Neural Comput & Applic (2013) 23:2231–2239
DOI 10.1007/s00521-012-1174-0

ORIGINAL ARTICLE

Local sparse representation projections for face recognition

Zhihui Lai · Yajing Li · Minghua Wan · Zhong Jin

Received: 18 November 2011 / Accepted: 11 September 2012 / Published online: 25 September 2012
© Springer-Verlag London Limited 2012

Abstract How to define the sparse affinity weight matrices is still an open problem in existing manifold learning algorithm. In this paper, we propose a novel supervised learning method called local sparse representation projections (LSRP) for linear dimensionality reduction. Differing from sparsity preserving projections (SPP) and the recent manifold learning methods such as locality preserving projections (LPP), LSRP introduces the local sparse representation information into the objective function. Although there are no labels used in the local sparse representation, it still can provide better measure coefficients and significant discriminant abilities. By combining the local interclass neighborhood relationships and sparse representation information, LSRP aims to preserve the local sparse reconstructive relationships of the data and simultaneously maximize the interclass separability. Comprehensive comparison and extensive experiments show that LSRP achieves higher recognition rates than principle component analysis, linear discriminant analysis and the state-of-the-art techniques such as LPP, SPP and maximum variance projections.

Keywords Sparse representation · Manifold learning · Dimensionality reduction · Feature extraction

Z. Lai (✉)
Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China
e-mail: lai_zhi_hui@163.com

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

Y. Li
School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

M. Wan
School of Information Engineering, Nanchang Hangkong University, Nanchang, Jianxi 330063, China

1 Introduction

Low-dimensional representation of high-dimensional data is an important problem in many application fields. The goal of dimensionality reduction is to discover the intrinsic structure from the raw data. There are many classical approaches for dimensionality reduction such as principle component analysis (PCA) [1–5], linear discriminant analysis (LDA) [3–5] and their kernelized variations [6, 7]. These kinds of techniques measure the Euclidean distance between the data points and obtain the global representation with the assumption of Gaussian distribution in the data space. These approaches are often based upon the assumption that the training data are drawn from the same underlying distribution as the test data. Unfortunately, due to the limitations in data collection and the high complexity of the data, it is usually difficult to guarantee that the training data have the desired characteristics in a statistically sufficient way. This issue becomes more prominent in high-dimensional small sample size problems.

Principle component analysis is an unsupervised method which preserves the maximal scatter of the data set. LDA is a supervised method which searches for a discriminative subspace such that patterns belonging to the same class are as close as possible while patterns belonging to different classes are as far away as possible. Because of using class information, LDA-based algorithms often perform better than PCA-based algorithms. However, both PCA and LDA take

the global Euclidean structure into account instead of the local geometry structure of original data. Recently, more and more nonlinear techniques based on manifold learning have been proposed to learn the local geometry structure. The representative spectral methods are Isomap [8], locally linear embedding (LLE) [9], local tangent space alignment [10], Laplacian Eigenmap [11], etc. These kinds of nonlinear methods aim to preserve local structures in small neighborhoods and successfully derive the intrinsic feature of nonlinear manifolds. However, they are implemented restrictedly on the training sets and cannot give explicit maps on new test data points for recognition problems. Since these nonlinear methods are only defined on the training data space, we have to use the out-of-sample extension [12] technique to deal with the test data in practical applications. In order to address the out-of-sample problem, He et al. [13] proposed a linear method named locality preserving projections (LPP) to approximate the eigenfunctions of the Laplace–Beltrami operator on the manifold, and thus, the new samples can be explicitly mapped to the learned subspace. By integrating the local neighborhood information and class label information, many methods [14–18] and some 2D/kernel variants were proposed and can achieve good performance. In fact, these supervised manifold learning methods use the same graph embedding framework, that is, unnormalized graph Laplacian [19], for feature extraction. However, how to define the sparse affinity weight matrices is still an open problem.

A new hot issue of the state-of-the-art technique for classification is sparse representation developed in recent years. In [20, 21], the discriminative nature of sparse representation was exploited to perform classification. This representation for a fixed sample is naturally sparse, involving only a small fraction of the training samples from the same class of the sample [21]. Therefore, although sparse representation is an unsupervised learning method, it implicitly includes the discriminative nature. Thus, we can take advantage of this property and put the sparse representation into the open problem of how to define the affine weighted matrix in most existing manifold learning methods. Based on this idea, Qiao et al. [22] extended neighborhood preserving embedding (NPE) [23] and proposed sparsity preserving projection (SPP) to avoid selecting the neighborhood parameters. However, there are two main drawbacks in SPP. On the one hand, when a training sample is sparsely represented by the remained training set, the sparse representation progresses are time-consuming, particularly when there are a large number of high-dimensional training samples. On the other hand, since SPP only focuses on unsupervised learning, neglecting the label information will degrade the performances.

