Neural Comput & Applic (2011) 20:565–573
DOI 10.1007/s00521-011-0577-7

ORIGINAL ARTICLE

Locality preserving embedding for face and handwriting digital recognition

Zhihui Lai · MingHua Wan · Zhong Jin

Received: 3 December 2008 / Accepted: 11 March 2011 / Published online: 1 April 2011

Abstract Most supervised manifold learning-based methods preserve the original neighbor relationships to pursue the discriminating power. Thus, structure information of the data distributions might be neglected and destroyed in low-dimensional space in a certain sense. In this paper, a novel supervised method, called locality preserving embedding (LPE), is proposed to feature extraction and dimensionality reduction. LPE can give a low-dimensional embedding for discriminative multi-class sub-manifolds and preserves principal structure information of the local sub-manifolds. In LPE framework, supervised and unsupervised ideas are combined together to learn the optimal discriminant projections. On the one hand, the class information is taken into account to characterize the compactness of local sub-manifolds and the separability of different sub-manifolds. On the other hand, at the same time, all the samples in the local neighborhood are used to characterize the original data distributions and preserve the structure in low-dimensional subspace. The most significant difference from existing methods is that LPE takes the distribution directions of local neighbor data into account and preserves them in low-dimensional subspace instead of only preserving the each local sub-manifold's original neighbor relationships. Therefore, LPE optimally preserves both the local sub-manifold's original neighborhood relationships and the distribution direction of local neighbor data to separate different sub-manifolds as far as possible. The criterion, similar to the classical Fisher criterion, is a Rayleigh quotient in form, and the optimal linear projections are obtained by solving a generalized Eigen equation. Furthermore, the framework can be directly used in semi-supervised learning, and the semi-supervised LPE and semi-supervised kernel LPE are given. The proposed LPE is applied to face recognition (on the ORL and Yale face databases) and handwriting digital recognition (on the USPS database). The experimental results show that LPE consistently outperforms classical linear methods, e.g., principal component analysis and linear discriminant analysis, and the recent manifold learning-based methods, e.g., marginal Fisher analysis and constrained maximum variance mapping.

Keywords Manifold learning · Face recognition · Marginal Fisher analysis (MFA) · Linear discriminant analysis (LDA) · Principal component analysis (PCA) · Constrained maximum variance mapping (CMVM)

Z. Lai (✉) · M. Wan · Z. Jin
School of Computer Science,
Nanjing University of Science and Technology,
210094 Nanjing, Jiangsu, People's Republic of China
e-mail: lai_zhi_hui@163.com

1 Introduction

In pattern recognition and machine learning fields, many applications, such as appearance-based image recognition and document categorization, face the high-dimensional problem. Finding a low-dimensional representation of the high-dimensional data is a basic task. Using the reduced features, classification can be much faster and more robust [1–3]. Thus, some dimensionality reduction approaches have been developed [1, 4–8]. For unsupervised methods, e.g., principal component analysis (PCA) [4, 6] and locality preserving projection (LPP) [7], the low-dimensional representation should discover the structure information of the data point cloud [9]. For supervised classification problem, e.g., linear discriminant analysis (LDA) [4, 6, 10, 11] and maximum margin criterion (MMC) [12], the reduced

low-dimensional features should contain most discriminative information based on the labeled data.

Recently, graph-based algorithms became an active topic. Yan et al. [13] proposed a new general framework called graph embedding for dimensionality reduction from which many algorithms, such as PCA, LDA, LPP, ISOMAP [14], LLE [15], and Laplacian eigenmap [16], can all be reformulated. Using the graph embedding framework as a platform, they developed a novel dimensionality reduction algorithm called marginal Fisher analysis (MFA) to overcome the limitation of LDA. The powerful strength of their algorithm comes from the intrinsic graph and the penalty graph, defined in local neighborhood. Each manifold learning algorithm, such as LPP, ISOMAP, LLE, attempts to preserve different geometrical properties of the underlying manifold. It was shown that these methods have yielded impressive results on artificial and real-world data sets, especially for data visualization. However, the original manifold learning-based techniques might be unsuitable for classification because of only concerning with the training data. Furthermore, it is known that the samples of different classes lie in different sub-manifolds and the samples in the same class might lie in two or more sub-manifolds in different areas. Therefore, it is necessary for us to distinguish individual images from different sub-manifolds. In order to achieve a better recognition result, the recovered embeddings corresponding to different sub-manifolds should be as separable as possible in the final embedding space. This proposes a problem called classification-oriented multi-manifolds learning. This problem cannot be solved by current manifold learning algorithms, including some supervised versions [17–21], because they are all only based on the characterization of local scatter for unsupervised learning or local class separation information, i.e., local interclass scatter. As is shown in Fig. 1, current manifold learning methods seem to find the projections approximating to PCA or LDA (the projection of the proposed LPE seems more effective than the other methods since different kinds of properties are taken into consideration and the four classes data are almost not overlapping on the LPE direction).

