Neurocomputing 260 (2017) 321–330

Contents lists available at ScienceDirect

Neurocomputing

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

Sparse locality preserving discriminative projections for face recognition

Jianbo Zhang[a,b,*], Jinkuan Wang[b], Xi Cai[c]

[a] School of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[b] School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China
[c] School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China

ARTICLE INFO

Article history:
Received 16 September 2016
Revised 1 April 2017
Accepted 21 April 2017
Available online 11 May 2017

Communicated by Deng Cai

MSC:
65D18
62H35

Keywords:
Dimensionality reduction
Sparse representation
Manifold learning
Maximum margin criterion
Face recognition

ABSTRACT

Recently, the construction of intrinsic graph using sparse representation (SR) has attracted considerable interest. Comparing with the traditional construction methods like k-NN and ε-ball which can well preserve the manifold structure of samples, SR method is more robust to data noise and parameter-free. To exploit the merits of robustness of sparse representation and manifold learning, we propose a new algorithm called sparse locality preserving discriminative projections (SLPDP), which utilizes sparse representation to construct the intrinsic weighted matrix of training samples and incorporates "locality" and "sparsity" into objective function. Simultaneously, SLPDP takes into account the global information of samples like LDPD and DSNPE, and integrates maximum margin criterion (MMC) into the optimal functions for dimensionality reduction. Experiments on PIE, AR, Extended Yale B and Yale face image databases demonstrate the effectiveness of the proposed approach.

1. Introduction

Face recognition has attracted considerable attentions for the past over twenty years in computer vision and pattern recognition [1–3]. In this issue, one of the most important challenges is to project the face images in high-dimensional space into low-dimensional space before recognition due to its high costs in time and space. In order to enhance the classification performance, dimensionality reduction is usually applied to most of the face recognition tasks [4,5]. And an enormous volume of literature has been devoted to investigate various data-dependent dimensionality reduction methods for projecting the high-dimensional data into low-dimensional feature spaces such as principal component analysis (PCA) [3–6], linear discriminant analysis (LDA) [4,5], maximum margin criterion (MMC) [7,8] and so on. These dimensionality reduction methods get better performance in processing general type (real-world or non-real-world) data. However, they only capture the global Euclidean structure of images and cannot well characterize the local structure [9,10].

To address the above problem, many manifold learning algorithms, based on the idea that the data points are actually samples from a low-dimensional manifold embedded in a high-dimensional space, have been proposed for face recognition. The most representative methods about manifold learning include locally linear embedding (LLE) [9], isometric feature mapping (ISOMAP) [10], and Laplacian eigenmaps (LE) [11]. These manifold learning algorithms attempt to preserve a different geometrical property of the underlying manifold. However, they cannot be well applied for classification tasks since it is difficult to find new sample images in the embedding space by utilizing the low-dimensional embedding results of the training data set. Thus, some improved methods, such as locality preserving projections (LPP) [12], neighborhood preserving embedding (NPE) [13] and so on, have been proposed to solve this problem. Since these improved methods only keep the locality structure while ignore the global information, a dozen of local discriminant approaches, including local discriminant embedding (LDE) [14], discriminant locality preserving projections based on maximum margin criterion (DLPP/MMC) [15], locality preserving discriminant projections (LPDP) [16] and so on, have been developed for image classification. These methods integrate "global" with "locality" information of samples by different ways, among which LPDP [16] integrates the maximum margin criterion

* Corresponding author at: College of Information Science and Engineering, Northeastern University, Shenyang 110819, China.
E-mail addresses: zjb510@126.com, zjb@neuq.edu.cn (J. Zhang).

http://dx.doi.org/10.1016/j.neucom.2017.04.051
0925-2312/

(MMC) [7] to the LPP [12] objective function for dimensionality reduction and obtains better effectiveness.

