Neural Process Lett (2016) 43:1–16
DOI 10.1007/s11063-014-9398-x

# Embedded Manifold-Based Kernel Fisher Discriminant Analysis for Face Recognition

**Guoqiang Wang · Nianfeng Shi · Yunxing Shu · Dianting Liu**

Published online: 15 November 2014

**Abstract** Manifold learning algorithms mainly focus on discovering the intrinsic low-dimensional manifold embedded in the high-dimensional Euclidean space. Among them, locally linear embedding (LLE) is one of the most promising dimensionality reduction methods. Though LLE holds local neighborhood information, it doesn't fully take the label information and the global structure information into account for classification tasks. To enhance classification performance, this paper proposes a novel dimensionality reduction method for face recognition, termed embedded manifold-based kernel Fisher discriminant analysis, or EMKFDA for short. The goal of EMKFDA is to emphasize the local geometry structure of the data while utilizing the global discriminative structure obtained from linear discriminant analysis, which can maximize the between-class scatter and minimize the within-class scatter. In addition, by optimizing an objective function in a kernel feature space, nonlinear features can be extracted. Thus, EMKFDA, which combines manifold criterion and Fisher criterion, has better discrimination, and is more suitable for recognition tasks. Experiments on the ORL, Yale, and FERET face databases show the impressive performance of the proposed method. Results show that this proposed algorithm exceeds other popular approaches reported in the literature and achieves much higher recognition accuracy.

**Keywords** Face recognition · Dimensionality reduction · Manifold learning · Locally linear embedding · Kernel discriminant analysis

## 1 Introduction

Face recognition is one of the most active and challenging research topics in computer vision and pattern recognition due to its wide-ranging applications in many areas, such as identity authentication, access control, surveillance, and human-computer interaction.

G. Wang · N. Shi · Y. Shu (✉)
Department of Computer and Information Engineering, Luoyang Institute of Science and Technology, Luoyang 471023, People's Republic of China
e-mail: wgq2211@163.com

D. Liu
Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA

During the past two decades, considerable progress has been made in face recognition and many new methods have been proposed. Up to now, the appearance-based method with subspace learning is the main subject and direction for current research. For the appearance-based method, facial images can be modeled as points in high-dimensional image space, causing the curse of dimensionality. A common way to resolve this problem is to project the data onto low-dimensional subspace through dimensionality reduction techniques. Wang et al. [1] presents four reasons for dimensionality reduction of the observation data: (1) to compress the data to reduce the storage requirement; (2) to extract features from the data for face recognition; (3) to eliminate noise; and (4) to project the data to a lower-dimensional space so as to be able to discern the data distribution.

In many existing dimensionality reduction algorithms, Principal Component Analysis (PCA) [2] and Linear Discriminant Analysis (LDA) [3] are two of the most popular dimensionality reduction methods. PCA aims to find a set of mutually orthogonal basis vectors that capture the directions of maximum variance in the data. PCA is optimum for the minimization of reconstruction error. However, it is not ideal for classification purpose because it doesn't take the class information of the sample into account. In contrast with PCA, LDA is a supervised learning method. LDA can produce an optimal discriminative subspace by maximizing the ratio of the trace of the between-class scatter to the trace of the within-class scatter. Because the class information is available, LDA usually outperforms PCA for recognition tasks. However, LDA is optimal only when the data of each class is in an approximate Gaussian distribution. Since the dimensionality of samples is usually higher than the number of training samples, an intrinsic limitation of traditional LDA is that it fails to work when the within-class scatter matrix becomes singular, which is known as the small sample size (SSS) problem. To address this problem, many effective and efficient methods have been proposed in the literature [4–7].

Recently, a lot of studies have shown that facial images may lie on a nonlinear submanifold [11,26,27]. Although PCA, LDA, and their extension methods have been successfully applied to linear data, they are all designed to deal with the flat global Euclidean structure and may fail to discover the essential curved or nonlinear structure of the original input data with nonlinear distribution. In order to overcome this problem, a number of nonlinear dimensionality reduction methods, including kernel based techniques and manifold learning based ones, have been developed. The basic idea of kernel based techniques is to implicitly map the observed patterns into potentially much higher dimensional feature space by a nonlinear mapping determined by a kernel trick. It is possible that the nonlinear structure data will be linearly separable in the kernel space. The most widely used kernel techniques are Kernel Principal Component Analysis (KPCA) [8] and Kernel Discriminant Analysis (KDA) [9], which can be viewed as the kernel versions of PCA and LDA. KPCA and KDA have been proven effective in real world applications.

Unlike kernel based techniques, manifold learning based methods are straightforward in finding the intrinsic low-dimensional nonlinear data structure hidden in the observation space. In addition, kernel based techniques do not consider the structure of the manifold on which the images may lie. In the past few years, many manifold learning based methods have been presented. Among the various methods, the most well-known ones include ISOmetric feature MAPping (ISOMAP) [10], Locally Linear Embedding (LLE) [11], Laplacian Eigenmaps (LE) [12,13], Maximum Variance Unfolding (MVU) [14], Local Tangent Space Alignment (LTSA) [15], Local Coordinates Alignment (LCA) [16], and Local Spline Embedding (LSE) [17]. These algorithms differ in the representations of the local geometries that they attempt to preserve. ISOMAP preserves the global geodesic distances of all pairs of samples. LLE is based on the linear coefficients of local reconstructions. LE preserves proximity relations based on the undirected weighted graph. LTSA uses the local tangent space as a representation of local structure. LCA obtains the local coordinates as representations of a local neighborhood. LSE exploits the local tangent space as a representation of the local geometry. These manifold learning methods yield impressive results on some benchmark artificial and real world datasets. However, they yield maps that are defined only on the sample data, and it is unclear how to construct the mapping function for out-of-sample data. Later, zhang et al. [33] proposed a unified patch alignment framework for subspace learning. It divided the subspace learning into the patch optimization step and whole alignment step. Most subspace learning methods, such as PCA, LDA, LLE, ISOMAP, and LE, can be formulated in this framework. All algorithms have an almost identical whole alignment stage and intrinsic differences of them are how to build patches and the associated optimization.