In order to solve the problem of classification-oriented multi-manifolds learning, Yang et al. [22] proposed an unsupervised discriminant projection (UDP) algorithm considering the nonlocal scatter and local scatter at the same time. As an unsupervised learning algorithm, UDP did not take the label information into account and could be viewed as a simplified or regularized version of LPP [23]. Since there are different sub-manifolds that might be very close to each other, it is difficult for unsupervised learning algorithm to separate this kind of different sub-manifolds. Therefore, PCA, LPP, and UDP might be problematic for solving the problem of classification-oriented multi-manifolds learning. Recently, many supervised learning algorithms were developed to learn the embedding subspace in the problem of classification-oriented multi-manifolds learning, including local discriminant embedding (LDE) [24], maximum variance projection (MVP) [25], marginal Fisher analysis (MFA) [13], and constrained maximum variance mapping (CMVM) [26]. These kinds of supervised algorithms only take the locally discriminant information (we call them locally discriminant direction in this paper) into account, and the original structure of local data distribution in the final embedding subspace may be disregarded or destroyed. That is to say, these kinds of algorithms do not simultaneously take the local principle direction into account and the faithful distribution of the data in final space may be disregarded or destroyed. Thus, it is difficult for these methods to achieve higher classification accuracy.

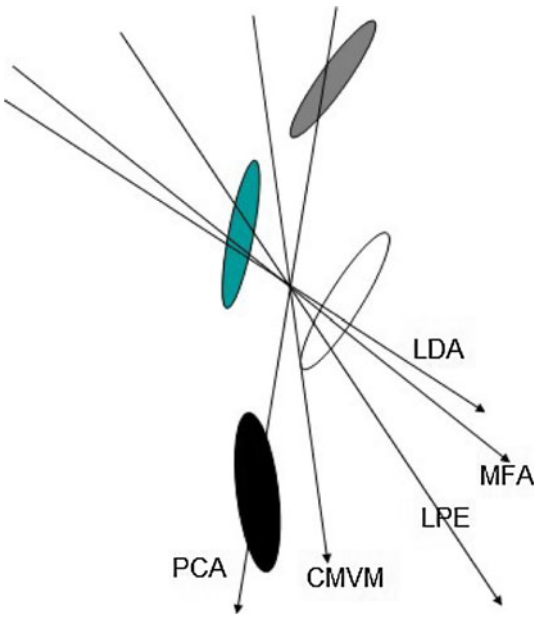

Fig. 1 Illustration of multi-classes of samples in two-dimensional space and the projection directions of PCA, LDA, MFA, CMVM, and proposed LPE

In this paper, motivated by the manifold learning-based methods mentioned previously, a novel supervised method, called locality preserving embedding (LPE), is proposed to feature extraction and dimensionality reduction. LPE focuses on solving the problem of classification-oriented multi-manifolds learning. Some parts of the LPE algorithm were presented earlier by us in [27] as a conference paper, and more details and its extensions are shown in this paper. The characteristics of LPE can be summarized as follows: Firstly, we adopted the strategy of linear approximation method to avoid the out-of-sample extension problem. Secondly, local discriminant directions and locally principal directions of the data set are analyzed and combined to

find the optimal subspace. Thirdly, in order to preserve the locality of each sub-manifold, a constrained condition is appended to the objective function to separate different sub-manifolds as far as possible. In LPE, the criterion, which is similar to the classical Fisher criterion, is a Rayleigh quotient in form, and the optimal linear projections are obtained by solving a generalized Eigen equation. Furthermore, the framework is generalized to semi-supervised learning, and semi-supervised LPE and semi-supervised kernel LPE are also shown in this paper.

The rest of this paper is organized as follows: in Sect. 2, the LPE algorithm is described and the extensions of LPE are also given. In Sect. 3 the proposed algorithm is examined on some databases. The conclusions are given in Sect. 4.

2 Locality preserving embedding

For a classification problem, the training sample is represented as a matrix $X = [x_1, x_2, \ldots, x_N]$, where $x_i \in R^m (i = 1, 2, \ldots, N)$, N is the sample number and m is the feature dimension. For supervised learning problems, the class label of the sample x_i is assumed to be $c_i \in \{1, 2, \ldots, N_c\}$, where N_c denotes the number of class. Let n_c denote the number of the samples belonging to the cth class. In practice, the feature dimension m is often very high. The goal of the proposed algorithm is to transform the data from the original high-dimensional space to a low-dimensional one, i.e., $Y \in R^{d\times N}$ with $d \ll m$. That is to say, we want to find an optimal linear transform $P = (\omega_1, \omega_2, \ldots, \omega_d)$, such that $Y = P^{\mathrm{T}}X$, where ω_i $(i = 1, \ldots, d)$is an m-dimension column vector.

