Neural Process Lett (2011) 33:267–282
DOI 10.1007/s11063-011-9177-x

Locally Minimizing Embedding and Globally Maximizing Variance: Unsupervised Linear Difference Projection for Dimensionality Reduction

Minghua Wan · Zhihui Lai · Zhong Jin

Published online: 2 May 2011

Abstract Recently, many dimensionality reduction algorithms, including local methods and global methods, have been presented. The representative local linear methods are locally linear embedding (LLE) and linear preserving projections (LPP), which seek to find an embedding space that preserves local information to explore the intrinsic characteristics of high dimensional data. However, both of them still fail to nicely deal with the sparsely sampled or noise contaminated datasets, where the local neighborhood structure is critically distorted. On the contrary, principal component analysis (PCA), the most frequently used global method, preserves the total variance by maximizing the trace of feature variance matrix. But PCA cannot preserve local information due to pursuing maximal variance. In order to integrate the locality and globality together and avoid the drawback in LLE and PCA, in this paper, inspired by the dimensionality reduction methods of LLE and PCA, we propose a new dimensionality reduction method for face recognition, namely, unsupervised linear difference projection (ULDP). This approach can be regarded as the integration of a local approach (LLE) and a global approach (PCA), so that it has better performance and robustness in applications. Experimental results on the ORL, YALE and AR face databases show the effectiveness of the proposed method on face recognition.

Keywords Locally linear embedding (LLE) · Linear preserving projections (LPP) · Dimensionality reduction · Principal component analysis (PCA) · Face recognition

M. Wan · Z. Lai · Z. Jin
School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China

M. Wan (✉)
School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China
e-mail: wmh36@sina.com

Z. Lai
Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518005, China

1 Introduction

Techniques for dimensionality reduction in linear and nonlinear learning tasks have attracted much attention in the areas of pattern recognition and computer vision. Linear dimensionality reduction seeks to find a meaningful lower-dimensional subspace in a higher-dimensional input space. The subspace can provide a compact representation of the input data when the structure of data embedded in the input space is linear. Two of the most fundamental linear dimensionality reduction methods are principal component analysis (PCA) [1–3] and linear discriminant analysis (LDA) [3–5].

PCA aims to find a linear mapping, which preserves the total variance by maximizing the trace of the feature covariance matrix. The optimal projection of PCA is corresponding to the first d-largest eigenvalues of the data's total covariance matrix. LDA is used to find the optimal set of projection vectors by maximizing the ratio between the interclass and intraclass scatters. However in applications, the dimension of vectors is high and the number of samples is small. 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 avoid the singularity problem of LDA, Li et al. [6] used the difference of both between-class scatter and within-class scatter as discriminant criterion, called maximum margin criterion (MMC), to find the optimal projections. Since the inverse matrix is not necessary to compute in MMC, the SSS problem in traditional LDA is alleviated. So far many effective and efficient methods [7–17] have been explored to solve the SSS problem. Both PCA and LDA and their extensions have been successfully applied to linear data. However, they may fail to explore the essential structure of the data with nonlinear distribution.

In the real world applications, nonlinear data include non-Gaussian and manifold-value data. We usually deal with non-Gaussian data from local patches because it can be viewed locally Gaussian, and the curved manifold-value data can be viewed locally Euclidean [18,19]. In order to deal with nonlinear data, many nonlinear feature extraction methods have been developed such as kernel-based techniques and manifold learning based techniques. Kernel-based technique is implicitly mapping the observed patterns into potentially much higher-dimensional feature space by a kernel trick such that the nonlinear structure data will be linearly separable in the kernel space. The widely used kernel techniques are kernel principal component analysis (KPCA) [20] and kernel Fisher discriminant analysis (KFDA) [21], 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 high computational requirements with increasing number of dimensions. Furthermore, due to introducing the kernel trick, how to select the effective kernels and how to assign the optimal parameters in kernel techniques remain unclear.

Unlike kernel-based methods, manifold learning-based methods are straightforward in finding the inherent nonlinear structure hidden in the observe space. Recently, many manifold learning-based algorithms with locality preserving abilities have been presented. Among them, isometric feature mapping (ISOMAP) [22], locally linear embedding (LLE) [23,24], Laplacian eigenmap (LE) [25,26] and local tangent space alignment (LTSA) [27] are widely used. He et al. [28,29] proposed locality preserving projections (LPP), which is a linear subspace learning method derived from Laplacian eigenmap. LPP preserved local information and best detected the essential face manifold structure. The optimal projection axes of LPP preserves the local structure of the underlying distribution in the L2 Euclidean space. LPP finds an embedding space that preserves local information, and it is an unsupervised method. Many modified LPP algorithms have been put forward to consider the discriminant

information of recognition task in recent years [30–33]. LPP is modeled based on the characterization of "locality". However, this modeling has no direct connection to classification. The objective function of LPP is to minimize the local quantity, i.e., the local scatter of the projected data. This criterion cannot guarantee to yield a good projection for classification in some cases where the "nonlocality" provides dominant information for discrimination. So, Yang et al. [34] proposed an unsupervised discriminant projection (UDP) algorithm considering the nonlocal and local quantities at the same time, which could be viewed as a simplified or regularized version of LPP. In contrast with LPP, UDP has intuitive relations to classification since it utilizes the information of the "nonlocality". Provided that each cluster of samples in the observation space is exactly within a local neighborhood, UDP can yield an optimal projection for clustering in the projected space, while LPP cannot.