In 2007, Cai et al. [17] introduced the framework of sparse subspace learning (SSL) for the first time. After that, a large number of new dimensionality reduction methods based on SSL have been proposed. The main methods with representativeness include [18–21]. In these methods, sparsity preserving projections (SPP) algorithm [19] preserves the sparse reconstructive relationship of the data by minimizing the l_1-regularization-related objective function and achieves good recognition effect. However, SPP does not consider the global information of all samples. Thus, many discriminant sparse subspace learning methods are proposed, such as discriminant sparse neighborhood preserving embedding (DSNPE) [22] which combines maximum margin criterion (MMC) [7] with SPP and obtains better performance for face recognition. Although preserving the sparse reconstruction and global structure of samples, DSNPE impairs the inherent manifold structure of training samples [23]. Similar work can be found in [24].

Sparse representation (SR) has received considerable interest in recent years. The main idea is that the given sample can be represented as a linear combination of the others. The coefficients obtained by SR reflect the contributions of the samples to reconstruct the given sample. It has been reported in [23] that the bigger coefficients are the more likely these samples belong to the same class. Therefore, the reconstruction coefficients can be considered as a similarity measurement of samples, and the adjacency graph matrix of samples can be constructed by SR method. Compared with the graph matrix constructed by k-NN or ε-ball methods, SR method has three advantages: robustness to data noise, sparsity and parameter-free [25]. Due to the above considerations, Lou et al. [26] proposed graph regularized sparsity discriminant analysis (GRSDA) which constructs the intra-class graph and inter-class graph by sparse representation. Similar works can also be found in [23,27].

Although the construction of intrinsic graph via SR is more robust to data noise and parameter-free, the traditional methods like k-NN and ε-ball can better preserve the manifold structure of samples. In this paper, to further exploit the merits of robustness of sparse representation and manifold learning, we propose a new algorithm called sparse locality preserving discriminative projections (SLPDP), which utilizes sparse representation to construct the intrinsic weighted matrix of samples and incorporates "locality" and "sparsity" into objective function for dimensionality reduction. Simultaneously, SLPDP takes into account the global information of samples like LDPD and DSNPE and integrates maximum margin criterion (MMC) into the optimal functions. Thus, SLPDP builds the discriminant sparsity manifold learning objective function for dimensionality reduction. Experiments on PIE, AR, Extended Yale B and Yale face image databases indicate the better performance than LPP, SPP, LPDP, DSNPE and GRSDA.

The remainder of the paper is organized as follows. Section 2 introduces the related works. We introduce the details of the SLPDP algorithm in Section 3. Section 4 shows the experimental results in four face image datasets. We conclude this paper in Section 5.

2. The related works

Given n samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from c classes in high-dimensional space, and each sample $\mathbf{x}_i \in \mathbf{R}^m$ is a column vector pattern of one image in sample set, the label of $\mathbf{x}_i$ is $l_i \in \{1, 2, \ldots, c\}$. The aim of linear dimensionality reduction algorithm is to find a transformation matrix $\mathbf{V}$ that can map each sample $\mathbf{x}_i$ to low-dimensional space, i.e. $\mathbf{y}_i = \mathbf{V}^T\mathbf{x}_i \in \mathbf{R}^d$ and $d \ll m$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, then $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] = \mathbf{V}^T\mathbf{X}$. Obviously, different dimensionality reduction algorithm can gain different transformation matrix.

2.1. Sparse representation

Recently, sparse representation (SR) has received a great deal of attentions, which was initially proposed as an extension of traditional signal processing methods such as Fourier and wavelet and has been successfully used in image recognition [18].