2.1 Locally discriminant direction

Focusing on manifold learning and pattern classification, our embedding method attempts to achieve good discriminating performance by integrating the information of neighbor and class relations between data points. Since there are many sub-manifolds in the high-dimensional space, how to distinguish one sub-manifold from the others will heavily depend on the sub-manifold labels. As mentioned earlier, the data distributed on a sub-manifold are belonging to the same class. So, we can construct a label matrix H to mark the label information of each point, where H is defined as follows:

$$H_{ij}^I = \begin{cases} 1, & \text{if } i \in N_K^-(j) \text{ or } j \in N_K^-(i) \\ 0, & \text{else} \end{cases} \tag{1}$$

where $N^-(i)$ indicates the index in the K nearest neighbors of the sample x_i but with different class labels. We call H as local interclass graph in this paper. Then the local interclass scatter can be defined as the following equation:

$$\begin{aligned} S_I &= \sum_{ij} H_{ij}^I \left\|y_i - y_j\right\|^2 = \sum_{ij} H_{ij}^I \left\|\varphi^{\mathrm{T}}x_i - \varphi^{\mathrm{T}}x_j\right\|^2 \\ &= 2\varphi^{\mathrm{T}}X(D^I - H^I)X^{\mathrm{T}}\varphi = 2\varphi^{\mathrm{T}}XL^IX^{\mathrm{T}}\varphi \end{aligned} \tag{2}$$

where $L^I = D^I - H^I$, $D_{ii}^I = \sum H_{ij}^I$.

Let $\varphi^* = \arg\max_{\varphi} S_I$. Then, φ^* is called locally discriminant direction of the sub-manifolds, which characterizes the distribution properties of the data set in different classes, i.e., in different sub-manifolds. Similar to PCA, the optimal projection vector, which is corresponding to the maximal eigenvalue, can be obtained from solving the Eigen equation:

$$XL^IX^{\mathrm{T}}\varphi = \lambda\varphi. \tag{3}$$

In fact, S_{I} characterizes the local dissimilarities between sub-manifolds. The optimal projection φ^* indicates along which direction the sub-manifold can be separated as far as possible, see Fig. 2a.

2.2 Local compactness direction of sub-manifold

Since the data set in the same class is supposed to be down-sampled from a manifold and might form one or more sub-manifolds, we should characterize the compactness of the down-sampled data set for effectiveness to separate the different sub-manifolds. Intraclass compactness is characterized by the term

$$\begin{aligned} S_C &= \sum_{ij} W_{ij}^C \left\|\phi^{\mathrm{T}}x_i - \phi^{\mathrm{T}}x_j\right\|^2 = 2\phi^{\mathrm{T}}X(D^C - W^C)X^{\mathrm{T}}\phi \\ &= 2\phi^{\mathrm{T}}XL^CX^{\mathrm{T}}\phi \end{aligned} \tag{4}$$

where

$$L^C = D^C - W^C, \quad D_{ii}^C = \sum\nolimits_j W_{ij}^C,$$

and

$$W_{ij}^C = \begin{cases} 1, & i \in N_K^+(j) \text{ or } j \in N_K^+(i), \\ 0, & \text{else.} \end{cases}$$

where $N_K^+(i)$ indicates the index set of the K nearest neighbors of the sample x_i in the same class. We call S_C as local intraclass graph in this paper. From the form of S_C, it is easy to see that it exactly characterizes the sum scatter of the samples in K-neighborhood in the same class.

Let $\phi^* = \arg\min_{\phi} S_C$. Then, ϕ^* is called local compactness direction of the sub-manifolds, which characterizes the distribution properties of the data set in the same classes, i.e., in the same sub-manifold. Similar to LPP, the optimal projection vector corresponding to the minimal eigenvalue can be obtained by solving the Eigen equation:

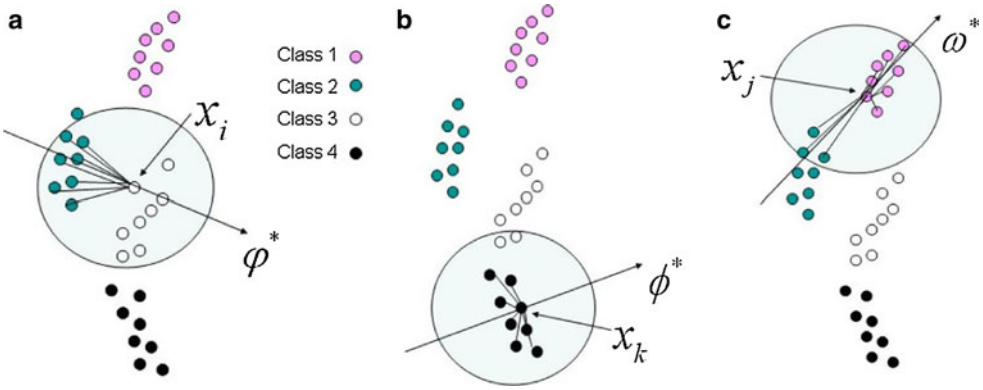

Fig. 2 Three graphs and its optimal projections used in this paper. The circle denotes the local neighborhood of x_i, x_k, x_j, respectively. **a** Local interclass graph and corresponding locally discriminant direction of the sub-manifolds. **b** Local intraclass graph and local compactness direction of the sub-manifolds. **c** Local adjacency graph and locally principal direction of the sub-manifolds