Recently, He et al. [35] proposed another linear dimensionality reduction technique neighborhood preserving embedding (NPE), which is the linearization of the LLE algorithm and aims at finding a low-dimensional embedding that optimally preserves the local neighborhood reconstruction relationships on the original data manifold. Some extension methods of NPE [36,37] were introduced to feature extraction. However, locality preserving methods fail to nicely deal with the sparsely sampled or noise contaminated datasets, particularly, when the local neighborhood structure is critically distorted.

In order to integrate the locality preserving (LLE) and globality preserving (PCA) together and avoid the drawbacks in LLE and PCA, in this paper, we propose a new subspace learning algorithm, named unsupervised linear difference projection (ULDP) for face recognition. LLE is an efficient dimensional reduction algorithm for nonlinear data, and the low dimensional data can maintain topological relations in the original space. On the contrary, PCA is a global linear method which preserves the total variance by maximizing the trace of feature variance matrix. In addition, LLE as well as PCA are unsupervised learning methods, so ULDP is an unsupervised method. So, ULDP is a linear difference projection algorithm that aims at preserving both local and global information by capturing both local and global geometry of the manifold.

The rest of this paper is organized as follows: We review the ideas of linear methods in Sect. 2. In Sect. 3, we propose the idea of ULDP algorithm in detail. In Sect. 4, we introduce the connections between LLE, NPE and ULDP. Experiments are presented to demonstrate the effectiveness of ULDP on face recognition in Sect. 5. Finally, we give concluding remarks and a discussion of future work in Sect. 6.

2 Outline of Linear Methods

Let us consider a set of N data vectors $X = \{x_1, x_2, \ldots, x_N\}, x_i \in R^n$ taking values in an n-dimensional image space. Let us also consider a linear transformation mapping the original n-dimensional space into an d-dimensional feature space $Y = \{y_1, y_2, \ldots, y_N\}$, where $y_i \in R^d$ and $n > d$. The new feature vectors $y_i \in R^d$ are defined by the following linear transformation:

$$y_i = U^T x_i, \quad i = 1, \ldots, N \tag{1}$$

where $U \in R^{n \times d}$ is a transformation matrix. In this section, we briefly review how the LDA, LPP and UDP algorithms realize subspace learning.

2.1 Linear Discriminant Analysis

Linear discriminant analysis [2] is a supervised learning algorithm. Let c denote the total number of classes and l_i denote the number of training samples in the ith class. Let x_i^j, denote the jth sample in ith class, $\bar{x}$ be the mean of all the training samples, $\bar{x}_i$ be the mean of the ith class. The between-class and within-class scatter matrices can be evaluated by:

$$S_b = \sum_{i=1}^{c} l_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{2}$$

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{l_i} \left(x_i^j - \bar{x}_i\right) \left(x_i^j - \bar{x}_i\right)^T \tag{3}$$

Linear discriminant analysis aims to find an optimal projection U such that the ratios of the between-class scatter to within-class scatter is maximized, i.e.

$$U = \arg\max_U \frac{\left|U^T S_b U\right|}{\left|U^T S_w U\right|} \tag{4}$$

where $U = \{u_i | i = 1, 2, \ldots, d\}$ is the set of generalized eigenvectors of S_b and S_w corresponding to the d largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \ldots, d\}$, i.e.

$$S_b u_i = \lambda_i S_w u_i, \quad i = 1, 2, \ldots, d. \tag{5}$$

2.2 Linear Preserving Projection

The matrix W of linear preserving projection (LPP) [25,26] is a similarity matrix, which can be Gaussian weight or uniform weight of Euclidean distance using k-neighborhood or ε-neighborhood. W is defined as:

$$W_{ij} = \begin{cases} 1, & \left\|x_i - x_j\right\|^2 < \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Hence, the objective function of LPP is defined as:

$$\begin{aligned} &\min \sum_{i,j} \left\|y_i - y_j\right\| W_{ij} \\ &\text{s.t. } U^T X D X^T U = 1 \end{aligned} \tag{7}$$

where $\|\cdot\|$ means the L_2 norm. In order to avoid the trivial solution, we have added a constraint which can make $y_i \neq 0$;. After some matrix analysis steps, the minimization problem becomes

$$\begin{aligned} &\arg\min_U U^T X L X^T U \\ &\text{s.t. } U^T X D X^T U = 1 \end{aligned} \tag{8}$$

where $X = [x_1, x_2, \ldots, x_N]$ is the training space of size $n \times N$, and D is a diagonal matrix whose entries are column or row sums of S. $L = D - S$ is the Laplacian matrix.

The optimal d projection vectors that minimizes the objective function can be computed by the minimum eigenvalues solutions to the generalized eigenvalues problem

$$X L X^T u_i = \lambda_i X D X^T u_i \tag{9}$$

2.3 Unsupervised Discriminant Projection

In UDP [31] algorithm, the local k-neighbor adjacency matrix W^l is defined as:

$$W_{ij}^l = \begin{cases} 1, & x_i \in N_{k_w}^+(x_j) \quad \text{or} \quad x_j \in N_{k_w}^+(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $x_i \in N_{k_w}^+(x_j)$ *or* $x_j \in N_{k_w}^+(x_i)$ indicates the index set of the k nearest neighbors of the sample x_i or x_j.