Recently, He et al. proposed Neighborhood Preserving Embedding (NPE) [18] and Locality Preserving Projections (LPP) [19], which are linear manifold learning methods, derived from LLE and LE respectively. However, these methods are only designed to best preserve data locality or similarity in the embedding space. For face recognition problems, these methods are not optimal because they ignore the class label information of the sample data, which can be effectively applied to further enhance classification performance. Based on the idea of LPP, many extension manifold learning methods such as Local Discriminant Embedding (LDE) [20], Marginal Fisher Analysis (MFA) [21], and Locality Sensitive Discriminant Analysis (LSDA) [22] are proposed to address the classification problem. In addition, based on LLE and Maximum Margin Criterion [30], Li et al. proposed local linear discriminant embedding (LLDE), which is a supervised learning method [28]. Although these methods are successful in many circumstances, they often fail to deliver good performance when facial images are subject to complex nonlinear changes due to large pose, expression, or illumination variations, for they are linear methods in nature.

Global dimensionality reduction techniques such as PCA, LDA, KPCA, and KDA do not preserve the local manifold of the facial structure, whereas the local dimensionality reduction techniques such as LLE, ISOMAP, LE, and LTSA don't preserve global features of facial images. To take advantage of these methods while avoiding their disadvantages, we present a new algorithm called Embedded Manifold-based Kernel Fisher Discriminant Analysis, or EMKFDA. The objective function of EMKFDA is formulated by combining the ideas of LLE and LDA. That is to say, EMKFDA builds patches globally with considering the local geometry. By mapping the input data to a high-dimensional feature space using the kernel technique, nonlinear features are extracted. EMKFDA can not only preserve local geometry information, but also preserve global structure information. At the same time, EMKFDA captures the discriminate features among the samples, considering the different classes in the subjects, which produces the considerably improved results in face recognition. Experiments based on the ORL, Yale, and FERET face databases demonstrate that the proposed algorithm has more discriminating power than other methods.

The rest of the paper is organized as follows: In Sect. 2, we review briefly the LLE and KDA. The algorithm detail of EMKFDA is introduced in Sect. 3. Experimental results for recognition using the ORL, Yale, and FERET face database are given in Sect. 4. Finally, conclusions are summarized in the last Section.

## 2 Related Work

Before describing the proposed method, we will firstly review briefly LLE and KDA in this section, which are relevant to the proposed method. We begin with a description of the dimensionality reduction problem. Consider $N$ data points $X = [x_1, x_2, \ldots, x_N] \in R^D$ sampled from one underlying manifold $M$. The goal of dimensionality reduction is to map the high-dimensional data into a low-dimensional feature space. Let us denote the corresponding set of $N$ points in the reduced space as $Y = [y_1, y_2, \ldots, y_N] \in R^d$, where $d << D$, in which $y_i$ is a low-dimensional representation of $x_i (i = 1, 2, \ldots, N)$.

### 2.1 Locally Linear Embedding (LLE)

Locally Linear Embedding is a nonlinear manifold learning method. The LLE algorithm is based on simple geometric intuitions, where the input data is composed of $N$ points $x_i,\ x_i \in R^D,\ i \in [1, N]$, each of dimensionality $D$, which were obtained by an underlying manifold. As an output, it provides $N$ points $y_i,\ y_i \in R^d,\ i \in [1, N]$ where $d << D$. Suppose there are sufficient data (such that the manifold is well sampled), each data point and its neighbors are expected to lie on or near a locally linear patch of the manifold. A brief description of LLE algorithm is as follows:

*Step I* For each individual point, its $K$ nearest neighbor points in the original high-dimensional space are found, typically based on the Euclidean distance.

*Step II* After identifying the $K$ nearest neighbor points, this step seeks the optimal reconstruction weights. The reconstruction cost function to be minimized is defined as:

$$\varepsilon(W) = \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{N} W_{ij} x_{ij} \right\|^2, \tag{1}$$

where $x_{ij}$ is the neighbors of $x_i$. Considering the constraint term $\sum_{j=1}^{N} W_{ij} = 1$, and if $x_i$ and $x_{ij}$ are not neighbors, $W_{ij} = 0$, compute the weight matrix $W$ according to the least squares.

*Step III* The weight matrix $W$ is fixed and the embedded coordinates $y_i,\ i = 1, 2, \ldots, N$ are computed by minimizing the following embedding cost function:

$$\phi(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{N} W_{ij} y_{ij} \right\|^2 = tr(YMY^T), \tag{2}$$

subject to $1/N \cdot \sum_{i=1}^{N} y_i y_i^T = I_{d \times d}$, and $\sum_{i=1}^{N} y_i = 0$. In Eq. (2), $Y$ is the output matrix assembled by $y_i$, $M = (I - W)^T (I - W)$ is a sparse, symmetric, and semi-positive definite matrix, $tr$ denotes the operation of trace.

The optimal embedding is provided by the $d$ bottom eigenvectors of $M$ starting from the second one and $Y$ is the rows of these eigenvectors.

The detail of theoretical justification about LLE can be found in Ref. [31].

### 2.2 Kernel Discriminant Analysis (KDA)

Kernel Discriminant Analysis is a nonlinear extension of Linear Discriminant Analysis via kernel trick to deal with nonlinear feature extraction and classification. The main idea of KDA is to first map the original data $x$ into a high-dimensional feature space $F$ by a nonlinear mapping $\phi : x \in R^D \rightarrow \phi(x) \in F$, and then implement LDA in the feature space $F$. Note that the feature space $F$ could have a much higher, possibly infinite, dimensionality.

Suppose there are $c$ known pattern classes $\omega_1, \omega_2, \ldots, \omega_c$ with every class $\omega_i$ having $n_i$ elements where $\sum_{i=1}^{c} n_i = N$, the within-class scatter matrix $S_w^\phi$ and the between-class scatter matrix $S_b^\phi$ in the feature space can be denoted as:

$$S_w^\phi = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\phi(x_j) - \mu_i^\phi)(\phi(x_j) - \mu_i^\phi)^T, \tag{3}$$

$$S_b^\phi = \sum_{i=1}^{c} n_i (\mu_i^\phi - \mu^\phi)(\mu_i^\phi - \mu^\phi)^T, \tag{4}$$

where $\mu_i^\phi$ is the mean vector of class $i$ in the feature space and $\mu^\phi$is the total mean vector in the feature space.