$$XL^C X^{\mathrm{T}}\phi = \lambda\phi. \tag{5}$$

In fact, S_C characterizes the local similarities on the sub-manifolds. The optimal projection φ^* indicates that along which direction the sub-manifold of each class can be compacted as close as possible, see Fig. 2b.

2.3 Locally principal directions

Principal component analysis seeks to find a projection axis such that the global scatter is maximized when the samples are projected to the low-dimensional subspace. We define the local scatter S_L to characterize the local data structure.

$$\begin{aligned} S_L &= \sum_{ij} F_{ij}\left\|y_i - y_j\right\|^2 = \sum_{ij} F_{ij}\left\|\omega^{\mathrm{T}}x_i - \omega^{\mathrm{T}}x_j\right\|^2 \\ &= 2\omega^{\mathrm{T}}X(D-F)X^{\mathrm{T}}\omega = 2\omega^{\mathrm{T}}XLX^{\mathrm{T}}\omega \end{aligned} \tag{6}$$

where

$$L = D - F, \quad D_{ii} = \sum\nolimits_j F_{ij},$$

and

$$F_{ij} = \begin{cases} 1, & \text{if } i \in N(j) \text{ or } j \in N(i) \\ 0, & \text{else.} \end{cases},$$

$N(i)$ indicate the index set of the K nearest neighbors of the sample x_i. We call S_L as local adjacency graph in this paper.

Let $\omega^* = \arg\max_{\omega} S_L$. Then, ω^* is called locally principal direction of the sub-manifolds, which characterizes the distribution properties of the data set. Similar to PCA, the optimal projection vector, which is corresponding to the maximal eigenvalue, can be obtained from solving the Eigen equation:

$$XLX^{\mathrm{T}}\omega = \lambda\omega. \tag{7}$$

It is clear that ω^* characterizes the principle direction of the local data. The optimal projection ω^* indicates that along which direction the sub-manifold of each local neighborhood can be preserved as more as possible, see Fig. 2c.

2.4 The motivation and justification of LPE

Based on the assumption that data belonging to the same class are resided on one or more sub-manifold and data with different classes are distributed on different sub-manifolds, LPE algorithm is proposed to separate sub-manifolds farther and preserve the local structures of the data set at the same time. In other words, the LPE algorithm can obtain the optimal discriminant features without destroying the local geometry of each sub-manifold. We want to find the optimal projections that can preserve the local geometry structure but have more discriminant abilities. This means that the optimal projections can separate different sub-manifolds as farther as possible without destroying localities. The idea is to approximate the local principal directions and the local discriminant directions under the constraint of preserving the local compactness directions. In LPE framework, supervised and unsupervised idea are combined together to learn the optimal discriminant directions. On the one hand, the class information is taken into account to characterize the compactness of local sub-manifolds and the separability of different sub-manifolds. On the other hand, at the same time, all the samples in the local neighborhood are used to characterize and preserve the natural data distribution structure in low-dimensional subspace, which have the unsupervised learning property. That is to say, locally unsupervised learning is used to enhance or help the locally supervised learning for feature extraction and dimensionality reduction. Also, this framework can be viewed as a new graph embedding algorithm since a local graph is used in each term of the optimization problem. This is the reason why the proposed method is named as locality preserving embedding (LPE).

With these above aspects of consideration, we want to find the suitable (or common) projection V that can take

these properties into consideration, thus we arrive at the following constrained optimization problem:

$$\text{Maximze } J(V) = V^{\mathrm{T}}XLX^{\mathrm{T}}V + V^{\mathrm{T}}XL^{\mathrm{I}}X^{\mathrm{T}}V \tag{8}$$

$$\text{Subject to } V^{\mathrm{T}}XL^{C}X^{\mathrm{T}}V = 1 \tag{9}$$

This constraint optimization problem can be figured out by enforcing Lagrange multiplier. Firstly, a function $J'(V)$ can be constructed by the objective function and the constraint:

$$J'(V) = V^{\mathrm{T}}XLX^{\mathrm{T}}V + V^{\mathrm{T}}XL^{\mathrm{I}}X^{\mathrm{T}}V - \lambda(V^{\mathrm{T}}XL^{C}X^{\mathrm{T}}V - 1) \tag{10}$$

Secondly, the optimal transformation matrix V can be obtained from

$$\frac{\partial J'(V)}{\partial V} = 2XLX^{\mathrm{T}}V + 2XL^{I}X^{\mathrm{T}}V - \lambda 2XL^{C}X^{\mathrm{T}}V = 0. \tag{11}$$

Then, we have

$$[XLX^{\mathrm{T}} + XL^{\mathrm{I}}X^{\mathrm{T}}]V = \lambda XL^{C}X^{\mathrm{T}}V. \tag{12}$$

From (12), it can be found that the optimal projections are composed of the eigenvectors associated with the d top eigenvalues by solving the corresponding generalized Eigen equation.