The local scatter matrix is defined as:

$$S_l = \frac{1}{2}\frac{1}{NN}\sum_{i,j} W_{ij}^l \left\| x_i - x_j \right\|^2 = \frac{1}{NN} X(D^l - W^l)X^T = \frac{1}{NN} X L^l X^T \tag{11}$$

where D^l is a diagonal matrix whose elements on diagonal are column sum of W^l, i.e. $D_{ii}^l = \sum_{i \neq j} W_{ij}^l$, $L^l = D^l - W^l$.

The nonlocal scatter matrix is defined as:

$$S_n = \frac{1}{2}\frac{1}{NN}\sum_{i,j}\left(1 - W_{ij}^l\right)\left\| x_i - x_j \right\|^2 = S_T - S_l \tag{12}$$

S_T is the total scatter matrix:

$$\begin{aligned} S_T &= \frac{1}{2NN}\sum_{i,j}(x_i - x_j)(x_i - x_j)^T \\ &= \frac{1}{2NN}\left[N\sum_i x_i x_i^T - \left(\sum_i x_i\right)\left(\sum_j x_i^T\right)\right] \\ &= \frac{1}{N}\sum_i (x_i - \bar{x})(x_i - \bar{x})^T \end{aligned} \tag{13}$$

where $\bar{x} = \frac{1}{N}\sum_i x_j$ is the mean vector.

We can obtain just a projection matrix U by maximizing the following criterion:

$$U = \arg\max_U \frac{\left|U^T S_n U\right|}{\left|U^T S_l U\right|} \tag{14}$$

The optimal d projection vectors that maximizes the objective function can be computed by the maximum eigenvalues solutions of the generalized eigenvalues problem

$$S_n u_i = \lambda_i S_l u_i \tag{15}$$

3 Unsupervised Linear Difference Projection

3.1 The Idea of ULDP

Principal component analysis is essentially a global technique which cannot discover the local structure of the data, whereas LLE is local relation of the data in embedding space which cannot preserve the global structure of the dataset. To take advantages of LLE and PCA, a new technique, called ULDP based on preserving local embedding and global variance is proposed, which is able to do nonlinear dimensionality reduction in an unsupervised way.

To gain more discriminative power, it is desirable to minimize the locality and maximize the globality simultaneously.

3.2 Locally Minimizing Embedding

To begin with, we propose to minimize the local scatter compactness of each data point by linear coefficients that reconstruct the data point from other points. The technique of local representation is the same as LLE [20,21]. LLE regards each data point and its nearest neighbors as the locality. The algorithm can be described in three steps.

The first step of LLE is to select k-nearest neighbors of each data points x_i using Euclidean distances.

The second step of LLE is to calculate the reconstructing weight matrix $W = \left[w_{ij}\right]_{N\times N}$, which reconstructs each point x_i from its k-nearest neighbors. We can obtain the coefficient matrix W by minimizing the reconstruction error:

$$\min J_L(W) = \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{N} w_{ij} x_j \right\|^2 \tag{16}$$

where $w_{ij} = 0$ if x_i and x_j are not neighbors, and the rows of W sum to 1: $\sum_{j=1}^{N} w_{ij} = 1$.

The reconstruction error for x_i can be converted to this form:

$$\begin{aligned} \xi_i &= \left\| x_i - \sum_{j=1}^{N} w_{ij} x_j \right\|^2 = \left\| \sum_{j=1}^{N} w_{ij} \left(x_i - x_j\right) \right\|^2 \\ &= \sum_{j=1}^{N} w_{ij} \left(x_i - x_j\right) \sum_{t=1}^{N} w_{it} \left(x_i - x_t\right) = \sum_{j=1}^{N} \sum_{t=1}^{N} w_{ij} w_{it} G^i_{jt} \end{aligned} \tag{17}$$

where $G^i_{jt} = \left(x_i - x_j\right)^T \left(x_i - x_t\right)$, called the local Gram matrix. By solving the least-squares problem with the constraint $\sum_{j=1}^{N} w_{ij} = 1$, the optimal coefficients are given:

$$w_{ij} = \frac{\sum_{t=1}^{N} \left(G^i\right)^{-1}_{jt}}{\sum_{p=1}^{N} \sum_{q=1}^{N} \left(G^i\right)^{-1}_{pq}} \tag{18}$$

After repeating the first step and the second step are performed on all the N data points, we can calculate the reconstruction weights to construct a weight matrix $W = \left[w_{ij}\right]_{N\times N}$.

The third step of LLE is to reconstruct represented y_i by the weight matrix W. To maintain the intrinsic geometrical feature of the data after the embedding process, the reconstruction error function must be minimized:

$$\min J_L(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{N} w_{ij} y_j \right\|^2 \tag{19}$$

where y_i is the mapping output of x_i, y_j is a neighbor of y_i.

Considering the map in Eq. 1, the objective function reduces to

$$\begin{aligned} J_L(U) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{N} w_{ij} y_j \right\|^2 &= \sum_{i=1}^{N} tr \left\{ \left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right) \left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right)^T \right\} \\ &= tr \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right) \left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right)^T \right\} \\ &= tr \left\{ Y \left(I - W^T \right) \left(I - W^T \right)^T Y^T \right\} \\ &= tr \left\{ Y \left(I - W \right)^T \left(I - W \right) Y^T \right\} \\ &= tr \left\{ U^T X M X^T U \right\} \end{aligned} \tag{20}$$

where $M = (I - W)^T (I - W)$.