SR has compact mathematical expression. Given a signal (or an image with vector pattern) $\mathbf{x} \in \mathbf{R}^m$ and a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ containing the elements of an over-complete dictionary in its column, the goal of SR is to represent $\mathbf{x}$ using as few of $\mathbf{X}$ as possible. The objective function can be described as follows:

$$\min_{\mathbf{s}_i} \|\mathbf{s}_i\|_0, \quad \text{s.t. } \mathbf{x}_i = \mathbf{X}\mathbf{s}_i \tag{1}$$

or

$$\min_{\mathbf{s}_i} \|\mathbf{s}_i\|_0, \quad \text{s.t. } \|\mathbf{x}_i - \mathbf{X}\mathbf{s}_i\| < \varepsilon \tag{2}$$

where $\|\mathbf{s}_i\|_0$ is the l_0-norm, denotes the number of nonzero entries in the vector $\mathbf{s}_i$, and $\mathbf{s}_i = [s_{i,1}, \ldots, s_{i,i-1}, 0, s_{i,i+1}, \ldots, s_{i,n}]^T$ is an n-dimensional vector in which the ith element is equal to zero (implying that the $\mathbf{x}_i$ is removed from $\mathbf{X}$) and the other elements $s_{i,j} (j \neq i)$ denote the contribution of each $\mathbf{x}_j$ to reconstructing $\mathbf{x}_i$.

The above optimization problem cannot be solved in polynomial time since the criterion is not convex, and finding the sparsest solution is NP-hard. Fortunately, recent efforts have been made and demonstrated that the l_0-norm is equivalent to the l_1-norm optimization problem if the solution is sparse enough [28].

2.2. Locality preserving projections

As one of the representative manifold learning methods, the objective function of LPP [12] is defined as

$$\min_{\mathbf{V}} \sum\nolimits_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} \tag{3}$$

where the adjacency matrix $\mathbf{W} = \left(W_{ij}\right)_{n\times n}$ is a structural matrix. The weight W_{ij} incurs a heavy penalty when neighboring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Thus, minimizing the objective function is an attempt to ensure that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" in high-dimensional space then $\mathbf{y}_i$ and $\mathbf{y}_j$ are close as well. A possible way, called heat kernel method, to calculate $\mathbf{W}$ is as follows:

$$W_{ij} = \begin{cases} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t\right), & \|\mathbf{x}_i - \mathbf{x}_j\| < \varepsilon; \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where t is the parameter that can be determined empirically, and $\|\mathbf{x}_i - \mathbf{x}_j\| < \varepsilon$, which called ε-ball method, can also be written as k-NN form. The detailed description can be found in [12]. Obviously, the weighted matrix $\mathbf{W}$ is symmetrical.

The objective function (3) can be reduced and written as [12]:

$$\frac{1}{2}\sum\nolimits_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} = \mathbf{V}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{V} \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix and D_{ii} is row sum of $\mathbf{W}$, i.e. $D_{ii} = \sum_i W_{ij}$, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. In addition, a constraint $\mathbf{V}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{V} = \boldsymbol{I}$ is imposed to avoid the divergence of the solution and $\boldsymbol{I}$ is unit matrix.

2.3. Maximization margin criterion

Maximum margin criterion (MMC) [7,8,29] is proposed to maximize the margin between classes after dimensionality reduction. In [7], the objective function of MMC is written as

$$\max \mathrm{tr}\left(\sum\nolimits_{ij} p_i p_j \left(d\left(\mathbf{m}_i, \mathbf{m}_j\right) - s(\mathbf{m}_i) - s\left(\mathbf{m}_j\right)\right)\right) \tag{6}$$

where p_i and p_j are the prior probability of class i and class j, respectively; m_i and m_j are the mean vectors of class i and class j, respectively. Here $d(\mathbf{m}_i, \mathbf{m}_j)$, $s(\mathbf{m}_i)$ and $s(\mathbf{m}_j)$ are defined as: $d(\mathbf{m}_i, \mathbf{m}_j) = \|\mathbf{m}_i - \mathbf{m}_j\|$, $s(\mathbf{m}_i) = \mathrm{tr}(\mathbf{S}_i)$, $s(\mathbf{m}_j) = \mathrm{tr}(\mathbf{S}_j)$, where $\mathbf{S}_i$ and $\mathbf{S}_j$ are the covariance matrix of class i and class j, respectively. The optimized function (6) can be derived as follows [7]:

$$\max \mathrm{tr}(\mathbf{S}_b - \mathbf{S}_w) \tag{7}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are called the between-class scatter matrix and the within-class scatter matrix, respectively.