It should be noted that the matrix $XL^{\mathrm{C}}X^{\mathrm{T}}$ might be singular, which stems from the SSS problem. In order to overcome the complication of a singular matrix $XL^{\mathrm{C}}X^{\mathrm{T}}$, we first project the data set to a PCA subspace so that the resulting matrix $XL^{\mathrm{C}}X^{\mathrm{T}}$ is nonsingular. Another consideration of using PCA as preprocessing is for noise reduction. The preprocessing must be performed when encountering the case mentioned earlier. Therefore, the final transformation matrix V can be expressed as follows:

$$V = V_{\mathrm{PCA}}V_{\mathrm{LPE}} \tag{13}$$

where V_{LPE} denotes the transformation matrix obtained from (12), and V_{PCA} denotes the transformation matrix of PCA.

The LPE algorithmic procedures can be summarized as follows:

Step 1: Project the original data into the PCA subspace to overcome the SSS problem by throwing away the smallest principal components.
Step 2: Compute the distance matrix between any two data points.
Step 3: Compute the matrix XLX^{T} by using the K nearest neighbors.
Step 4: Compute the matrices $XL^{\mathrm{I}}X^{\mathrm{T}}$ and $XL^{\mathrm{C}}X^{\mathrm{T}}$ by using the K nearest neighbors and labels.
Step 5: Compute the optimized resolutions by solving the generalized eigenvalue problem based on (12).
Step 6: Compute the final transformation matrix based on (13).
Step 7: Project the data points to the low-dimensional subspace and adopt a suitable classifier for classification.

2.5 Extensions of LPE

2.5.1 Kernel LPE

The classification power of linear projection algorithms may be limited and insufficient to deal with complicated problems. One possible attempt to elevate the classification performance is to transform input data to a higher-dimensional space via a nonlinear mapping. The kernel trick is widely used to enhance the separation ability of a linear dimensionality reduction algorithm. We investigate the kernel representation and establish a new algorithm (named kernel LPE) incorporating nonlinearity to separate the nonlinear multi-sub-manifolds.

Suppose that the original data X is mapped into a high-dimensional feature space by a nonlinear mapping Φ. Then, we have a new data set $\Phi(X)$. To simplify the discussion, we can obtain the following optimization problem with simple calculations in the kernel space:

$$\text{Maximze } J_{\Phi}(V) = V^{\mathrm{T}}\Phi(X)L\Phi^{\mathrm{T}}(X)V + V^{\mathrm{T}}\Phi(X)L^{I}\Phi^{\mathrm{T}}(X)V \tag{14}$$

$$\text{Subject to } V^{\mathrm{T}}\Phi(X)L^{C}\Phi^{\mathrm{T}}(X)V = 1. \tag{15}$$

Thus, the optimal problem is changed into the corresponding generalized Eigen equation:

$$[\Phi(X)L\Phi^{\mathrm{T}}(X) + \Phi(X)L^{I}\Phi^{\mathrm{T}}(X)]V = \lambda\Phi(X)L^{C}\Phi^{\mathrm{T}}(X)V. \tag{16}$$

2.5.2 Extensions of semi-supervised LPE and semi-supervised kernel LPE

In the last decades, semi-supervised learning has attracted an increasing amount of attention. Recently, there are considerable interests and successes on graph-based semi-supervised learning algorithms, which consider the graph over all the samples as a prior guide to decision making. All these algorithms considered the problem of classification. Here, we show that LPE can be directly used in semi-supervised learning.

Given a labeled set $X = \{(x_i, y_i)\}_{i=1}^{l}$ belonging to c classes and an unlabeled set $X_U = \{x_i\}_{i=l+1}^{u}$. The kth class have l_k samples, $\sum_{k=1}^{c} l_k = l$. Without loss of generality, we assume that the data points in X are ordered according to their labels. Let $X_T = (X, X_U)$, then according to Sect. 2.1, all the data, including the labeled and unlabeled data, are attended to learn the locally principal direction. On the contrary, only the labeled data set is attended to learn the locally discriminant direction and characterize the local

Fig. 3 The sample images of one person from ORL face database

compactness. Thus, we have the semi-supervised LPE generalized Eigen equation:

$$[X_T L X_T^{\mathrm{T}} + X L^I X^{\mathrm{T}}]V = \lambda X L^C X^{\mathrm{T}} V. \quad (17)$$