3.3 Globally Maximizing Variance

On the second hand, we propose to maximize the sum of pairwise squared distances between outputs, where PCA [1] preserves the global geometric structure of data in a transformed low-dimensional space. So, maximizing the global scatter of samples is considered:

$$\max J_G(Y) = \sum_{i=1}^{N} \| y_i - \bar{y} \|^2 \tag{21}$$

Considering the map equation (1), the objective function reduces to

$$\begin{aligned} J_G(U) = \sum_{i=1}^{N} \| y_i - \bar{y} \|^2 &= \sum_{i=1}^{N} \left\| U^T \left(x_i - \bar{x} \right) \right\|^2 \\ &= \sum_{i=1}^{N} tr \left\{ U^T (x_i - \bar{x})(x_i - \bar{x})^T U \right\} \\ &= tr \left\{ U^T S_T U \right\} \end{aligned} \tag{22}$$

where $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$, $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$, and S_T is the total scatter matrix in Eq. 13.

3.4 Optimization Criterion of ULDP

At last, when local minimizing embedding and global maximizing variance have been constructed, an intuitive motivation is to find a common projection that minimizes local scatter $J_L(W)$ and maximizes global scatter $J_G(U)$ at the same time. Actually, we can obtain such a projection by the following multi-object optimized problem, that is:

$$\begin{cases} \min \quad tr \left\{ U^T X M X^T U \right\} \\ \max \quad tr \left\{ U^T S_T U \right\} \\ \quad \text{s.t.} \;\; U^T X D X^T U = I \end{cases} \tag{23}$$

where D is a diagonal matrix in Eq. 11. The solution to the constrained multi-object optimized problem is to find a subspace which preserve the locality property and maximize the global scatter variance simultaneously. Motivated by the idea of MMC [3], ULDP seeks to minimize the difference, rather than the ratio, between the local minimizing embedding and the global maximizing variance. So it can be changed into the following constrained problem:

$$\begin{aligned} &\min tr\left\{U^T\left((1-\alpha)XMX^T-\alpha S_T\right)U\right\} \\ &\text{s.t. } U^TXDX^TU=I \end{aligned} \tag{24}$$

where α $(0 \leq \alpha < 1)$ is an adjustable parameter to balance the locality and globality.

Equation 24 can be solved by Lagrangian multiplier method:

$$\frac{\partial}{\partial u_i} tr\left\{u_i^T\left((1-\alpha)XMX^T-\alpha S_T\right)u_i-\lambda_i\left(u_i^TXDX^Tu_i-I\right)\right\}=0 \tag{25}$$

where λ_i is the Lagrangian multiplier. Thus we get:

$$\left[(1-\alpha)XMX^T-\alpha S_T\right]u_i=\lambda_i XDX^Tu_i \tag{26}$$

where u_i is generalized eigenvector correspondingly to generalized eigenvalue λ_i.

After the transformation matrix of ULDP is obtained, the samples can be projection to this low-dimensional subspace. Then a nearest-neighbor classifier can be used for classification.

Given two images x_1, x_2 represented by ULDP feature vectors y_1 and y_2, then the distance $d(y_1, y_2)$ is defined as:

$$d(y_1, y_2) = \|y_1 - y_2\|^2 \tag{27}$$

If the feature matrices of training images are $y_1, y_2, \ldots, y_N$(N is the total number of training images), and each image is assigned to a class $\pi_l (l = 1, \ldots, c)$. Then for a given test image y, if $d(y, y_1) = \min_j d(y, y_j)$ and $y_1 \in \pi_l$, the resulting decision is $y \in \pi_l$.

3.5 The Outline of ULDP

Based on the above descriptions, ULDP algorithm can be described as follows:

Step1: (PCA): PCA transformation is implemented on original image spaces using Eq. 22.
Step2: (ULDP): Calculate preserving local minimizing embedding using Eqs. 16–20 and preserving global maximizing variance using Eqs. 21–24 in PCA subspace.
Step3: (Feature extracting): Extract the sample feature using Eq. 26.
Step4: (Dimension reduction): Project all samples onto the obtained optimal discriminant vectors and yield the projected eigenvectors using Eq. 1.
Step5: (Recognition): Classify the projected eigenvectors with a classifier using Eq. 27.

4 Connection Between LLE, NPE and ULDP

In this section, ULDP will be shown to be formally similar to LLE and NPE [32]. However, ULDP is also obviously different from them. In order to investigate the similarity and the difference, we discuss the connections between LLE, NPE and ULDP.

LLE and NPE aim to discover the local structure of the data manifold. LLE is defined only on the training samples, and there are no natural maps of the testing sample. Instead, NPE is defined on both the training and test samples. NPE is a linear approximation to LLE.

In NPE, the matrix XMX^T is symmetric and semi-positive definite. In order to remove an arbitrary scaling factor in the projection, we impose a constraint as follows:

$$YY^T = I \Rightarrow U^T XX^T U = I \tag{28}$$

Finally, the minimization problem reduces to finding U:

$$\min_{U^T XX^T U=I} tr\left\{U^T XMX^T U\right\} \tag{29}$$

The transformation matrix U that minimizes the objective function is given by the minimum eigenvalue solution to the following generalized eigenvector problem:

$$XMX^T u_i = \lambda_i XX^T u_i \tag{30}$$

ULDP preserves local embedding and global variance. The total scatter matrix S_T in Eq. 13 can be written as:

$$\begin{aligned} S_T &= \frac{1}{2NN}\sum_{i,j}(x_i - x_j)(x_i - x_j)^T \\ &= \frac{1}{N}X\left(I - \frac{1}{N}ee^T\right)X^T \end{aligned} \tag{31}$$

where I is the identity matrix and e is a column vector taking 1 at each entry.