Motivated by the sparse representation and manifold learning, we propose a novel method called local sparse representation projections (LSRP) for linear dimensionality reduction. Differing from the recent manifold learning methods such as LPP and SPP, LSRP introduces the locally sparse representation information into the objective function instead of globally sparse representation in SPP. The idea of the proposed method is that the sparse weighted matrix is constructed from local sparse representation coefficients instead of other forms such as binary pattern or Gaussian kernel. By integrating the local interclass relationship and sparse representation information, the proposed method aims to preserve the sparse reconstructive relationship of the data and simultaneously maximize the interclass separability. In the literature, besides NPE, LPP and SPP, a supervised method named maximum variance projections (MVP) [17] is the most closest to the proposed method. The different points are that MVP uses L_2-norm to reconstruct within-class samples and locality is not introduced in between-class seperability.

The rest of the paper is organized as follows. In Sect. 2, LPP and SPP are briefly reviewed. LSRP algorithm is proposed in Sect. 3. In Sect. 4, experiments are carried out to evaluate the proposed LSRP algorithm. Finally, the conclusions are given in Sect. 5.

2 A brief review of LPP and SPP

2.1 Locality preserving projections

Let matrix $X = [x_1, x_2, \ldots, x_N]$ be the data matrix including all the training samples $\{x_i\}_{i=1}^{N} \in R^m$ in its columns. In practice, the feature dimension m is often very high. The goal of linear dimensionality reduction is to transform the data from the original high-dimensional space to a low-dimension one, that is, $y = A^T x \in R^d$ for any $x \in R^m$ with $d << m$, where $A = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ and $\alpha_i (i = 1, \ldots, d)$ is an m-dimension column vector.

Locality preserving projections aims to preserve the local geometric structure of the data set. The objective function of LPP is defined as follows:

$$\text{Min} \frac{1}{2} \sum_i \sum_j W_{ij} \|y_i - y_j\|^2 = \text{Min}\, \text{tr}(A^T X (D - W) X^T A) \tag{1}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix and $y_i = A^T x_i (i = 1, \ldots, N)$, $D_{ii} = \sum_j W_{ij}$ and the affinity weight matrix W is defined as

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / t), & \text{if } x_i \in N_K(x_j) \quad \text{or } x_j \in N_K(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $N_K(x_i)$ denotes the K nearest neighbors of x_i.

Minimizing Eq. (1) means that if two points are close to each other in the original space, then they should be kept close in the low-dimensional transformed space. By imposing a constraint $A^T XDX^T A = 1$, the optimal projections of LPP are given by the minimum eigenvalue solution to the following generalized eigenvalue problem:

$$X(D - W)X^T\alpha = \lambda XDX^T\alpha \tag{3}$$

where α is a column vector of A. Thus, the optimal transformation matrix A_{LPP} is composed by the eigenvectors corresponding to the minimum eigenvalue solutions of Eq. (3). It is obvious that LPP is effective in discovering the local geometric structure of the underlying manifold.

2.2 Sparsity preserving projections

SPP first seeks a sparse reconstructive weight vector s_i for each x_i through the following l_1 minimization problem:

$$\begin{aligned} &\mathrm{Min}\|s_i\|_1 \\ &s.t. x_i = Xs_i \\ &1 = e^T s_i \end{aligned} \tag{4}$$

where $s_i = [s_{i,1}, \ldots, s_{i,i-1}, 0, s_{i,i+1}, \ldots, s_{i,N}]^T$ is an N-dimensional vector in which the ith element is equal to zero (implying that the x_i is removed from X), and the elements $s_{i,j}(j \neq i)$ denote the contribution of each x_j to reconstruct x_i; e is a N-dimensional vector of all ones. The optimization problem of Eq. (4) can be solved by using l_1-magic [24]. Then, the optimal solution, denoted as $.\tilde{s}_i.$, is used to construct the following objective function which aims to preserve the optimal weight vector $\tilde{s}_i$.