The optimal projection matrix $w$ for KDA can be obtained by maximizing the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix in the feature space $F$:

$$J(w) = \arg\max_w \frac{\left| w^T S_b^\phi w \right|}{\left| w^T S_w^\phi w \right|}. \tag{5}$$

In general, the maximization problem for KDA can be converted to calculating the generalized eigenvectors of the generalized eigen-equation as follows:

$$S_b^\phi w = \lambda S_w^\phi w. \tag{6}$$

According to the theory of reproducing kernel [23], any $w$ must lie in the space, which is spanned by $\phi(x_1), \phi(x_2), \ldots, \phi(x_N)$, i.e.,

$$w = \sum_{i=1}^{N} a_i \phi(x_i) = \phi a. \tag{7}$$

After substituting Eqs. (3), (4), (7) into Eq. (6), we can obtain an equivalent eigenvalue problem as follows:

$$K_b a = \lambda K_w a, \tag{8}$$

where $K_b = \sum\limits_{i=1}^{c} n_i (m_i - m)(m_i - m)^T$, $K_w = \sum\limits_{i=1}^{c} \sum\limits_{j=1}^{n_i} n_i (\xi_j - m_i)(\xi_j - m_i)^T$,

$$\begin{aligned}
\xi_j &= [k(x_1, x_j), k(x_2, x_j), \ldots, k(x_N, x_j)]^T, \\
m_i &= \frac{1}{n_i} \left[ \sum_{j=1}^{n_i} k(x_1, x_j), \sum_{j=1}^{n_i} k(x_2, x_j), \ldots, \sum_{j=1}^{n_i} k(x_N, x_j) \right]^T, \\
m &= \frac{1}{N} \left[ \sum_{i=1}^{N} k(x_1, x_i), \sum_{i=1}^{N} k(x_2, x_i), \ldots, \sum_{i=1}^{N} k(x_N, x_i) \right]^T.
\end{aligned}$$

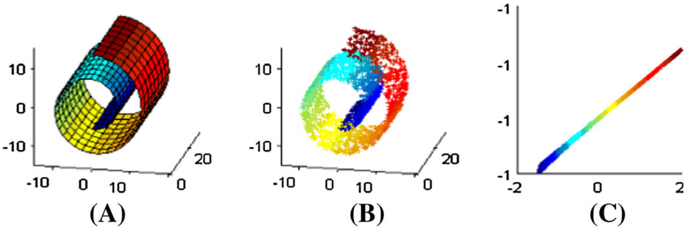

**Fig. 1** An example of the nonlinear manifold of LLE. **a** A two-dimensional manifold. **b** Three–dimensional data sampled from (**a**). **c** The 2D visualization of the 3D manifold by using LLE [11]

## 3 Embedded Manifold-based Kernel Fisher Discriminant Analysis (EMKFDA)

### 3.1 Basic Idea of EMKFDA

Manifold learning methods suppose that the input data lies on or nearly on a low-dimensional manifold embedded in the high-dimensional observation space. For visualization, the goal of manifold learning is to map the original data set into a low-dimensional space that preserves the intrinsic structure as well as possible. For classification, it aims to project the sample data into a feature space in which the samples from different classes could be clearly separated. On the basis of the assumption, Roweis and Saul [11] proposed a nonlinear manifold learning method— LLE. Figure 1 depicts the nonlinear manifold of LLE. LLE has demonstrated excellent results for exploratory analysis and visualization of multivariate data. But it is suboptimal from the perspective of pattern classification. In addition, LLE doesn't preserve the global feature of the input data. EMKFDA aims to make full use of the class information and kernel trick to improve the discriminant ability of the original LLE. EMKFDA combines local geometry structure and global discriminant structure of the data manifold to form the high quality feature set. EMKFDA also uses kernel trick to preserve the nonlinear structure of the data samples. At the same time, EMKFDA retains the local information and global information of the data, which makes the recognition insensitivity to absolute image intensity and insensitivity to contrasting and local facial expressions.

### 3.2 Theoretical Derivation of EMKFDA

Give a data set $X = [x_1, x_2, \ldots, x_N]$ in a $D$-dimensional space. Each data $x_i$ belongs to one of $C$ classes $\{X_1, X_2, \ldots, X_C\}$. Each class contains $n_i$ samples, $i = 1, 2, \ldots, C$. Then the data is mapped into a Hilbert space $F$ through a nonlinear mapping function $\phi : X \to F$. The problem that the proposed algorithm solves is to find a transformation matrix that map the set $X$ to the set $Y = [y_1, y_2, \ldots, y_N]$ in a $d$-dimensional space ($d << D$).

It is well-known that the original LLE algorithm might be unsuitable for pattern recognition task because it yields an embedding only based on the training data set. To begin with, the data is mapped into an implicit space $F$ using a nonlinear function $\phi : X \to F$. In order to circumvent the out-of-sample problem, an explicit linear map in $F$ from $\phi(X)$ to $Y$, i.e. $Y = V^T \phi(X)$, is constructed. The basic idea of LLE is that the same coefficients that reconstruct the point $\phi(x_i)$ in $F$ should also reconstruct its embedded counterpart $y_i$ in $R^d$. Thus the objective function for the original LLE can be converted to the following form:

$$J_1(V) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{K} W_{ij} y_j \right\| = \|Y(I - W)\|^2$$

$$
\begin{aligned}
&= tr(Y(I-W)(I-W)^T Y^T) \\
&= tr(V^T \phi(X) M \phi(X)^T V). \qquad (9)
\end{aligned}
$$

Mapping new data points to the low-dimensional space for LLE becomes trivial once linear transformation matrix $V$ is determined. However, LLE does not take class information and global structure information into account, which are important for face recognition problems. The linear transformation is not always the optimal one that the proposed method pursues. That is to say, we need a new criterion that can be used to automatically find an optimal linear transformation for classification.

To obtain optimal linear discriminant embedding, we introduce Fisher criterion to the objective function of LLE. According to Ref. [19], the total scatter matrix $S_t^\phi$, the within-class scatter matrix $S_w^\phi$, and the between-class matrix $S_b^\phi$ can be formulated, respectively, as follows:

$$
\begin{aligned}
S_t^\phi &= \frac{1}{N}\sum_{i=1}^{N}(\phi(x_i)-\mu)(\phi(x_i)-\mu)^T \\
&= \frac{1}{N}\phi(X)(I-ee^T/N)\phi(X)^T \\
&= \phi(X)G\phi(X)^T, \qquad (10) \\
S_w^\phi &= \sum_{i=1}^{C}\sum_{x\in X_i}(\phi(x)-\mu_i)(\phi(x)-\mu_i)^T \\
&= \phi(X)(I-E)\phi(X)^T = \phi(X)L\phi(X)^T, \qquad (11) \\
S_b^\phi &= S_t^\phi - S_w^\phi \\
&= \phi(X)(G-L)\phi(X)^T \\
&= \phi(X)B\phi(X)^T, \qquad (12)
\end{aligned}
$$

where $G = I - (1/N)ee^T$, $L = I - E$, and $B = G - L$. $I$ represents an identity matrix, $e = [1, 1, \ldots, 1]^T$, and $E_{ij} = 1/n_c$ if $x_i$ and $x_j$ belong to the $c$th class; otherwise, $E_{ij} = 0$.