Similarly, it is clear that the optimal projection matrix V is composed of the eigenvectors associated with the first d top eigenvalues by solving the generalized Eigen equation (17). Furthermore, we can obtain the generalized Eigen equation of the semi-supervised kernel LPE:

$$[\Phi(X_{\mathrm{T}}) L \Phi^{\mathrm{T}}(X_{\mathrm{T}}) + \Phi(X) L^I \Phi^{\mathrm{T}}(X)]V = \lambda \Phi(X) L^C \Phi^{\mathrm{T}}(X) V \quad (18)$$

3 Experiments and discussions

To evaluate the proposed LPE algorithm, we systematically compare it with the PCA, LDA, CMVM, and MFA algorithm in real-work databases: ORL and Yale face database and USPS handwriting digital database. The ORL database was used to evaluate the performance of LPE under the conditions where the pose and sample size are varied. The Yale database was used to examine the system performance when both facial expressions and illumination are varied. The USPS database was employed to test the performance of the system under conditions where there are large sample size problems. Nearest neighborhood classifier with Euclidean distance was used in all the experiments.

3.1 Experiment on ORL face database

The ORL database is used to evaluate the performance of LPE under conditions where the pose, face expression, and sample size vary. The ORL face database contains images from 40 individuals, each providing 10 different images. The facial expressions and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20°. Moreover, there is also some variations in the scale of up to about 10%. All images are normalized to a resolution of 56 × 46. Some sample images are shown in Fig. 3. In the experiment, the first five samples per class are used for training and the remaining for test. Note that LDA, MFA, CMVM, and LPE all involve a PCA phase. The results are shown in Table 1 and Fig. 4.

Table 1 The maximal recognition rates (percent) of PCA, LDA, MFA, LPE on the ORL database and the corresponding dimensions (shown in parentheses) when the first five samples per class are used for training and the remaining for test

Method	PCA	LDA	MFA	CMVM	LPE
Recognition rate (%)	89.00	91.50	94.00	91.50	95.50
Dimension	34	34	48	36	38

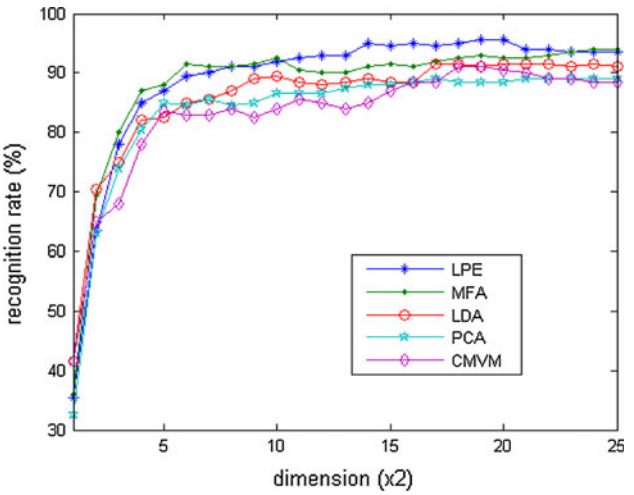

Fig. 4 The recognition rates (%) of PCA, LDA, MFA, CMVM, and LPE versus the dimensions when the first five images per person were used for training on the ORL face database

Fig. 5 Sample images of one person in the Yale database

From this experiment, we find that LPE can achieve higher recognition rate on ORL face database when the training sample number is small. The reason is that the local neighbor relationship characterized in LPE can provide important discriminant information. Moreover, the local adjacency graph and interclass graph of LPE can provide more discriminant information than the penalty graph; thus, the top recognition rates of LPE are higher than MFA and CMVM.

3.2 Experiment on Yale face database

The Yale face database contains 165 images of 15 individuals (each person providing 11 different images) under various facial expressions and lighting conditions. In our experiments, each image was manually cropped and resized to 100 × 80 pixels. Figure 5 shows sample images of one person. For computational effectiveness, we down sample it to 50 × 40 in this experiment.

In the experiment, we focus on the case that there are outliers (left-light and right-light images can be viewed as outliers) in training set and text set. The experiment was performed using the first six images (i.e., center-light, with glasses, happy, left-light, without glasses, and normal) per class for training and the remaining five images (i.e., right-light, sad, sleepy, surprised, and winking) for testing. For feature extraction, we used, respectively, PCA (eigenface), LDA (fisherface), MFA, CMVM, and the proposed LPE. Note that LDA, MFA, CMVM, and LPE all involve a PCA phase. The maximal recognition rate of each method and the corresponding dimension are given in Table 2.

Table 2 The maximal recognition rates (percent) of PCA, LDA, MFA, CMVM, and LPE on the Yale database and the corresponding dimensions when the first six samples per class are used for training

Method	PCA	LDA	MFA	CMVM	LPE
Recognition rates (%)	92	93.33	94.67	96.00	97.33
Dimensions	30	18	16	14	16

As it is shown in Table 2 and Fig. 6, the top recognition rate of LPE is significantly higher than the other methods. Why can LPE significantly outperform the other algorithms? An important reason may be that LPE not only finds the local discriminant direction but also preserves the adjacency relationship and the local principle direction of data points at the same time, thus eliminating more negative influence of outliers.