$$\begin{aligned} \mathrm{Min}\sum_{i=1}^{N} \left\|\alpha^T x_i - \alpha^T X\tilde{s}_i\right\|^2 = \mathrm{Min}\alpha^T X(I - S - S^T \\ + S^T S)X^T\alpha \end{aligned} \tag{5}$$

where $S = [\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_N]$. The optimal projections, called SPPs, are the eigenvectors of the following generalized eigenvalue problem:

$$X(I - S - S^T + S^T S)X^T\alpha = \lambda XX^T\alpha \tag{6}$$

An advantage of SPP is that the affinity weight matrix of the data set can be automatically given by sparse representation. However, when a training sample is sparsely represented by the whole training set (excluded the training sample itself) with polynomial time by standard linear programming methods [25], the sparse representation procedures are time-consuming, particularly when there are a large number of high-dimensional training samples. Note that, in order to obtain the sparse solution in Eq. (4), the sparse representation should be operated on a lower PCA subspace because the number of samples N is usually less than feature dimension m.

Since all the samples except for the represented points itself are use for sparse representation, we call them as global sparse representation in this paper. In the following Sect. 3.1, local sparse representation is proposed to accelerate the computational speed in the procedures of sparse representation.

3 Local sparse representation projections

In LLE [9] algorithm, each data point is represented by its neighbors in least square error sense, where sparseness is not imposed in the reconstruction coefficients. Traditional sparse representation [20–22] is to sparsely represent samples by the whole training set. With the inspirations of LLE and traditional sparse representation, local sparse representation is introduced in the proposed algorithm.

3.1 Local sparse representation

Assume that samples belonging to the same class are resided on a sub-manifold and samples in different classes are distributed on different sub-manifolds. Therefore, each training sample can be sparsely represented with its K nearest neighbors instead of the whole remained training data set without the represented sample itself. Thus, computational speed of the local sparse representation will be faster than that of the global sparse representation since the number of the nearest samples is much less than the number of the whole training data set, that is, $K << N$. Local sparse representation is to use the following objective function to obtain the optimal local sparse representation coefficients:

$$\begin{aligned} &\mathrm{Min}\|\hat{s}_i\|_1 \\ &s.t. x_i = \hat{X}_i\hat{s}_i \\ &\|\hat{s}_i\|_2 = 1 \end{aligned} \tag{7}$$

where $\hat{X}_i$ only includes the local K nearest neighbors of x_i. Denote the optimal local sparse representation coefficients, that is, the optimal solution of (7), as K-dimensional vector $s_i^* = [s_{ii_1}^*, \ldots, s_{ii_K}^*]$, where $s_{ii_p}^*$ denotes the contribution of the pth nearest neighbor x_{ii_p} of x_i to construct x_i and $\{i_1, \ldots, i_K\}$ denotes the index set of K nearest neighbors of x_i. Let $\bar{s}_i = [\bar{s}_{i1}, \bar{s}_{i2}, \ldots, \bar{s}_{iN}]^T$ be an N-dimensional vector and the elements in it are defined as

$$\bar{s}_{ij} = \begin{cases} s_{ij}^*, & \text{if } j \in \{i_1, \ldots, i_K\} \\ 0, & \text{otherwise} \end{cases}.$$

Following the same way as SPP, we aim to minimize the following local sparse reconstruction error:

$$J_s(A) = \sum_{i=1}^{N} \left\| \alpha^T x_i - \alpha^T X \bar{s}_i \right\|_2^2 = \alpha^T X (I - \bar{S} - \bar{S}^T + \bar{S}^T \bar{S}) X^T \alpha \quad (8)$$

where $\bar{S} = [\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_N]$ is the local sparse representation matrix.

It is clear that only local nearest neighbors of a data point are attended in the sparse representation in Eq. (7) in our proposed method. However, in SPP, all the remained training data except the represented data point itself are used for sparse representation. This is the difference between our method and the approach used in SPP. Thus, the nonzero coefficients in our proposed method are strictly located on local nearest neighbors of the represented points.

Since most of the nonzero coefficients in $\bar{s}_i$ are associated with the same class of the represented samples, we take advantage of this property and use $I - \bar{S} - \bar{S}^T + \bar{S}^T \bar{S}$ to characterize the local affinity weight matrix in our algorithm. Obviously, minimizing $J_s(A)$, that is, minimizing the local sparse reconstruction error, will preserve the local sparse representation relationship of the data.

Note that, similar to SPP, in the step of local sparse representation, the samples should be projected into a lower PCA subspace since the dimension m of the samples must be smaller than K in order to obtain the sparse solution in Eq. (7). When $K = N - 1$, local sparse representation becomes global/traditional sparse representation in SPP. Thus, SPP is a special case of LSRP.