Thus, Fisher criterion can be rewritten as:

$$
J_2(V) = \frac{tr\left(V^T\phi(X)B\phi(X)^T V\right)}{tr\left(V^T\phi(X)L\phi(X)^T V\right)}. \qquad (13)
$$

By combining Eq. (9) and Eq. (13), we can formulate a final objective function as follows:

$$
J(V) = \max\left(\frac{tr\left(V^T\phi(X)B\phi(X)^T V\right)}{tr\left(V^T\phi(X)L\phi(X)^T V\right)} - \mu\cdot tr\left(V^T\phi(X)M\phi(X)^T V\right)\right), \qquad (14)
$$

where $\mu$ is a parameter to balance the global discriminant information and the local geometry information, and $\mu > 0$.

Since LLE wants to find a projection direction $v$ to make $tr\left(V^T\phi(X)M\phi(X)^T V\right)$ as small as possible, we can instead here choose $-tr\left(V^T\phi(X)M\phi(X)^T V\right)$ and make it as large as possible in the low-dimensional space.

Alternatively, we can reformulate Eq. (14) as:

$$
\begin{aligned}
&J(V) = \max\left(tr\left(V^T\phi(X)B\phi(X)^T V\right) - \mu\cdot tr\left(V^T\phi(X)M\phi(X)^T V\right)\right) \\
&s.t.\quad V^T\phi(X)L\phi(X)^T V = I. \qquad (15)
\end{aligned}
$$

As each column of $V$ should lie in the span of all training samples in $F$, there exist coefficients $a_j$ $(j = 1, 2, \ldots, N)$ such that $v = \sum_{j=1}^{N} a_j \phi(x_j) = \phi(X)a$, where $a = [a_1, a_2, \ldots, a_N]^T$. Therefore, Eq. (15) can be rewritten as:

$$\begin{aligned} &J(V) = \max \left( tr \left( A^T K (B - \mu \cdot M) K A \right) \right) \\ &s.t. \quad A^T KLKA = I, \end{aligned} \tag{16}$$

where $A = [a_1, a_2, \ldots, a_N]$, $K = \phi(X)^T \phi(X)$ is a kernel matrix with elements $K_{ij} = k(x_i, x_j)$, and $k$ is a kernel function which satisfies $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) = \phi(x_i)^T \phi(x_j)$.

The above constrained maximization problem can be solved by Lagrange multiplier method:

$$L(a_i) = a_i^T K (B - \mu \cdot M) K a_i + \lambda_i \left( 1 - a_i^T KLK a_i \right).$$

Compute the gradients with respect to $a_i$ and set the gradients to zero, we have the following maximization eigenvalue problem:

$$K(B - \mu \cdot M) K a_i = \lambda_i KLK a_i. \tag{17}$$

The transformation matrix $A$ can be constituted by the $d$ eigenvectors corresponding to the first $d$ largest eigenvalues of Eq. (17). Once $A$ is obtained, for any data $x$ in the input space, the nonlinear feature is given as $y = V^T \phi(x) = A^T [k(x_1, x), k(x_2, x), \ldots, k(x_N, x)]^T$.

### 3.3 The outline of EMKFDA

The algorithmic procedure of EMKFDA can be described in Table 1.

**Table 1** The algorithm of EMKFDA

**Step 1. Construct the kernel matrix**:

Compute the kernel matrix $K = \phi(X)^T \phi(X)$ whose elements are $K_{ij} = k(x_i, x_j) = [\phi(x_i) \cdot \phi(x_j)]$.

**Step 2. Compute the weight matrix:**

The optimal reconstruction weights can be computed by minimizing the reconstruction cost function: $J(W) = \sum_{i=1}^{N} \left\| \phi(x_i) - \sum_{j=1}^{K} W_{ij} \phi(x_j) \right\|^2$, where $\sum_{j=1}^{K} W_{ij} = 1$, and $W_{ij} \neq 0$ if $\phi(x_j)$ is one of the $K$ same class nearest neighbors of $\phi(x_i)$; otherwise, $W_{ij} = 0$.

**Step 3. Compute matrix** $M$ , $G$ , $L$ , $B$:

$M = (I - W)(I - W)^T$, $G = I - (1/N)ee^T$, $L = I - E$, $B = G - L$, where $I$ is an identity matrix, $e = [1, 1, \ldots, 1]^T$, $E_{ij} = 1/n_c$ if $x_i$ and $x_j$ both belong to the $c$ th class; otherwise $E_{ij} = 0$.

**Step 4. Compute the transformation matrix:**

Solve the generalized eigenvalue problem: $K(B - \mu \cdot M) K a_i = \lambda_i KLK a_i$, with $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$, and compute the matrix $A = [a_1, a_2, \ldots, a_d]$.

**Step 5. Extract EMKFDA features:**

For any data point $x$ in $R^D$, the embedded feature in $R^d$ is given by $y = V^T \phi(x) = A^T [k(x_1, x), k(x_2, x), \ldots, k(x_N, x)]^T$.

## 4 Experimental Results

To evaluate the efficiency of the EMKFDA algorithm, we compared the recognition rate of the proposed method with that of other methods, such as Eigenfaces (PCA) [2], Fisherfaces (LDA) [3], NPP [29], LPP [19], MFA [21], LSDA [22], KDA [9], and LLDE [28] on three well-known face databases: the ORL [24], Yale [25], and FERET [32] face databases. Note that, NPP and LPP were implemented in a supervised setting. In short, the recognition process had three steps. First, we calculated the face subspace from the training samples; then the new facial image to be identified was projected into low-dimensional subspace via a projection matrix; finally, the new facial image was identified using a nearest neighbor classifier with Euclidean distance. All of our experiments were performed on the Windows XP Platform with 2.00 GHz Intel Pentium Dual CPU and 2GB RAM, and MatLab (Version R2008a) programming environment.

### 4.1 Face Database

The ORL database contains 400 facial images of 40 individuals (each individual has ten images) with variations in pose, illumination, facial expression (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The images were taken with a tolerance for tilting and rotation of face up to 20°. Moreover, there was also some variation in the scale of up to 10 %. The Yale database contains 165 facial images of 15 individuals. There are 11 facial images per individual, and these images demonstrated variations in facial expression (happy, normal, sad, sleepy, surprised, and winking) and lighting condition (center-light, left-light, right-light). All the images in ORL database and Yale database were manually aligned, cropped, and resized to a size of 32 × 32 pixels with 256 gray levels per pixel. Figures 2 and 3 show the cropped and resized image samples of one individual in ORL database and Yale database respectively.