Because the dimension of $\mathrm{tr}(\mathbf{S}_b)$ and $\mathrm{tr}(\mathbf{S}_w)$ may be different in practical applications, it is more reasonable to add a rescaling coefficient to $\mathbf{S}_w$ in order to balance the difference in dimension. This method is called modified MMC (MMMC) [8,29]. So, the MMMC can be obtained and rewritten in the following form:

$$\max \mathrm{tr}(\mathbf{S}_b - \alpha \mathbf{S}_w) \tag{8}$$

where α is the rescaling coefficient.

3. Sparse locality preserving discriminative projections (SLPDP)

3.1. The proposed algorithm

Generally, the weight matrix can characterize data geometry (e.g. manifold) and thus plays an important role in data analysis including machine learning, such as dimensionality reduction [30] and spectral clustering [31]. However, how to establish high-quality weight matrix is still an open problem [32].

Since the traditional manifold learning methods need to manually construct adjacency graph by k-NN or ε-ball methods which use a given parameter to determine the neighborhoods for all the data and is sensitive to data noise [18]. In this paper, inspired by the observation that the most compact expression of a certain face image is generally given by the face images from the same class [18], we employ the sparse representation method, which is parameter-free and robust to data noise, to construct the intrinsic matrix.

Firstly, to construct the sparse weighted matrix like GRSDA, we modify the original objective function of SR [See Eq. (1) or Eq. (2)] as

$$\min_{\mathbf{s}_i^k} \|\mathbf{s}_i^k\|_1, \quad \text{s.t. } \mathbf{x}_i = \mathbf{X}_k \mathbf{s}_i^k, \ \text{label}(\mathbf{x}_i) = k \tag{9}$$

or

$$\min_{\mathbf{s}_i^k} \|\mathbf{s}_i^k\|_1, \quad \text{s.t. } \|\mathbf{x}_i - \mathbf{X}_k \mathbf{s}_i^k\| < \varepsilon, \ \text{label}(\mathbf{x}_i) = k \tag{10}$$

where label $(\mathbf{x}_i)$ denotes the label of $\mathbf{x}_i$, $\mathbf{s}_i^k = [s_{i,1}^k, \ldots, s_{i,i-1}^k, 0, s_{i,i+1}^k, \ldots, s_{i,n_k}^k]^T$, n_k is the number of samples whose label is k, and $n = \sum_{k=1}^{c} n_k$.

After repeating the above optimization problem to the all data samples, the sparse weight matrix $\mathbf{S}$ can be expressed as:

$$\mathbf{S} = (S_{ij})_{n \times n} = \mathrm{diag}(\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^c) = \begin{bmatrix} \mathbf{S}^1 & & \\ & \mathbf{S}^2 & \mathbf{0} \\ & & \ddots \\ \mathbf{0} & & \mathbf{S}^c \end{bmatrix} \tag{11}$$

where $\mathbf{S}^k = [\mathbf{s}_1^k, \mathbf{s}_2^k, \ldots, \mathbf{s}_{n_k}^k]$, $k = 1, 2, \ldots, c$.

An interpretation for the element S_{ij} is that sample x_i's contribution to the reconstruction of sample x_j. if $S_{ij} = 0$, the contribution from x_i to x_j is zero. As we can see $S_{ii} = 0$ by definition of $\mathbf{S}$, there is no contribution from a sample to itself. In general, the levels of reconstruction contribution between a pair of samples are different. In other words, S_{ij} and S_{ji} are typically not equal, implying that weighted matrix $\mathbf{S}$ is usually asymmetrical.