3.3 Experiments on USPS handwriting database

The USPS handwriting digital data [28] include 10 classes from "0" to "9". Each class has 1,100 examples. In our experiment, we select a subset from the original database. We cropped each image to be of size 16 × 16. There are 100 images for each class in the subset, and the total number is 1,000. The first 50 images are used as training samples, and the rest 50 images are applied to test. Figure 7

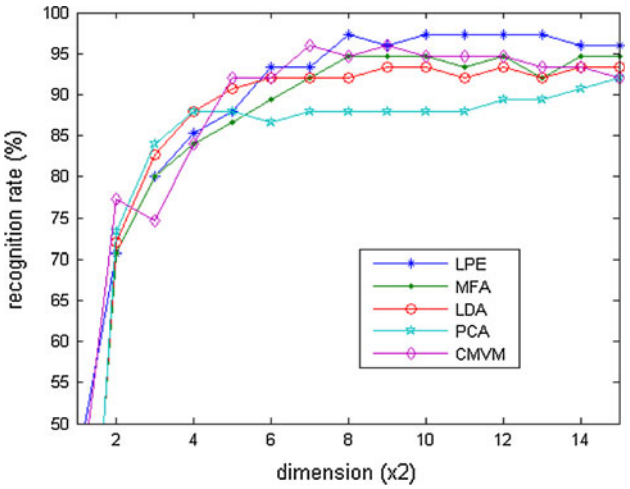

Fig. 6 The recognition rates (%) of PCA, LDA, MFA, CMVM, and LPE versus the dimensions when the first six images per person were used for training on the Yale face database

Fig. 7 The sample digital images "2" from USPS handwriting database

Table 3 The maximal recognition rates (percent) of PCA, LDA, MFA, CMVM, and LPE on the USPS database and the corresponding dimensions when 50 samples per class are used for training

Method	PCA	LDA	MFA	CMVM	LPE
Recognition rates (%)	88.0	89.0	90.2	84.60	90.8
Dimensions	12	8	8	12	8

displays a subset of digital "2" from original USPS handwriting digital database.

In this experiment, we first perform PCA on the data and then apply LDA, MFA, CMVM, and LPE on the PCA subspace for feature extraction. Table 3 shows the best recognition rates and the corresponding dimensions after carrying out PCA, LDA, MFA, CMVM, and LPE. Figure 8 displays the recognition rate curves versus the subspace dimensions by performing these approaches. It can be found that LPE also gains the best classification result compared with the other methods.

3.4 Discussions

According to the experimental results systematically performed on the three databases, we can find that:

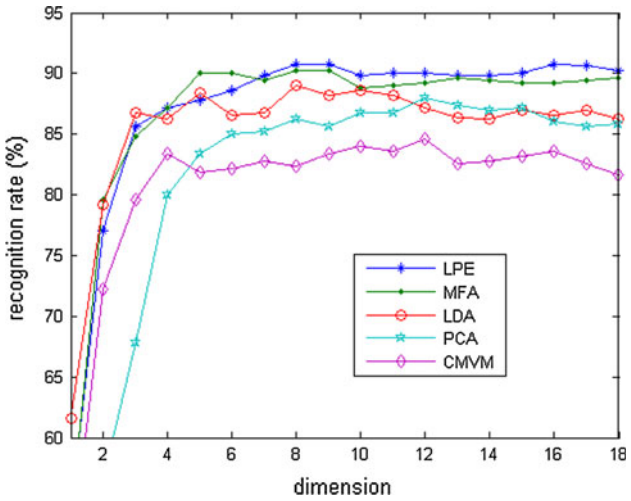

Fig. 8 The recognition rates (%) of PCA, LDA, MFA, CMVM, and LPE versus the dimensions when 50 images per class were used for training on the USPS database

1. Compared with some feature extraction methods, the proposed LEP can gain better recognition rates. For manifold learning-based methods, MFA and LPE consistently outperform classical linear method, such as PCA and LDA. But the performance of CMVM is related to databases. In Yale database, CMVM gains a good recognition rate. However, CMVM performs not so well on ORL and USPS databases.
2. Except PCA and LDA, the remaining three methods involved in all the experiments are manifold learning-based approaches, where K nearest neighborhood search is contained. For most manifold learning-based methods, how to select the K nearest neighborhood is very important. In [26], CMVM can achieve its best performance with a small K among 3–10 corresponding to different databases. Similarly, in [13], MFA should search a big range to obtain the optimal K, which is varied on different databases and hard to be chosen. In LPE framework, we figure out that the top recognition rate of LPE is relatively robust to K. Usually, when $2 \leq K \leq S$ (S is the number of training samples in each class), LPE will perform well and achieve the higher recognition rates. Therefore, LPE is more robust to the parameter than that of CMVM and MFA.
3. It should be noted that how to construct the local neighbor graph is very important for manifold learning-based algorithms. In our three experiments mentioned previously, it seems that local neighbor graph of sub-manifold defined in MFA and LPE is superior to characterize the local data than the dissimilarities graph in CMVM. Thus, MFA and LPE perform better than CMVM.