The FERET database contains 14,126 face images from 1199 individuals. In the experiments, we tested the proposed algorithm on a subset of the FERET database. This subset contains 1400 images of 200 individuals (each individual has seven images). The subset has variations in facial expression, illumination, and pose. Original images on the FERET database were normalized such that the two eyes were aligned at the same position. Then, the facial areas were cropped to 40 × 40 pixels for matching. Some example images of one person on FERET database are shown in Fig. 4.

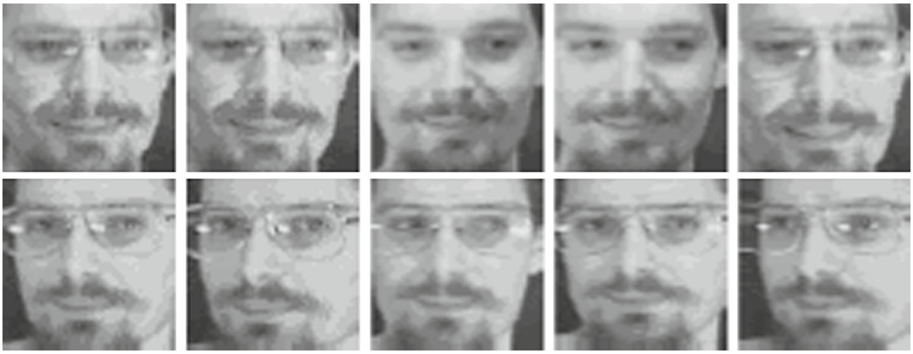

**Fig. 2** The cropped and resized image samples of one individual in the ORL database

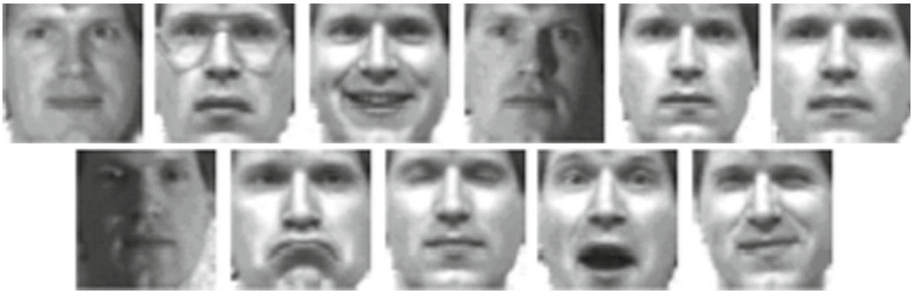

**Fig. 3** The cropped and resized image samples of one individual in the Yale database

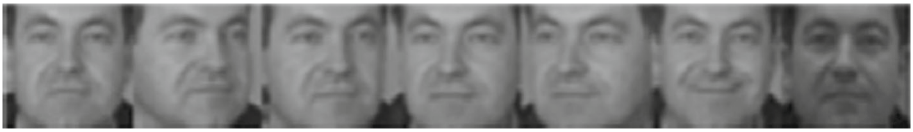

**Fig. 4** Images of one individual in the FERET database

### 4.2 Experiments on ORL Database

In the experiments, $l\,(=3,4,5,6)$ facial image samples were randomly selected from the image gallery of each individual to form the training sample set. The remaining images were used for testing. For each given $l$, the results were averaged over 50 random splits. Note that, for LDA, there are at most $c-1$ non-zero generalized eigenvalues, and so an upper bound on the dimension of the reduced space is $c-1$, where $c$ is the number of the classes. All the algorithms except PCA, KDA, and EMKFDA involved a PCA phase. In the PCA phase of LDA, NPP, LPP, MFA, LSDA and LLDE, we kept 100 % image energy and selected all principal components corresponding to the non-zero eigenvalues for each method. In general, the performance of all these methods varied with the number of dimensions. Firstly, we tested the impact of selecting different dimensions in the reduced subspace on the recognition rate. Figure 5 illustrates the recognition rates versus the variation of subspace dimensions when 3, 4, 5, and 6 images per individual were randomly selected for training. At the beginning, the recognition rates improved with the increase of the dimensions. However, more dimensions would not lead to higher recognition rate after these methods attained the best results. Secondly, the experiments were conducted to examine the effect of the training number on the performance. The maximal average recognition accuracy of each method and the corresponding standard deviation and the reduced dimension are given in Table 2 when the 3, 4, 5, and 6 samples per class were randomly selected for training and the remaining 7, 6, 5, and 4 images were respectively for testing. The proposed EMKFDA algorithm performed the best among all the cases. Moreover, the optimal dimensionality obtained by EMKFDA and LDA was much lower than that obtained by PCA.

At last, we tested the impact of parameter coefficient on the recognition rate. We randomly chose 5 training samples of each subject from the ORL database to form the training set and the rest are the testing set. The coefficient, i.e. $\mu$, was set to 0.01, 0.1, 1, 10, and 100, respectively. The maximal average recognition rates for different coefficients are stated in the Table 3. The optimal recognition rates can be obtained with different coefficients and the corresponding dimensions, for example, when coefficient was 0.01, the recognition was