Secondly, we hope that samples from the same class in original space can retain the "locality" and "sparsity" information in low-dimensional space. Therefore, we construct the objective function as follows:

$$\min_{\mathbf{V}} \frac{1}{2} \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 S_{ij} = \min_{\mathbf{V}} \frac{1}{2} \sum_{ij} \|\mathbf{V}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{x}_j\|^2 S_{ij}. \tag{12}$$

Eq. (12) can be simplified as:

$$\min_{\mathbf{V}} \mathrm{tr}(\mathbf{V}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{V}) \tag{13}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{T}$ is Laplacian matrix, $\mathbf{T} = (\mathbf{S} + \mathbf{S}^T)/2$, $\mathbf{D}$ is a diagonal matrix and $D_{ii} = \sum_j (S_{ij} + S_{ji})/2$. From Eq. (13), We can see that the Laplacian matrix $\mathbf{L}$ in SLPDP is similar to that in LPP [See Eq. (5)]. However, the weighted matrix $\mathbf{S}$ in SLPDP is different to the weighted matrix $\mathbf{W}$ in LPP, since $\mathbf{W}$ is always symmetrical while $\mathbf{S}$ may be symmetrical or not.

To avoid the divergence of the solution, a constraint $\mathbf{V}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{V} = \boldsymbol{I}$ is imposed on Eq. (13).

Thirdly, if matrix $\mathbf{V}$ is the optimal transformation matrix which can project a pattern closer to patterns in the same class but farther from those in different classes, which is exactly the goal for classification. Then, to find an optimal linear subspace for classification means to maximize the following optimized function:

$$\max_{\mathbf{V}} \mathrm{tr}(\mathbf{V}^T (\mathbf{S}_b - \alpha \mathbf{S}_w) \mathbf{V}). \tag{14}$$

Finally, like LPDP [16] and DSNPE [22], by combining Eqs. (13) and (14), we can get the following optimization problem:

$$\begin{cases} \min_{\mathbf{V}} \mathrm{tr}(\mathbf{V}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{V}); \\ \max_{\mathbf{V}} \mathrm{tr}(\mathbf{V}^T (\mathbf{S}_b - \alpha \mathbf{S}_w) \mathbf{V}); \\ \quad \text{s. t. } \mathbf{V}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{V} = \boldsymbol{I}. \end{cases} \tag{15}$$

The solution to the constrained multi-object optimization problem (15) is to find a subspace which preserves the sparsity locality property and maximizes the margin between different classes simultaneously. So, we can change Eq. (15) into the following constrained problem:

$$\begin{aligned} &\min_{\mathbf{V}} \mathrm{tr}(\mathbf{V}^T (\mathbf{X} \mathbf{L} \mathbf{X}^T - \gamma (\mathbf{S}_b - \alpha \mathbf{S}_w)) \mathbf{V}), \\ &\text{s. t. } \mathbf{V}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{V} = \boldsymbol{I} \end{aligned} \tag{16}$$

where γ is a parameter to balance the sparsity locality and the discriminant information.

Eq. (16) can be solved by Lagrangian multiplier method:

$$\frac{\partial}{\partial \mathbf{V}} \mathrm{tr}(\mathbf{V}^T (\mathbf{X} \mathbf{L} \mathbf{X}^T - \gamma (\mathbf{S}_b - \alpha \mathbf{S}_w)) \mathbf{V}) - \lambda_i (\mathbf{V}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{V} - \boldsymbol{I}) = 0$$

where λ_i is the Lagrangian multiplier. Then, we can get

$$(\mathbf{X} \mathbf{L} \mathbf{X}^T - \gamma (\mathbf{S}_b - \alpha \mathbf{S}_w)) \mathbf{v}_i = \lambda_i \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{v}_i \tag{17}$$

where $\mathbf{v}_i$ is the generalized eigenvector of $\mathbf{X} \mathbf{L} \mathbf{X}^T - \gamma (\mathbf{S}_b - \alpha \mathbf{S}_w)$ and $\mathbf{X} \mathbf{D} \mathbf{X}^T$; λ_i is the corresponding eigenvalue.