3.2 Characterization of the local interclass separability

Focusing on manifold learning and pattern classification, our proposed method is expected to achieve good discriminating performance by integrating the neighborhood information and class relations among data points. Since there are many sub-manifolds in the high-dimensional space, in order to distinguish one sub-manifold from the others, labels are directly used to construct an interclass similarity matrix H. Here H is defined as follows:

$$H_{ij} = \begin{cases} 1, & \text{if } x_i \in N_K^-(x_j) \quad \text{or } x_j \in N_K^-(x_i) \\ 0, & \text{otherwise} \end{cases}. \quad (9)$$

where $N_K^-(x_i)$ indicates the K nearest neighbors of the sample x_i but with different labels. Suppose that we get the low-dimensional training samples $y_1, y_2, \ldots, y_N$, then the local between-class separability can be defined as the following equation:

$$\begin{aligned} J_b(A) &= \frac{1}{2} \sum_i \sum_j H_{ij} \left\| y_i - y_j \right\|^2 \\ &= \frac{1}{2} \sum_i \sum_j H_{ij} \left\| \alpha^T x_i - \alpha^T x_j \right\|^2 \\ &= \alpha^T X (\bar{D} - H) X^T \alpha = \alpha^T X L X^T \alpha \end{aligned} \quad (10)$$

where L is Laplacian matrix, $L = \bar{D} - H$, $\bar{D}_{ii} = \sum_j H_{ij}$.

XLX^T characterizes the separability of the data set in different classes, that is, in different sub-manifolds. Maximizing $J_b(A)$ means that samples in different classes are separated as far as possible, which is similar to LDA maximizing the between-class scatter.

3.3 The novel objective function

The goal of our proposed algorithm is to separate different sub-manifolds as far as possible and preserve the local sparse representation relationship of the data set. From Eqs. (8) and (10), a novel objective functions are given as follows:

$$\text{Maximize } J_b(A) = \alpha^T X L X^T \alpha \quad (11)$$

$$\text{Subject to } J_s(A) = \alpha^T X (I - \bar{S} - \bar{S}^T + \bar{S}^T \bar{S}) X^T \alpha = 1 \quad (12)$$

This constrained optimization problem can be figured out by enforcing Lagrange multiplier. Then, the optimal projections are given by the maximum eigenvalue solution to the following generalized eigenvalue problem:

$$XLX^T \alpha = X (I - \bar{S} - \bar{S}^T + \bar{S}^T \bar{S}) X^T \alpha \quad (13)$$

From Eq. (13), it can be found that the optimal projection matrix $A_{\text{LSRP}} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is composed of the eigenvectors associated with the d largest eigenvalues by solving the generalized eigenequation of Eq. (13). The optimal projection $A_{\text{LSRP}} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is called LSRP.

For further understanding the constrained optimization problem, Eqs. (11) and (12) can be equivalently rewritten as maximizing the following function:

$$J(a) = \frac{\alpha^T X L X^T \alpha}{\alpha^T X (I - \bar{S} - \bar{S}^T + \bar{S}^T \bar{S}) X^T \alpha} \quad (14)$$

It is known that the criterion of LDA is to maximize the ratio of the between-class scatter to the within-class scatter. From the Eq. (11), we find that, similar to LDA, the criterion of LSRP is to maximize the ratio of local between-class separability to sparse representation errors. The criterion $J(a)$ indicates that we can find the projections/maps by simultaneously maximizing the local between-class separability and minimizing sparse representation errors. Since most of the nonzero coefficients are associated with the same class of the represented sample, in the supervised learning point of view, minimizing $X(I - \bar{S} - \bar{S}^T + \bar{S}^T \bar{S}) X^T$ implicitly characterizes the local within-class compactness, and $I - \bar{S} - \bar{S}^T + \bar{S}^T \bar{S}$ can be viewed as within-class affinity weight matrix which is different from the ones existing in current manifold learning algorithm. Therefore, our method implicates the idea of LDA. This is also the primary motivation of our proposed method.

3.4 The algorithm

The LSRP algorithmic procedures can be summarized as follows:

Step 1 Project the original data into the PCA subspace to overcome the small sample size problem and constructs matrix H with Eq. (9)

Step 2 Compute the optimal sparse reconstruction coefficients on a lower PCA subspace and construct local sparse representation matrix $\bar{S}$.