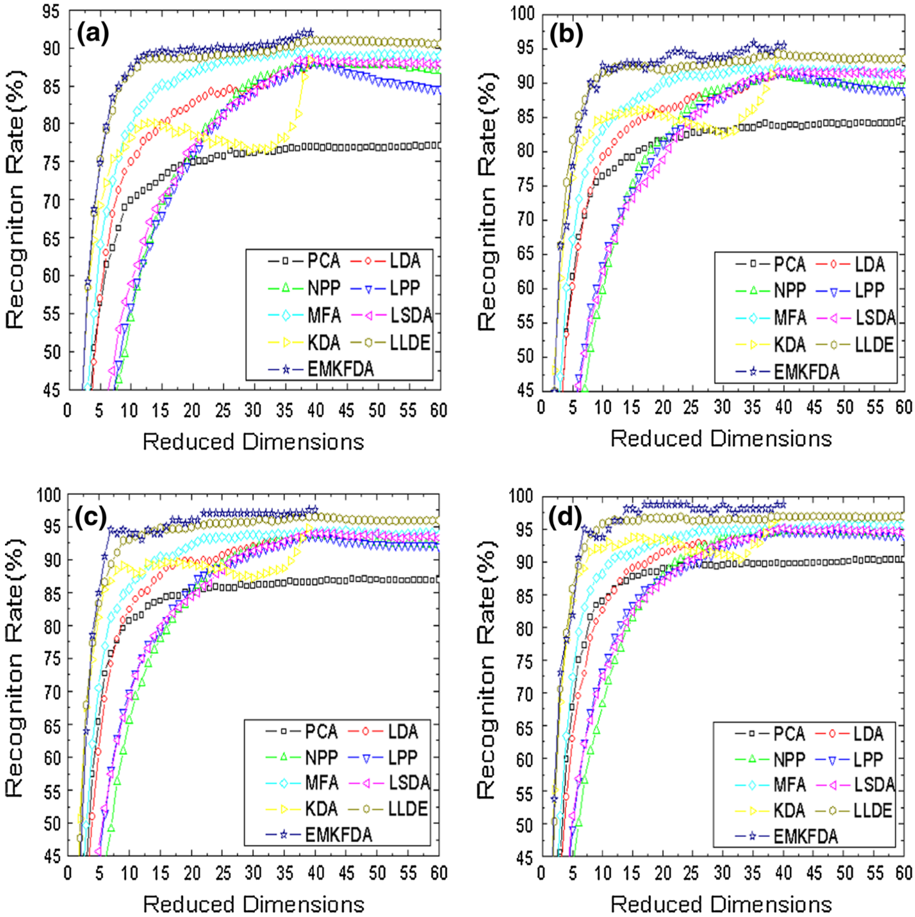

**Fig. 5** Recognition rate versus dimensionality reduction on ORL face database. **a** Three samples for training. **b** Four samples for training. **c** Five samples for training. **d** Six samples for training

97.50 % at 39 dimensions. However, the recognition rate reached 97.50 % at 39 dimensions with $\mu$ equaling to 100. It can be found that the parameter coefficient shows few effects on the recognition rate on ORL face database.

### 4.3 Experiments on Yale Database

The experimental design was the same as before. For each individual, $l$ (= 3, 4, 5, 6) facial image samples were randomly selected for training and the rest were used for testing. For each given $l$, we averaged the results over 50 random splits. LDA, NPP, LPP, MFA, LSDA and LLDE involved a preceding PCA stage to avoid the singularity problem and 100 % image energy was kept in PCA stage. In this experiment, we also tested the impact of selecting different dimensions in the reduced subspace on the recognition rate. Figure 6 shows the average recognition rates (%) of PCA, LDA, NPP, LPP, MFA, LSDA, KDA, LLDE and the proposed method versus the dimensions when the 3, 4, 5, and 6 samples per class were randomly selected for training and the remaining 7, 6, 5, and 4 images were respectively for testing. Figure 6 shows that the discrimination power of these methods will be enhanced with

**Table 2** The maximal average recognition rate (%) and the corresponding standard deviations with the reduced dimensions on ORL database

| Method | 3Train | 4Train | 5Train | 6Train |
|---|---|---|---|---|
| PCA | 78.42 ± 2.89 (119) | 84.96 ± 1.98 (159) | 87.80 ± 1.57 (199) | 90.63 ± 2.04 (239) |
| LDA | 87.60 ± 3.08 (39) | 91.29 ± 2.56 (39) | 94.00 ± 1.90 (39) | 94.63 ± 2.15 (39) |
| NPP | 88.00 ± 2.41 (39) | 91.50 ± 2.21 (39) | 93.65 ± 2.26 (39) | 94.69 ± 2.17 (42) |
| LPP | 87.92 ± 2.47 (39) | 91.54 ± 2.14 (40) | 93.50 ± 1.63 (39) | 94.63 ± 1.91 (39) |
| MFA | 89.50 ± 2.38 (35) | 92.08 ± 1.88 (39) | 94.30 ± 1.60 (42) | 95.13± 1.71 (40) |
| LSDA | 88.50 ± 2.13 (39) | 91.58 ± 2.31 (39) | 94.00 ± 1.96 (40) | 95.06 ± 1.16 (40) |
| KDA | 88.43 ± 2.61 (39) | 93.50 ± 1.70 (39) | 94.70 ± 1.93 (39) | 96.44 ± 1.10 (39) |
| LLDE | 90.96 ± 2.06 (39) | 94.25 ± 2.11 (39) | 96.50 ± 1.58 (39) | 96.93 ± 1.23 (39) |
| EMKFDA | 91.92 ± 1.17 (38) | 95.41 ± 2.03 (39) | 97.50 ± 1.66 (39) | 98.75 ± 1.18 (39) |

**Table 3** Dimensions versus recognition rate by varying parameter coefficient in the proposed algorithm

| | | | | | |
|---|---|---|---|---|---|
| Parameter coefficient | 0.01 | 0.1 | 1 | 10 | 100 |
| The best average recognition rate (%) | 97.50 | 97.50 | 97.50 | 97.50 | 97.50 |
| Dimensions | 39 | 38 | 39 | 39 | 39 |

the increase of final projected dimension, but they will not increase all the time. Moreover, the effect of the training sample number was also tested in the experiment. Table 4 shows the maximal average recognition rates and the corresponding standard deviations with dimensions after carrying out PCA, LDA, NPP, LPP, MFA, LSDA, KDA, LLDE and the proposed method. As with the ORL database, the proposed algorithm also outperformed all other methods with the Yale database while PCA method performed the worst among all the cases.

### 4.4 Experiments on FERET Database

In the FERET experiment, we investigated the performances of different algorithms under different numbers of training samples. $l(= 3, 5)$ images of each class were randomly selected to form the training images and the remaining images for testing. We randomly chose the training set and performed the experiments 50 times. The final result was the average recognition rate over 50 random training sets. For subspace learning, we used PCA, LDA, NPP, LPP, MFA, LSDA, KDA, LLDE, and the proposed method, respectively. Note that LDA, NPP, LPP, MFA, LSDA, LLDE all involved a PCA phase. In this phase, we kept 100 % image energy. Figure 7 demonstrates the recognition rates of different algorithms over the variance of the dimensionality of subspaces. In addition, we also investigated the performances of different algorithms over different sizes of the training dataset. The highest recognition accuracies and the corresponding standard deviations

with dimensions of different algorithms on FERET database are reported on Table 5. The proposed algorithm performed superior to the other methods. The recognition accuracy improved from 88.3 % with three train images of each individual to 93.85 % with five train of each person.