As a result, ULDP is formulated as the following constrained minimization problem:

$$\min_{U^T XDX^T U=I} tr\left\{U^T X\left((1-\alpha)M + \alpha\left(\frac{1}{N^2}ee^T - \frac{1}{N}I\right)\right)X^T U\right\} \tag{32}$$

Thus we have:

$$X\hat{M}X^T u_i = \lambda_i XX^T u_i \tag{33}$$

where $\hat{M} = (1-\alpha)M + \alpha\tilde{M}$, $\tilde{M} = \frac{1}{N^2}ee^T - \frac{1}{N}I$. It is easy to see that NPE is a special case of ULDP (i.e. when $\alpha = 0$, $D = I$).

From above discussed, NPE and ULDP yield mappings that are defined not only on the training data points but also on novel testing points. The essence of NPE is the linear approximation to LLE and the essence of ULDP is weighted NPE integrating globality-based PCA. As we know, the graph construction of LLE and NPE fails to use the global discriminative information while the graph construction of PCA fails to utilize the locality information. However, we can see from $\hat{M}$ first that ULDP preserves the locality characteristic since M still exists and second that it adds the global information through $\tilde{M}$. From what has been discussed above, it can be concluded that ULDP builds a new graph with different edge weight assignment method, integrating both local information and global information. Thus, by integrating the globality into the objective function, ULDP will be more robust than LLE and NPE.

5 Experiments and Results

To evaluate the proposed ULDP algorithm, we systematically compare it with the PCA [1], LDA [2], LLE [20,21], NPE [32], LPP [25,26] and UDP [31] algorithm in three face databases: ORL, YALE and AR. When the projection matrix was computed from the training part, all the images including the training part and the test part were projected to feature space.

Euclidean distance and nearest neighborhood classifier are used in all the experiments. The experiments were carried out on the same PC (CPU: P4 2.8 GHz, RAM: 1024 MB).

5.1 Database

The ORL face database (http://www.uk.research.att.com/facedatabase.html) contains images from 40 individuals, each providing 10 different images where the pose, face expression and sample size vary. 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 degrees. Moreover, there is also some variation in this scale of up to about 10%. All images normalized to a resolution of 56×46. We test the recognition performances of the seven methods: PCA, LDA, LLE, NPE, LPP, UDP and ULDP. In the experiments, l images (l varies from 2 to 6) are randomly selected from the image gallery of each individual to form the training sample set. The remaining $10 - l$ images are used for testing. For each l, we independently run 50 times. In the PCA phase of LDA, LLE, NPE, LPP, UDP and ULDP, we keep 95% image energy.

The YALE face database (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 gray scale images of 15 individuals, each individual has 11 images. The images demonstrate variations in lighting condition, facial expression (normal, happy, sad, sleepy, surprised, and wink). In this experiment, each image in Yale database was manually cropped and resized to 50×40. In the PCA phase of LDA, LLE, NPE, LPP, UDP and ULDP, we keep 95% image energy. In the experiments, l images (l varies from 2 to 6) are randomly selected from the image gallery of each individual to form the training sample set. The remaining $11 - l$ images are used for testing. For each l, we independently run 50 times.

The AR face database (http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html) contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of 120 individuals (65 men and 55 women) were taken in two sessions (separated by 2 weeks) and each session contains 13 color images. The face portion of each image is manually cropped and then normalized to 50×40 pixels. These images vary as follows: (1) neutral expression, (2) smiling, (3) angry, (4) screaming, (5) left light on, (6) right light on, (7) all sides light on, (8) wearing sum glasses, (9) wearing sun glasses and left light on, and (10) wearing sun glasses and right light on. In this experiment, l images (l varies from 2 to 6) are randomly selected from the image gallery of each individual to form the training sample set. The remaining $20 - l$ images are used for testing. For each l, we independently run 10 times. In the PCA phase of LDA, LLE, NPE, LPP, UDP and ULDP, the number of principal components is set as 150. The dimension steps are set to be 5 in final low-dimensional subspaces obtained by the seven methods.

Figures 1, 2 and 3 show the sample images from the three databases.

5.2 Experimental Results and Analysis

Except PCA and LDA, the local methods involved in the experiments are manifold learning based approaches, where k nearest neighborhood search is employed. Thus how to select parameter k is an important problem in feature extraction. If the value of k is too small, it is very difficult to preserve the topologic structure in the low-dimensional space. On the contrary, if the value of k is too large, it is very difficult to depict the assumption of local linearity in the high dimensional space. So it will affect the dimensionality manifold reduction

Fig. 1 Images of one person on the ORL database

Fig. 2 Images of one person on the YALE database

Fig. 3 Images of one subject of the AR database. The *first line* and the *second line* images were taken in different time (separated by 2 weeks)

result by the value of k. In the first experiment, we investigate the performance of the ULDP algorithm over the reduced dimensions versus the varied the value of the k. To find how k affects the recognition performance, we change k from 1 to 20 with step 1. Figure 4 displays the average recognition rate of LPP, UDP and ULDP with varied the value of k by carrying out ULDP when only six images per class were randomly selected for training on the ORL, YALE and AR face databases.