Step 3 Compute the matrices XLX^T and $X(I-\bar{S}-\bar{S}^T+\bar{S}^T\bar{S})X^T$.

Step 4 Compute the optimized solutions by solving the generalized eigenvalue problem based on Eq. (13).

Step 5 Project samples to the LSRP subspace and adopt a suitable classifier for classification.

It should be noted that the matrix $X(I-\bar{S}-\bar{S}^T+\bar{S}^T\bar{S})X^T$ might be singular, which stems from the small sample size problem. In order to overcome the singularity of $X(I-\bar{S}-\bar{S}^T+\bar{S}^T\bar{S})X^T$, we first project the data set to a PCA subspace so that the resulting matrix $X(I-\bar{S}-\bar{S}^T+\bar{S}^T\bar{S})X^T$ is nonsingular. Another consideration of using PCA as preprocessing is for noise reduction. Moreover, to perform sparse representation on a lower PCA subspace can also significantly save computational time, particularly when there are a lot of high-dimensional training samples. The preprocessing must be performed when encountering the case mentioned above. Therefore, the final transformation matrix A can be expressed as follows:

$$A = A_{\mathrm{PCA}}A_{\mathrm{LSRP}} \tag{15}$$

where A_{PCA} denotes the PCA transform.

4 Experiments

To evaluate the proposed LSRP algorithm, we compare it with PCA (Eigenface), LDA (Fisherface), LPP (Laplacianface), SPP [22] and MVP [17] in ORL, Yale and extended Yale-B face databases. The ORL database is used to evaluate the performance of LSRP under conditions where the pose, face expression and sample size vary. The Yale database was used to examine the performance when both facial expressions and illumination were varied. The Yale-B face database was employed to test the performance under conditions where there were large variations in facial expressions and lighting conditions. Nearest neighbor classifier with Euclidean distance is used in all the experiments. The experiments are completed in Matlab 7.0 on a platform of Pentium 4 3.20 GHz CPU and 1.5G memory.

4.1 Experiments on ORL face database

The ORL face database (http://www.uk.research.tt.com/facedatabase.html) is used to evaluate the performance of LSRP 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 degrees. Moreover, there is also some variation in the scale of up to about 10 %. All images are normalized to a resolution of 56×46. Sample images of one person are shown in Fig. 1.

In the experiment, T images (T varies from 2 to 5) are randomly selected from the image gallery of each individual to form the training sample set. The remaining 10 T images are used for test. For each T, experiments were repeated 50 times. PCA, LDA, LPP, SPP, MVP and LSRP are, respectively, used for feature extraction. Note that LDA, LPP and LSRP all involve a PCA phase. For fair comparisons, in the PCA phase of LDA, LPP, SPP and LSRP, the number of principle components is set as 50. The neighbor parameter K in the locality-based methods is varied from 3 to $N-1$ with step 3 to search for the best performance, where N is the number of training samples. The maximal average recognition rates of each method and the corresponding dimension are given in Table 1. From Table 1, it can be found that LSRP obtains the higher recognition rates in all cases.

Table 1 also shows that compared with PCA and LDA which attempt to preserve the global Euclidean structure, locality-based method such as LPP and LSRP can achieve higher recognition rates. LSRP is superior to SPP and MVP since LSRP directly characterizes local interclass seperability and sparse representation relationships of the local nearest neighbors. SPP is not superior to LDA when there are only 2 and 3 training samples per person. But, when there are more training samples, SPP are superior to PCA, LDA and LPP.

Focusing on the supervised manifold learning algorithm MVP and LSRP, we find that both of them characterize the interclass seperability and reconstruction relationships by using within-class sample points. Within-class samples with L_2-norm reconstruction are used in MVP for representation. However, local nearest samples with L_1-norm minimization are used in LSRP for representation, which is significantly different from MVP. Different representations result in different recognition rates. The facts that the recognition rates obtained by LSRP are higher than the

Fig. 1 Sample images of one person on the ORL face database

Table 1 The maximal average recognition rates (percent) of six methods on the ORL database and the corresponding dimensions (shown in parentheses) when the 2, 3, 4 and 5 samples per class are randomly selected for training and the remaining for test

#/class	2	3	4	5
PCA	74.91 (50)	82.23 (46)	84.53 (34)	86.71 (46)
LDA	77.40 (39)	85.09 (39)	86.17 (39)	87.23 (35)
LPP	72.05 (48)	81.78 (46)	87.42 (36)	90.82 (34)
SPP	73.04 (50)	82.69 (50)	88.04 (50)	91.53 (50)
MVP	76.93 (44)	84.70 (43)	88.38 (40)	91.48 (37)
LSRP	80.48 (50)	88.46 (50)	92.94 (50)	95.34 (50)

ones of MVP indicate that LSRP has more discriminant abilities, which are derived from local sparse representation relationships of the data, obviously.