Let the column vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ be the solutions of Eq. (17), ordered according to their first d smallest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$. Thus, the embedding is written as follows:

$$\mathbf{x}_i \to \mathbf{y}_i = \mathbf{V}^T \mathbf{x}_i, \ \mathbf{V}^T = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d] \tag{18}$$

where $\mathbf{y}_i$ is a d-dimensional vector and $\mathbf{V}$ is a $m \times d$ matrix.

The whole procedure of performing classification by SLPDP can be formally summarized as follows:

Input: Training data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ and the corresponding labels $\mathbf{C} = [l_1, l_2, \ldots, l_n]$; Testing data $\mathbf{X}^*$.

Step 1: Project the image set $\mathbf{X}$ into the PCA subspace by throwing away the smallest principal components.

Step 2: Obtain the optimizer $\mathbf{s}_i^k$ to problem (9) or (10) and calculate the weight matrix $\mathbf{S}$ using Eq. (11).
Step 3: Compute the optimization projection matrix $\mathbf{V}$ using Eq. (17).
Step 4: Calculate the low-dimensional embedding $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$ for training data by Eq. (18).
Step 5: Calculate the low-dimensional embedding $\mathbf{Y}^*$ for testing data by Eq. (18).
Step 6: According to the minimization distance to classify the testing data.

3.2. Connection to LPDP, DSNPE and GRSDA

LPDP combines the global discriminant information with local structure of samples by introducing MMC into the objective function of LPP. While DSNPE combines the global discriminant information with sparse structure of samples by introducing MMC into the objective function of SPP. The proposed algorithm combines the global discriminant information with the inter-class sparse structure of data by adding MMC into the objective function (13). So, from a structural point of view, LPDP, DSNPE and SLPDP have the similar objective function.

LPDP can preserve the locality information of samples, but its adjacent weight is calculated by using the k-nearest neighbor or the ε-ball method to measure the neighborhood size which is non-parameter-free. Although the calculation of DSNPE's weight matrix is parameter-free by employing the sparse representation, but DSNPE emphasizes the sparse reconstruction information of samples and ignores the locality information of data. The proposed algorithm preserves the sparse locality embedding using sparse representation and is parameter-free. So clearly, SLPDP is the extension of LPDP and DSNPE.

GRSDA minimizes the intra-class sparse scatter and maximizes the inter-class sparse scatter. Structurally, the objective function of GRSDA is similar to that of linear discriminant analysis (LDA) [4,5]. In SLPDP, we think the reconstruction coefficients can be seen as a measurement of similarity, in other word, if the coefficient of one sample to reconstruct another is larger, they are more likely from the same class. So, the first portion [See Eq. (12)] of the objective function in SLPDP shares the similar idea with GRSDA. As the second portion [See Eq. (14)] of the objective function, SLPDP maximizes the average margin between classes, which is different from GRSDA.

4. Experimental results

In this section, the proposed approach SLPDP is experimented using four image databases: PIE, AR, Extended Yale B and Yale. In order to evaluate the performance of our approach, we compare SLPDP with several representative dimensional reduction methods such as LPP [12], LPDP [16], SPP [19], DSNPE [22] and GRSDA [26].

For simplicity of representation, the experiments were named as p-train, which means that p images per individual were selected for training and the remaining images for test. To robustly evaluate the performance of different algorithms in different training and testing conditions, we selected images randomly and repeated the experiment 20 times in each condition and exhibit the results in the form of average recognition rate with standard deviation.

At the beginning of each experiment, we set the parameters of some methods. For LPP and LPDP, the number of nearest neighbors k is taken to be $p-1$ as done in [33]. For SLPDP algorithms, we set α is 1 like LPDP [16] and DSNPE [22] and thus we can compare the proposed method with LPDP and DSNPE in the same condition. As for γ, we set it different values in different databases. To make some matrices nonsingular, PCA is employed as a preprocessing step and 98% of the energy of images is retained. At the learning stage, we solve the objective function (9) by L1-Ls-matlab packages (http://www.eecs.berkeley.edu/~yang/software/l1benchmark/). At the stage of classification, for simplicity of implementation, we employ the nearest neighbor classifier to classify the testing images as the class of training images.