From Fig. 4 we can get the following conclusions:

(1) The average recognition rates of ULDP algorithm is not sensitive to parameter k when only six images per class were randomly selected for training on the ORL, YALE and AR face databases.
(2) The k-nearest neighborhood parameter in LPP and UDP [31] is chosen as $k = l - 1$ where they can get best recognition rates.
(3) Why LPP and UDP are affected and ULDP is not by the parameter k? Because of the essence of LPP and UDP are locality-based methods, where ULDP can preserve local and global information by capturing the local and global geometry of the manifold.

In the second experiment, we also test the impact of α on the performance when only six images per class were randomly selected for training on the YALE face database, which can be found in Fig. 5. We varied α from 0 to 0.9 with step 0.1. Figure 5 displays the maximal average recognition rates with varied parameter α by carrying out ULDP. From Fig. 5, it can be found that the effectiveness of the ULDP algorithm is sensitive to the value of the parameter α. ULDP obtains the best average recognition rate 95.09% when $\alpha = 0.5$. This indicates that the locality and the globality are with the same importance. In the next experiment, the value of adjustable parameter α is taken to be 0.5.

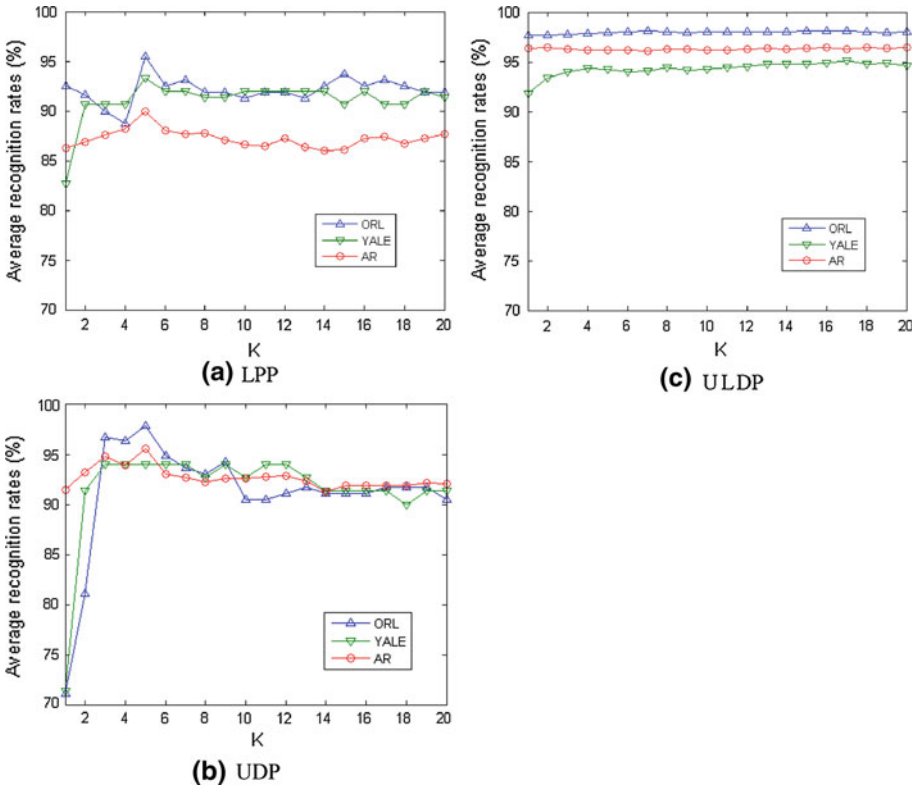

Fig. 4 The average recognition rate (%) of LPP, UDP and ULDP versus the varied the value of k when only six images per class were randomly selected for training on the ORL, YALE and AR face databases

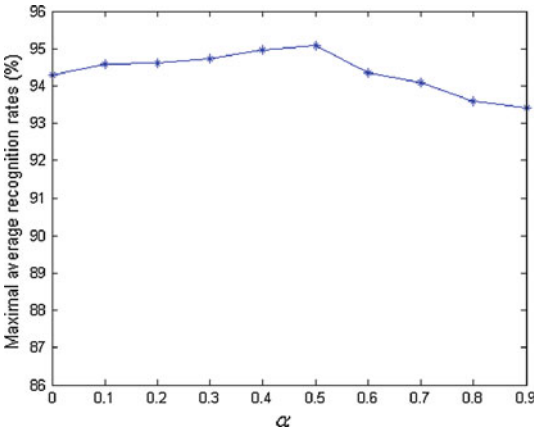

Fig. 5 The maximal average recognition rates (%) of ULDP versus the varied the value of α when only six images per class were randomly selected for training on the YALE face database

In the third experiment, we randomly select l images (l varies from 2 to 6) of each individual for training, and the remaining ones are used for testing. We compare the performances of different algorithms. The maximal average recognition rates obtained by different algorithms as well as the corresponding dimensionality of reduced subspace (the numbers in parentheses)

Table 1 The maximal average recognition rates (%) of the seven methods on the ORL face database versus the variation of the training sample sizes