4 Conclusions

In this paper, we develop a supervised learning technique, called locality preserving embedding (LPE), for dimensionality reduction in high-dimensional data. LPE can give a low-dimensional embedding result for the discrimination of multi-class sub-manifolds and preserve principal structure information of local sub-manifolds. The projection of LPE can be viewed as a linear approximation of the nonlinear map that uncovers and separates embeddings corresponding to different manifolds in the final embedding space. In LPE framework, supervised and unsupervised idea are combined together to learn the optimal discriminant projections. On the one hand, the class information is taken into account to characterize the compactness of local sub-manifolds and the separability of different sub-manifolds. On the other hand, at the same time, all the samples in the local neighborhood are used to characterize and preserve the natural data distribution in low-dimensional subspace. Thus, LPE optimally preserves the local sub-manifold's structure to separate different sub-manifolds as far as possible. The experimental results on Yale, ORL, and USPS database show that LPE consistently outperforms the classical linear methods and the recently proposed manifold learning-based methods.

Acknowledgments This work is partially supported by the National Science Foundation of China under grant No. 60503026, 60632050, 60473039, 60873151, 61005005 and Hi-Tech Research and Development Program of China under grant No. 2006AA01Z119.

References

1. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22(1):4–37
2. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceeding of the fourth international conference of face and gesture recognition, Grenoble, France, pp 46–53
3. Dubuisson S, Davoine F, Masson M (2002) A solution for facial expression representation and recognition. Signal Process Image Commun 17(9):657–673
4. Joliffe I (1986) Principal component analysis. Springer, New York
5. Fukunnaga K (1991) Introduction to statistical pattern recognition, 2nd edn. Academic Press, London
6. Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
7. He X, Niyogi P (2003) Locality preserving projections. In: Proceedings of 16th conference neural information processing systems
8. Goh A, Vidal R (2008) Clustering and dimensionality on Riemannian manifolds. IEEE Int Conf Comput Vis Pattern Recogn 1:1–7
9. Chung F (1997) Spectral graph theory. Regional conference series in mathematics, no. 92
10. Jin Z, Yang J, Hu Z, Lou Z (2001) Face recognition based on the uncorrelated discrimination transformation. Pattern Recogn 34(7):1405–1416
11. Belhumeour PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
12. Li H, Jiang T, Zhang K (2004) Efficient and robust feature extraction by maximum margin criterion. In: Proceedings of the advances in neural information processing systems, vol 16. MIT Press, Vancouver, Canada
13. Yan S, Xu D, Zhang B, Zhang H, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell (T-PAMI) 29(1):40–51
14. Tenenbaum JB, deSilva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
15. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
16. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396
17. Kouropteva O, Okun O, Pietikainen M (2003) Supervised locally linear embedding algorithm for pattern recognition. Lect Notes Comput Sci 2652:386–394
18. Ridder D, Loog M, Reinders M (2004) Local fisher embedding. In: Proceedings of the 17th international conference on pattern recognition
19. Vlassis N, Motomura Y, Krose B (2002) Supervised dimension reduction of intrinsically low dimensional data. Neural Comput 14(1):191–215
20. Geng X, Zhang DC, Zhou ZH (2005) Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Trans Syst Man Cybern B 35(6):1098–1107
21. Zhao HT, Sun SY, Jing ZL, Yang JY (2006) Local structure based supervised feature extraction. Pattern Recogn 39:1546–1550
22. Yang J, Zhang D, Yang J-y, Niu B (2007) Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics. IEEE TPAMI 29(4):650–664
23. Deng W, Hu J, Guo J, Zhang H, Zhang C (2008) Comments on 'globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics'. IEEE PAMI 30(8):1503–1504
24. Chen H-T, Chang H-W, Liu T-L (2005) Local discriminant embedding and its variants. IEEE Conf Comput Vis Pattern Recogn 2:846–853
25. Zhang T, Yang J, Wang H, Du C (2007) Maximum variance projection for face recognition. Opt Eng 46(6):1–8
26. Bo L, Huang D-S, Wang C, Liu K-H (2008) Feature extraction using constrained maximum variance mapping. Pattern Recogn 41:3287–3294
27. Lai Z, Wan M, Jin Z (2009) Locality preserving embedding. In: Proceedings of the first international conference on information science and engineering, Nanjing, China, pp 895–899
28. http://www.cs.nyu.edu/~roweis/data.html