### 4.5 Discussion

Several experiments on three standard face databases were conducted. The following observations should be noted:

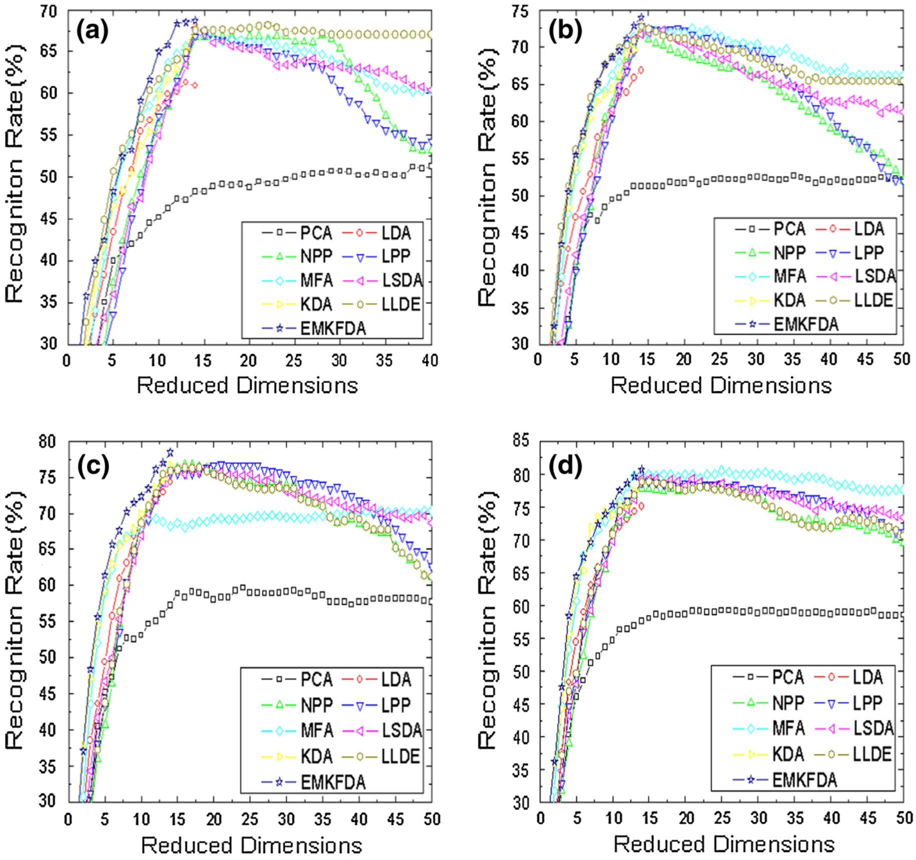

**Fig. 6** Recognition rate versus dimensionality reduction on Yale face database. **a** Three samples for training. **b** Four samples for training. **c** Five samples for training. **d** Six samples for training

**Table 4** The maximal average recognition rate (%) and the corresponding standard deviations with the reduced dimensions on Yale database

| Method | 3Train | 4Train | 5Train | 6Train |
|---|---|---|---|---|
| PCA | 52.00 ± 3.22 (44) | 53.05 ± 2.25 (59) | 58.89 ± 3.63 (74) | 59.73 ± 2.65 (89) |
| LDA | 60.93 ± 4.33 (14) | 66.86 ± 3.80 (14) | 74.33 ± 4.00 (14) | 75.07 ± 4.57 (14) |
| NPP | 66.67 ± 3.73 (14) | 71.71 ± 3.54 (14) | 76.67 ± 4.68 (14) | 78.13 ± 3.34 (22) |
| LPP | 67.00 ± 3.83 (15) | 72.76 ± 2.88 (14) | 75.56 ± 3.62 (14) | 78.80 ± 4.13 (15) |
| MFA | 67.83 ± 5.07 (14) | 72.47 ± 4.45 (13) | 69.93 ± 5.42 (12) | 80.00 ± 4.87 (15) |
| LSDA | 68.00 ± 3.02 (14) | 72.28 ± 3.43 (17) | 76.00 ± 4.03 (15) | 79.33 ± 3.84 (14) |
| KDA | 67.91 ± 2.97 (14) | 72.86 ± 3.65 (14) | 76.67 ± 3.95 (14) | 78.67 ± 4.12 (14) |
| LLDE | 68.16 ± 4.64 (22) | 72.38 ± 3.56 (15) | 76.33 ± 3.39 (16) | 78.80 ± 4.19 (16) |
| EMKFDA | 68.66 ± 3.96 (14) | 73.95 ± 3.94 (14) | 78.44 ± 3.99 (14) | 81.12 ± 2.19 (14) |

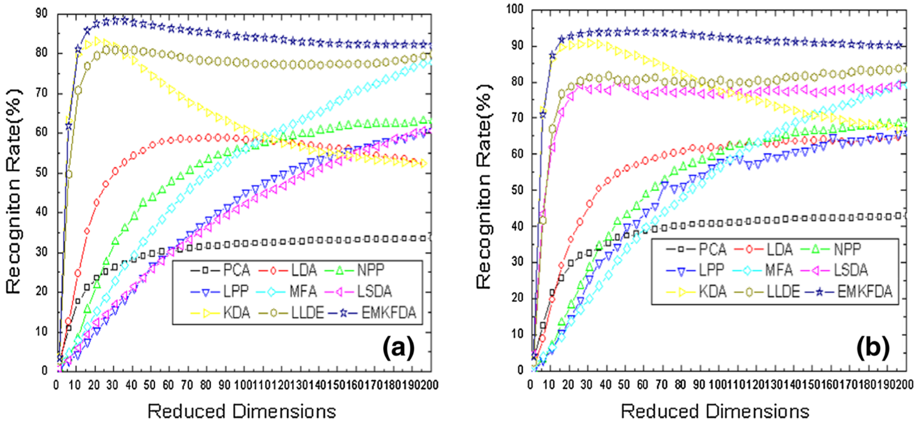

**Fig. 7** Recognition rate versus dimensionality reduction on FERET face database. **a** Three samples for training. **b** Five samples for training

**Table 5** The maximal average recognition rate (%) and the corresponding standard deviations with the reduced dimensions on FERET database

| Method | 3Train | 5Train |
|---|---|---|
| PCA | 34.16 ± 1.28 (566) | 43.57 ± 1.71 (681) |
| LDA | 58.81 ± 2.48 (81) | 64.57 ± 1.64 (199) |
| NPP | 63.25 ± 1.87 (201) | 68.57 ± 2.65 (199) |
| LPP | 60.57 ± 1.37 (201) | 65.75 ± 2.29 (199) |
| MFA | 77.81 ± 1.37 (206) | 79.12 ± 1.88 (199) |
| LSDA | 60.57 ± 1.15 (196) | 79.25 ± 1.54 (199) |
| KDA | 83.05 ± 1.40 (21) | 90.82 ± 1.56 (31) |
| LLDE | 80.90 ± 1.34 (26) | 83.75 ± 1.47 (199) |
| EMKFDA | 88.30 ± 1.10 (31) | 93.85 ± 1.02 (51) |

(1) The proposed EMKFDA algorithm consistently performed the best in all the experimental cases. The data sets used in this study were the ORL, Yale, and FERET face databases. The images for each individual varied from pose, illumination to expression. Some studies have demonstrated facial images likely reside on a low-dimensional submanifold. Compared to PCA and LDA which see only the global Euclidean structure of face space, EMKFDA explicitly considers the face manifold structure which is modeled by a neighborhood graph. Moreover, EMKFDA utilizes the label information and nonlinear structure information to enhance the classification performance.