4.1. Experiment using the PIE database

The CMU-PIE [34] database contains 68 subjects with 41368 face images as a whole. The face images are captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. We select pose-29 images as gallery, including 24 samples for each individual in the experiments. All images are cropped and resized to the resolution of 32× 32 pixels, and the gray values of all images are rescaled to [0, 1]. For each person, we randomly select p (from 7 to 10) images for training, the rest for testing, and we repeat the experiment 20 times in each condition. Fig. 1 shows the images of one person. In general, the recognition rates vary with the dimension of the face subspace. Fig. 2 plots the recognition accuracy vs. number of project vectors for our experiment. Table 1 lists the average recognition accuracy of 6 approaches and the corresponding standard deviation and dimensionality number on the PIE database. In experiments, we set parameter γ is 0.1.

4.2. Experiment using the AR database

The AR database [35] database contains over 4000 color face images of 126 people (56 women and 70 men) including frontal views of faces with different facial expressions, lighting conditions and occlusions. The pictures of most persons are taken in two sessions (separated by two weeks). Each session contains 13 color images and 120 individuals (65 men and 55 women) participate in both sessions. In our experiments, we use a subset of the AR face database provided and preprocessed by Martinez [35] which contains 1400 face images corresponding to 100 persons (50 men and 50 women) where each person has 14 different images. The facial portion of each image is manually cropped and then normalized to the size of 44 × 32, and the gray values are rescaled to [0, 1]. For each individual, we randomly selected p (from 5 to 8) images for training and the remaining images for testing and repeated the experiment 20 times. Fig. 3 shows all images of the first person in AR database. Fig. 4 plots the recognition accuracy vs. number of project vectors for one experiment. Table 2 lists the average recognition accuracy of 6 approaches and the corresponding standard deviation and dimensionality number on the AR database. In experiments, we set parameter γ is 0.1.

4.3. Experiment using the Extended Yale B database

The Extended Yale B database [36] database contains 2414 front-view face images of 38 individuals. For each individual, about 64 pictures were taken under various laboratory-controlled lighting conditions. In our experiments, images of 31 individuals are selected as gallery, and we use the cropped images with the resolution of 32× 32, and the gray values of all images are rescaled to [0, 1]. For each individual, we randomly selected p (from 8 to 11) images for training and the remaining images for testing and repeat the experiment for 20 times. Fig. 5 shows the images of one person. Fig. 6 plots the recognition accuracy vs. number of project vectors for one experiment. Table 3 lists the average recognition accuracy of 6 approaches and the corresponding standard deviation and dimensionality number on the Extended Yale B database. In experiments, we set parameter γ is 0.55.

Fig. 1. Sample images of one person in the PIE database.