Methods	Number of training samples of each class				
	2	3	4	5	6
PCA	74.91	82.23	84.53	86.71	87.40
	(50)	(46)	(34)	(46)	(42)
LDA	77.40	85.09	86.17	87.23	87.73
	(39)	(39)	(39)	(35)	(38)
LLE	70.60	73.39	77.46	80.00	89.99
	(43)	(36)	(50)	(27)	(50)
NPE	82.53	90.26	93.62	95.86	97.30
	(26)	(25)	(24)	(2)	(26)
LPP	72.05	81.78	87.42	90.82	93.21
	(48)	(46)	(36)	(34)	(32)
UDP	70.53	85.71	92.51	95.45	97.26
	(50)	(50)	(47)	(50)	(49)
ULDP	**83.66**	**91.31**	**94.63**	**96.77**	**98.06**
	(26)	**(25)**	**(25)**	**(25)**	**(24)**

The bold values denote the best results

Table 2 The maximal average recognition rates (%) of the seven methods on the YALE face database versus the variation of the training sample sizes

Methods	Number of training samples of each class				
	2	3	4	5	6
PCA	78.49	81.47	85.37	85.96	87.01
	(29)	(40)	(36)	(40)	(46)
LDA	81.93	85.61	88.30	88.84	89.36
	(14)	(14)	(14)	(14)	(14)
LLE	84.93	84.17	86.65	90.00	90.53
	(17)	(11)	(9)	(11)	(10)
NPE	89.76	92.95	91.94	93.36	94.29
	(26)	(29)	(49)	(46)	(47)
LPP	81.45	85.97	88.57	89.00	90.40
	(22)	(24)	(21)	(18)	(21)
UDP	88.36	90.77	90.50	92.76	94.75
	(50)	(50)	(50)	(27)	(29)
ULDP	**90.64**	**93.40**	**93.18**	**94.20**	**95.09**
	(25)	**(17)**	**(19)**	**(20)**	**(26)**

The bold values denote the best results

on the ORL, YALE and AR face databases are given in Tables 1, 2 and 3, respectively. Figure 6 shows the average recognition rates (%) of ULDP versus the varied dimensionality when only six images per class were randomly selected for training on the ORL, YALE and AR face databases. We change the number of eigenvectors from 2 to 50 with step 2 on the

Table 3 The maximal average recognition rates (%) of the seven methods on the AR face database versus the variation of the training sample sizes

Methods	Number of training samples of each class				
	2	3	4	5	6
PCA	67.79	71.83	78.87	79.84	82.58
	(150)	(150)	(150)	(150)	(150)
LDA	72.21	76.34	83.84	87.45	88.33
	(115)	(115)	(115)	(115)	(115)
LLE	71.29	75.47	84.84	86.54	87.38
	(135)	(120)	(120)	(135)	(135)
NPE	72.50	82.56	91.59	94.44	95.28
	(150)	(145)	(145)	(145)	(130)
LPP	72.45	76.39	84.39	88.68	89.67
	(140)	(130)	(130)	(110)	(110)
UDP	73.13	82.19	90.15	91.8	95.08
	(150)	(150)	(150)	(150)	(150)
ULDP	**73.45**	**83.21**	**91.74**	**92.44**	**96.00**
	(80)	**(145)**	**(145)**	**(125)**	**(120)**

The bold values denote the best results

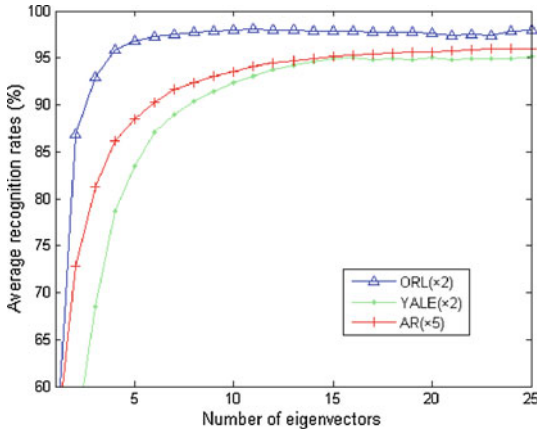

Fig. 6 The average recognition rates (%) of ULDP versus the varied dimensions when only six images per class were randomly selected for training on the ORL, YALE and AR face databases

ORL, YALE face databases and the number of eigenvectors from 5 to 125 with step 5 on the AR face database, respectively.

From Tables 1, 2 and 3, and Fig. 6, we can obtain the following conclusions:

(1) The above experiments showed that the maximal average recognition rates of all methods increase with the increase in training sample size in Tables 1, 2 and 3 respectively. The proposed ULDP algorithm consistently performs better than other methods in all experiments in three face databases.
(2) From Fig. 6 we can find that with the increasing number of eigenvectors on three face databases, the average recognition rates also improved.

(3) Local methods such as LLE, NPE, LPP and UDP retain the neighboring data points of k nearest distances, even though they might be from different classes. This information is inadequate to capture the locality of the real underlying data structure. Hence, local methods may not be optimal from a discrimination standpoint. On the other hand, ULDP seeks projection that preserves local embedding and global variance. This objective function leads to the enhancement of classification capability and this assumption is testified by the experimental results.