It is known that face images lie on a low-dimensional manifold embedded in the high-dimensional space. Due to the lack of training samples, the face images may have the property of multi-manifold structure [26–31], and thus, the images within the local neighborhood may have different labels. The local sparse representation can further explore the samples having the same labels with the represented images for representation, that is, the nonzero elements are given to the samples with the same label as the represented samples, which also provide helpful discriminative information. Therefore, preserving the reconstruction coefficients can further enhance the discriminative power. Thus, LSRP achieves best result.

4.2 Experiments on Yale database

The Yale face database (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 images of 15 individuals under various facial expressions and lighting conditions. For each individual, there are 11 images. In our experiments, each image was manually cropped and resized to 100 × 80 pixels. For computational effectiveness, we down sample it to 50 × 40 in experiments. Sample images of one person are shown in Fig. 2.

In experiments, T images (T varies from 2 to 5) were randomly selected from the image gallery of each individual to form the training sample set. The remaining 11 T images were used for test. For each T, experiments were repeated 50 times. PCA, LDA, LPP, SPP, MVP and LSRP were used for feature extraction. For fair comparisons, in the PCA phase of LDA, LPP, SPP, MVP and LSRP, the number of principle components are set as 30 (when there are two images per person for training) and 40 (when there are more than two images per person for training). Other parameters were set as in Sect. 4.1. The maximal average recognition rate of each method and the corresponding dimension are given in Table 2.

As it is shown in Table 2, the top recognition rate of LSRP is significantly higher than the other methods. Again, SPP is not superior to LDA when there are only 2, 3 and 4 training samples per person. The results are similar to the ones on ORL database. Moreover, as it is show in Table 2, SPP is not necessarily superior to LPP. However, from Tables 1 and 2, it can be found that when there are more training samples, SPP are superior to LDA and LPP.

When we focus on the recognition rates of MVP and LSRP, we can find that the recognition rates of LSRP are higher than the ones of MVP. This indicates that local sparse representation relationships of the data can provide more discriminative information than linear reconstruction with L_2-norm.

4.3 Experiments on the extended Yale-B face database

The Yale Face Database B contains 5,760 single-light-source images of 10 subjects, each under 576 viewing conditions (9 poses and 64 illumination conditions). The extended Yale Face Database B contains 16,128 images of 28 human subjects with the same condition and data format as in the previous database. We combine these two databases to include 38 subjects in total. Thus, the database contains 2,414 front-view images of 38 individuals. The images are cropped and resized to 32 × 32 pixels, with 256 gray levels per pixel. The database can be directly downloaded form http://www.cs.uiuc.edu/homes/dengcai. The feature of each image is represented by a 1,024-dimensional column vector. Large illumination variation is the property in Yale-B databases. Sample images of one person on the Yale-B face database are shown in Fig. 3.

In this experiment, T ($T = 5{:}5{:}30$) images are randomly selected from the image gallery of each individual to form the training sample set. The remaining images are used for

Fig. 2 Sample images of one person on the Yale database

Table 2 The maximal average recognition rates (percent) of six methods on the Yale database and the corresponding dimensions when 2, 3, 4 and 5 samples per class are randomly selected for training and the remaining for test

#/class	2	3	4	5
PCA	78.49 (29)	81.47 (40)	85.26 (37)	85.96 (40)
LDA	81.93 (14)	85.61 (14)	88.30 (14)	88.84 (14)
LPP	81.45 (22)	85.97 (24)	88.57 (21)	89.00 (18)
SPP	66.98 (29)	80.48 (40)	85.09 (39)	90.51 (39)
MVP	84.59 (24)	88.70 (28)	90.74 (28)	92.71 (23)
LSRP	87.97 (28)	90.15 (39)	92.69 (39)	94.31 (39)

test. For each T, the experiments were repeated 10 times. For fair comparisons, in the PCA phase of LDA, LPP, MVP and LSRP, the number of principle components is set as 150, and the remained parameters are set as in Sect. 4.1. The maximal average recognition rate of each method and the corresponding dimension are given in Table 3. As shown in Table 3, the top recognition rate of LSRP is significantly higher than the other methods. Both LSRP and SPP aim to preserve the sparse representation relationships. However, LSRP directly characterize local inter-class seperability and local sparse representation. Thus, LSRP is superior to SPP and MVP.