(2) In contrast with NPP and LPP, which only preserve the local neighborhood manifold information, EMKFDA preserves both local geometrical information and global discriminant structure information, so EMKFDA has stronger discriminant power.

(3) MFA, LSDA, and LLDE capture both the local geometry information and the discriminant information of the data, but they are linear manifold learning methods. EMKFDA is a local information and global discriminating information preserving nonlinear dimensionality reduction method.

(4) KDA is a nonlinear dimensionality reduction method; however, it does not possess local geometry preserving property. The EMKFDA feature set created using the combined

approach retains the local geometry information as well as the global discriminating information, which makes EMKFDA obtain better recognition rate.

## 5 Conclusions

In this paper, based on manifold criterion and Fisher criterion, we have proposed EMKFDA algorithm, which is a new dimensionality reduction method. EMKFDA preserves not only the local geometry structure of the data, but also the global discriminant structure of the data. The objective function of EMKFDA is implemented by mapping the input data into a high-dimensional feature space using a kernel matrix, so nonlinear feature can be extracted efficiently. Thus, EMKFDA has better capability of discrimination. Experimental results on ORL, Yale, and FERET face database have demonstrated the effectiveness of the proposed algorithm. In the future work, we will further research the proposed method in the tensor space to consider the image matrix structure information.

**Acknowledgments** This work is supported by the Young Core Instructor of Colleges and Universities in Henan Province Funding Scheme (No. 2011GGJS-173), the Science and Technology Project of Henan Province (No. 122102210138), Foundation of Henan Educational Committee (No. 14A520055), the Nation Scholarship Fund (No. 201308410352).

## References

1. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: a literature survey. ACM Comput Surv 35(4):399–458
2. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
3. Belhumeour PN, Hedpsnhs JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
4. Howland P, Wang JL, Park H (2006) Solving the small sample size problem in face recognition using generalized discriminant analysis. Pattern Recognit 39(2):277–287
5. Liang YX, Li CR, Gong WG, Pan YJ (2007) Uncorrelated linear discriminant analysis based on weighted pairwise Fisher criterion. Pattern Recognit 40(12):3606–3625
6. Zhao W, Zhao L, Zou C (2004) An efficient algorithm to solve the small sample size problem for LDA. Pattern Recognit 37(5):1077–1079
7. Ye J, Li Q (2005) A two-stage linear discriminant analysis via QR-decomposition. IEEE Trans Pattern Anal Mach Intell 27(6):929–941
8. Scholkopf B, Smola A, Muller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319
9. Baudat G, Anouar F (2000) Generalized discriminant analysis using a kernel approach. Neural Comput 12(10):2385–2404
10. Tenebaum J, Silva V, Langford J (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
11. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
12. Belkin M, Niyogi P (2001) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proceedings of neural information processing systems, Vancouver, pp 585–591
13. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396
14. Weinberger K, Saul L (2004) Unsupervised learning of image manifolds by semidefinite programming. In: Proceedings of the IEEE international conference computer vision and pattern recognition, vol 2, pp 988–985
15. Zhang Z, Zha H (2005) Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. SIAM J Sci Comput 26(1):313–318
16. Zhang TH, Li XL, Tao DC, Yang J (2008) Local coordinates alignment (LCA): a novel method for manifold learning. Int J Pattern Recognit Artif Intell 22(4):667–690

17. Xiang SM, Nie FP, Xiang SM, Zhuang YT, Wang WH (2009) Nonlinear dimensionality reduction with local spline embedding. IEEE Trans Knowl Data Eng 21(9):1285–1298
18. He X, Cai D, Yan S, Zhang H (2005) Neighborhood preserving embedding. In: Proceedings of the IEEE international conference computer vision, pp 1208–1213
19. He X, Yan S, Hu Y, Niyogi P, Zhang HJ (2005) Face recognition using Laplacianfaces. IEEE Trans Pattern Anal Mach Intell 27(3):328–340
20. Chen HT, Chang HW, Liu TL (2005) Local discriminant embedding and its variants. In: Proceedings of the conference on computer vision and pattern recognition, vol 2, pp 846–853
21. Yan SC, Xu D, Zhang BY, Zhang HJ (2005) Graph embedding: a general framework for dimensionality reduction. In: Proceedings of the conference on computer vision and pattern recognition, vol 2, pp 20–25
22. Cai D, He XF, Zhou K (2007) Locality sensitive discriminant analysis. In: Proceedings of the conference on artificial intelligence, pp 708–713
23. Vapnik V (1995) The nature of statistical learning theory. Springer, New York
24. The ORL face database. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. Accessed 2004
25. The Yale face database. http://cvc.yale.edu/projects/yalefaces/yalefaces.html. Accessed 2004
26. Seung HS, Lee DD (2000) The manifold ways of perception. Science 290:2258–2259
27. Zhang J, Li SZ, Wang J (2004) Manifold learning and applications in recognition. In: Intelligent multimedia processing with soft computing, vol 168, pp 281–300
28. Li B, Huang DS (2008) Locally linear discriminant embedding: an efficient method for face recognition. Pattern Recognit 41(12):3813–3821
29. Pang YW, Yu NH, Li HQ et al (2005) Face recognition using neighborhood preserving projections. In: Proceedings of pacific-rim conference on multimedia, vol 3768, pp 854–864
30. Li H, Jiang T, Zhang K (2006) Efficient and robust feature extraction by maximum margin criterion. IEEE Trans Neural Netw 17(1):157–165
31. Saul LK, Roweis ST (2003) Think globally, fit locally: unsupervised learning of low dimensional manifolds. J Mach Learn Res 4:119–155
32. The facial recognition technology (FERET) database. http://www.itl.nist.gov/iad/humanid/feret/feret_master.html. Accessed 2008
33. Zhang TH, Tao DC, Li XL, Yang J (2008) A unifying framework for spectral analysis based dimensionality reduction. In: Proceedings of the international joint conference on neural networks, pp 1670–1677