6 Conclusions

In pattern recognition, feature extraction techniques are widely employed to reduce the dimensionality of data and enhance the discriminatory information. In this paper, we proposed a new method for dimensionality reduction and feature extraction in face recognition, namely ULDP, which can be regarded as the integration of the local approaches (LLE) and the global approaches (PCA). ULDP can be used as a new graph construction and edge weight assignment method. The essence of this approach is to seek projection directions that preserve both local and global information. The experiments conducted on the ORL, YALE and AR face databases indicate the effectiveness of the proposed method under face recognition experimental conditions. The effectiveness of ULDP algorithm is not sensitive to the parameter of k when only six images per class were randomly selected for training on the ORL, YALE and AR face databases. For future work, we will extend ULDP to supervised and semi-supervised cases.

Acknowledgments This work is partially supported by the National Science Foundation of China under grant no. 60632050, 90820306, 60873151, 60973098 and 61005005.

References

1. Turk M, Pentland AP (1991) Face recognition using eigenfaces. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, Maui, Hawaii, June 1991, pp 586–591
2. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
3. Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
4. Ye J, Janardan R, Park C, Park H (2004) An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE Trans Pattern Anal Mach Intell 26(8):982–994
5. Belhumeur PN, Hespanha JP, Kriengman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
6. Li H, Jiang T, Zhang K (2006) Efficient and robust feature extraction by maximum margin criterion. IEEE Trans Neural Netw 17(1):1157–1165
7. Howland P, Wang J, Park H (2006) Solving the small sample size problem in face recognition using generalized discriminant analysis. Pattern Recognit 39:277–287
8. Zheng W, Zhao L, Zou C (2005) Foley–Sammon optimal discriminant vectors using kernel approach. IEEE Trans Neural Netw 16(1):1–9
9. Zheng W, Zhao L, Zou C (2004) An efficient algorithm to solve the small sample size problem for LDA. Pattern Recognit 37:1077–1079
10. Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84(405):165–175
11. Golub GH, Van Loan CF (1996) Matrix computations. Johns Hopkins University Press, Baltimore
12. Feng G, Hu D, Zhou Z (2008) A direct locality preserving projections (DLPP) algorithm for image recognition. Neural Process Lett 27:247–255
13. Swets DL, Weng J (1996) Using discriminant eigenfeatures for image retrieval. IEEE Trans Pattern Anal Mach Intell 18(8):831–836
14. Howland P, Jeon M, Park H (2003) Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. SIAM J Matrix Anal Appl 25(1):165–179

15. Ye J, Janardan R, Park CH, Park H (2004) An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE Trans Pattern Anal Mach Intell 26(8):982–994
16. Ye J, Li Q (2005) A two-stage linear discriminant analysis via QR-decomposition. IEEE Trans Pattern Anal Mach Intell 27(6):929–941
17. Chen L-F, Hong-Yuan X, Liao M, Ko M-T, Lin J-C, Yu G-J (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognit 33:1713–1726
18. Kirby M, Sirovich L (1990) Application of the KL procedure for the characterization of human faces. IEEE Trans Pattern Anal Mach Intell 12(1):103–108
19. Lee JM (1997) Riemannian manifolds: an introduction to curvature. Springer, Berlin
20. Scholkopf B, Smola A, Muller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319
21. Mika S, Ratsch G, Weston J, Scholkopf B, Smola A, Muller K-R (2003) Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. IEEE Trans Pattern Anal Mach Intell 25(5):623–628
22. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
23. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
24. Saul LK, Roweis ST (2003) Think globally, fit locally: unsupervised learning of low dimensional manifolds. J Mach Learn Res 4:119–155
25. Belkin M, Niyogi P, Dietterich TG, Becker S, Ghahramani Z (2000) Laplacian eigenmaps and spectral techniques for embedding and clustering. Adv Neural Inf Process Syst 14:873–878
26. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396
27. Zhang Z, Zha H (2004) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J Sci Comput 26(1):313–338
28. He X, Niyogi P (2003) Locality preserving projections. In: Proceedings of the 17th annual conference on neural information processing systems, Vancouver and Whistler, Canada, December 2003, pp 153–160
29. He X, Yan S, Hu Y, Niyogi P, Zhang H-J (2005) Face recognition using laplacianfaces. IEEE Trans Pattern Anal Mach Intell 27(3):328–340
30. Hu H (2008) Orthogonal neighborhood preserving discriminant analysis for face recognition. Pattern Recognit 41:2045–2054
31. Yu W, Teng X, Liu C (2006) Face recognition using discriminant locality preserving projections. Image Vis Comput 24:239–248
32. Yang L, Gong W, Gu X, Li W, Liang Y (2008) Null space discriminant locality preserving projections for face recognition. Neurocomputing 71:3644–3649
33. Lu GF, Lin Z, Jin Z (2010) Face recognition using discriminant locality preserving projections based on maximum margin criterion. Pattern Recognit 43:3572–3579
34. Yang J, Zhang D, Yang J, Niu B (2007) Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics. IEEE Trans Pattern Anal Mach Intell 29(4):650–664
35. He XF, Cai D, Yan SC, Zhang HJ (2005) Neighborhood preserving embedding. In: Proceedings of IEEE international conference on computer vision (ICCV), Beijing, China, October 2005, pp 1208–1213
36. Zeng XH, Luo SW (2007) A supervised subspace learning algorithm: supervised neighborhood preserving embedding. In: Proceedings of 3rd international conference on advanced data mining and applications, Harbin, China, August 2007, pp 81–88
37. Wang Y, Wu Y (2010) Complete neighborhood preserving embedding for face recognition. Pattern Recognit 43:1008–1015