Experiments on ORL and Yale face database show that SPP might not be superior to LDA and LPP when there are only small numbers of training samples such as $T = 2{:}4$. However, when there are more training samples such as $T > 4$ in all experiments presented above, the recognition rates of SPP are significantly higher than LDA and LPP, which are consistent with [22]. Moreover, once again, when we compare the recognition rates of MVP and LSRP, we can draw the same conclusion that local sparse representation provides significant discriminant information.

Moreover, in order to show the high efficiency of the proposed local sparse representation, we compare the computational time in 50 dimensional PCA subspace when 30 samples of each class are used for training. l_1-magic [24] with the same parameters are used in SPP and LSRP to obtain the sparse representation coefficients. For each sample, local sparse representation with $K = 100$ in LSRP only takes 0.187 s. However, global sparse representation in SPP costs 2.469 s in sparse representation for each sample. This indicates that SPP are more time-consuming than LSRP. The reason is that the computational complexity of sparse representation is polynomial time, that is, at least $O(n^3)$ where n denotes the number of the training samples used in sparse representation. Since $K \ll N$, LSRP is more efficient than SPP which uses $N - 1$ training

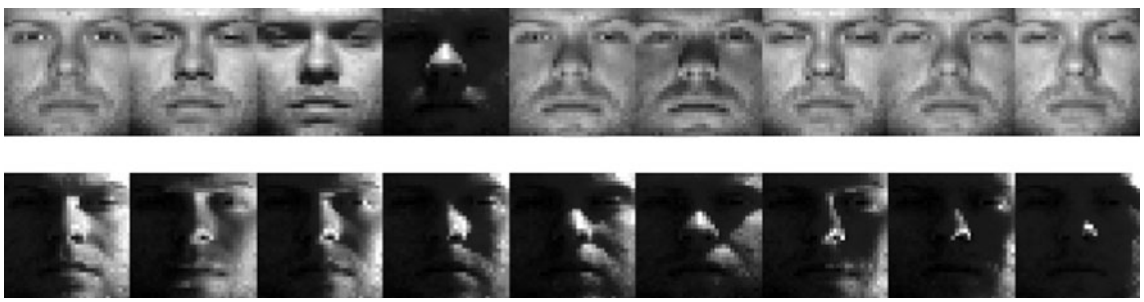

Fig. 3 Sample images of one person on the Yale-B face database

Table 3 The maximal average recognition rates (percent) of six methods on the Yale-B face database and the corresponding dimensions (shown in parentheses) when T ($T = 5{:}5{:}30$) samples per class are randomly selected for training and the remaining for test

#/class	5	10	15	20	25	30
PCA	46.77 (150)	60.05 (150)	67.60 (150)	71.81 (150)	74.95 (150)	78.34 (150)
LDA	63.84 (37)	80.08 (37)	85.37 (37)	88.20 (37)	85.19 (37)	84.18 (37)
LPP	64.17 (130)	80.20 (140)	87.65 (130)	89.96 (100)	90.87 (145)	91.42 (145)
SPP	65.68 (150)	80.90 (150)	87.15 (27)	90.72 (145)	92.61 (140)	93.10 (135)
MVP	67.91 (90)	81.93 (45)	88.30 (30)	90.98 (25)	92.13 (35)	93.35 (30)
LSRP	71.69 (150)	83.64 (150)	89.64 (150)	91.61 (150)	93.03 (100)	94.51 (125)

samples for sparse representation. Thus, introducing the locality into the sparse representation is helpful for improving the efficiency.

5 Conclusion

In this paper, we develop a supervised learning technique called LSRP for linear dimensionality reduction of high-dimensional data. But, differing from the recent manifold learning methods such as LPP, SPP and MVP, LSRP introduces the local sparse representation information into the objective function. By combining the local interclass neighborhood relationship and local sparse representation information, LSRP aims to preserve the sparse reconstructive relationship of the data and simultaneously maximize the interclass separability. Although no label information is given in constructing affinity matrix $I - \bar{S} - \bar{S}^T + \bar{S}^T\bar{S}$, the sparse representation can provide better measure coefficients which are different from the current manifold learning algorithm. Moreover, comparing with global sparse representation, local sparse representation greatly saves computational time. Our experiments also show that local sparse representation without label information provides more significant discriminant abilities than L_2-norm representation with label information used in MVP. Therefore, the recognition rates of LSRP are higher than that of LDA, LPP, SPP and MVP. The experimental results on ORL, Yale and extended Yale-B face databases show the effectiveness and efficiency of LSRP.

