local Fisher embedding, in: The 17th International Conference on Pattern Recognition, vol. 2, 2004, pp. 295–298.
[33] X. He, D. Cai, S. Yan, H.-J. Zhang, in: Tenth IEEE International Conference on Computer Vision (ICCV'2005), Beijing, China, October 2005.
[34] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Networks 17 (1) (2006) 157–165.
[35] T.K. Kim, J. Kittler, Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 318–327.
[36] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[37] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 696–710.
[38] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 696–710.
[39] Yale University Face Database, ⟨http://cvc.yale.edu/projects/yalefaces/yalefaces.html⟩, 2002.
[40] Y. Chang, C. Hu, M. Turk, Manifold of facial expression, in: Proceedings of the IEEE International Workshop Analysis and Modeling of Faces and Gestures, October 2003.
[41] H.S. Seung, D.D. Lee, The manifold ways of perception, Science 290 (2000) 2268–2269.
[42] K.-C. Lee, J. Ho, M.-H. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2003, pp. 313–320.
[43] D. DeCoste, Visualizing Mercer kernel feature spaces via kernelized locally-linear embeddings, in: Proceedings of the Eighth International Conference on Neural Information Processing, Shanghai, China, 14–18 November 2001.

About the Author—BO LI was born in Hubei province, China, in 1975. He received M.Sc. degree in Mechanical and Electronic Engineering in 2003, from Wuhan University of Technology. He is now in pursuit for Ph.D. degree in Pattern Recognition and Intelligent System in University of Science and Technology of China. His research interests include pattern recognition, manifold learning and image processing.

About the Author—CHUN-HOU-ZHENG received Ph.D. degree in Pattern Recognition & Intelligent System, from University of Science and Technology of China, in 2006. He is currently an associate professor at College of Information and Communication Technology, Qufu Normal University. His research interests include Artificial Neural Networks, Intelligent Computing, and Intelligent Information Processing.

About the Author—DE-SHUANG HUANG (SM'98) received the B.Sc., M.Sc. and Ph.D. degrees all in Electronic Engineering from Institute of Electronic Engineering, Hefei, China, National Defense University of Science and Technology, Changsha, China and Xidian University, Xian, China, in 1986, 1989 and 1993, respectively. During 1993–1997 he was a postdoctoral student, respectively, in Beijing Institute of Technology and in National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In September 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences as the Recipient of "Hundred Talents Program of CAS". From September 2000 to March 2001, he worked as Research Associate in Hong Kong Polytechnic University. From April 2002 to June 2003, he worked as Research Fellow in City University of Hong Kong. From August to September 2003, he visited the George Washington University as visiting professor, Washington DC, USA. From October to December 2003, he worked as Research Fellow in Hong Kong Polytechnic University. From July to December 2004, he worked as the University Fellow in Hong Kong Baptist University. Dr. Huang is currently a senior member of the IEEE.