Acknowledgments This work is partially supported by the Natural Science Foundation of China under grant No. 61203376, 61203243, 61005005, 61005008, 61105054, Hi-Tech Research and Development Program of China under grant No. 2006AA01Z119 and China Postdoctoral Science Foundation funded project 2012M511479.

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. Joliffe I (1986) Principal component analysis. Springer, New York
3. Fukunnaga K (1991) Introduction to statistical pattern recognition, 2nd edn. Academic Press, London
4. Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
5. 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
6. Scholkopf B, Smola A, Muller KR (1998) Nonlinear component analysis as a Kernel eigenvalue problem. Neural Comput 10(5):1299–1319
7. Yang J, Frangi AF, Zhang D, Yang J-Y, Zhong J (2005) KPCA plus LDA: a complete Kernel fisher discriminant framework for feature extraction and recognition. IEEE Trans Pattern Anal Mach Intell 27(2):230–244
8. Tenenbaum JB, deSilva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
9. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
10. Zhang Z, Zha H (2004) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J Sci Comput 26(1):313–338
11. Belkin M, Niyogi P (2001) Laplacian eigenmaps and spectral techniques for embedding and clustering. Proc Adv Neural Inf Process Syst Vancouver Can 14:585–591
12. Bengio Y, Paiement JF, Vincent P, Delalleau O, Roux N, Ouimet M (2003) Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. Adv Neural Inf Process Syst 16:177–184
13. He X, Niyogi P (2003) Locality preserving projections. In: Proceedings of the seventeenth annual conference on neural information processing systems, Vancouver and Whistler, Canada
14. Chen H-T, Chang H-W, Liu T-L (2005) Local discriminant embedding and its variants. Proc IEEE Conf Comput Vis Pattern Recognit 2:846–853
15. Yan S, Xu D, Zhang B, Zhang H-J, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51
16. Fu Y, Yan S, Huang TS (2008) Classification and feature extraction by simplexization. IEEE Trans Inf Forensics Secur 3(1):91–100
17. Zhang T, Yang J, Wang H, Du C (2007) Maximum variance projections for face recognition. Opt Eng 46(6):0672061–0672068
18. Fu Y, Yan S, Huang TS (2008) Correlation metric for generalized feature extraction. IEEE Trans Pattern Anal Mach Intell 30(12):2229–2235
19. Chung F (1997) Spectral graph theory, CBMS Regional Conference Series in Mathematics, no. 92
20. Wright J, Yang A, Sastry S, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227

21. Huang K, Aviyente S (2006) Sparse representation for signal classification. Adv Neural Inf Process Syst 19:609–616
22. Lishan Q, Songcan C, Xiaoyang T (2010) Sparsity preserving projections with applications to face recognition. Pattern Recognit 43:331–341
23. He X, Cai D, Yan S, Zhang H (2005) Neighborhood preserving embedding. Proc Int Conf Comput Vis (ICCV) 2:1208–1213
24. Candes E., Romberg J (2005) l_1-magic: recovery of sparse signals via convex programming. http://www.acm.caltech.edu/l1magic/
25. Chen S, Donoho D, Sarnders M (2001) Atomic decomposition by Basis pursuit. SIAM Rev 43(1):129–159
26. Yang W, Sun C, Zhang L (2011) A multi-manifold discriminant analysis method for image feature extraction. Pattern Recognit 44(8):1649–1657
27. Zhao C, Liu C, Lai Z (2011) Multi-scale gist feature manifold for building recognition. Neurocomputing 74(17):2929–2940
28. Zhao H, Sun S, Jing Z, Yang J (2006) Local structure based supervised feature extraction. Pattern Recognit 39(8):1546–1550
29. Zhao H, Wong W (2012) Supervised optimal locality preserving projection. Pattern Recognit 45(1):1546–1550
30. Xu Y, Zhong A, Yang J, Zhang D (2010) LPP solution schemes for use with face recognition. Pattern Recognit 43(12):4165–4176
31. Lai Z, Wan M, Jin Z (2011) Locality preserving embedding for face and handwriting digital recognition. Neural Comput Appl 20(4):565–573