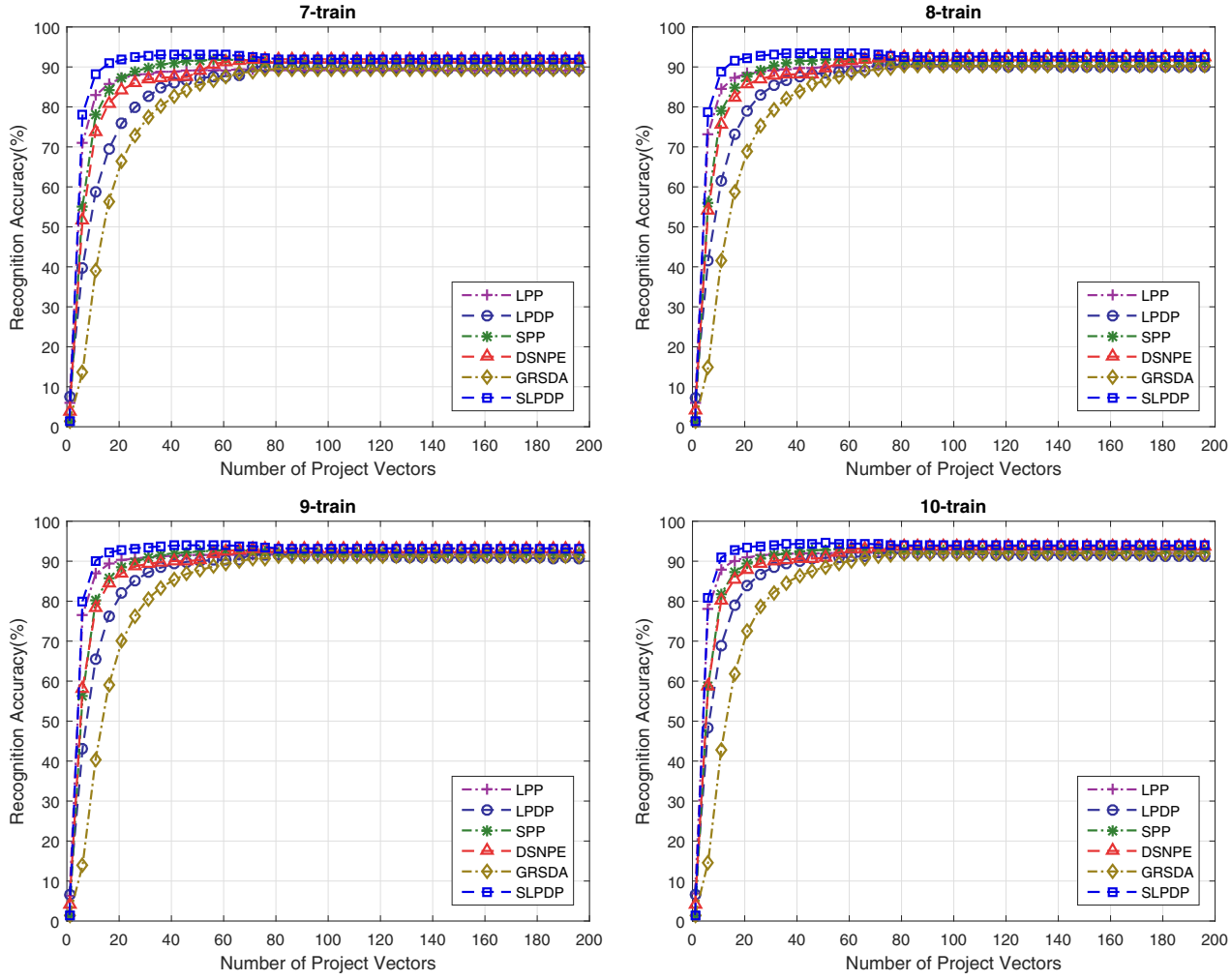


Fig. 2. Recognition accuracy vs. number of project vectors on PIE database with 7, 8, 9, 10 images for each individual randomly selected for training.

Table 1
The maximal average recognition accuracy rates (%) of six approaches across 20 runs on the PIE database and the corresponding standard deviations (std) and dimensionality (shown in parentheses).

Methods	7-train	8-train	9-train	10-train
LPP	89.43 ± 0.012(69)	90.26 ± 0.009(74)	91.72 ± 0.009(76)	92.18 ± 0.007(71)
LPDP	90.32 ± 0.011(67)	91.40 ± 0.007(67)	92.29 ± 0.008(67)	92.83 ± 0.009(67)
SPP	91.86 ± 0.009(62)	91.98 ± 0.007(65)	92.76 ± 0.007(61)	93.17 ± 0.006(65)
DSNPE	91.95 ± 0.006(78)	92.56 ± 0.006(81)	93.20 ± 0.007(82)	93.68 ± 0.006(83)
GRSDA	89.55 ± 0.009(80)	90.19 ± 0.008(81)	91.16 ± 0.009(83)	92.01 ± 0.007(83)
SLPDP	93.06 ± 0.006(43)	93.55 ± 0.008(52)	94.02 ± 0.006(49)	94.53 ± 0.007(52)

Fig. 3. Sample images of one person in the AR database.

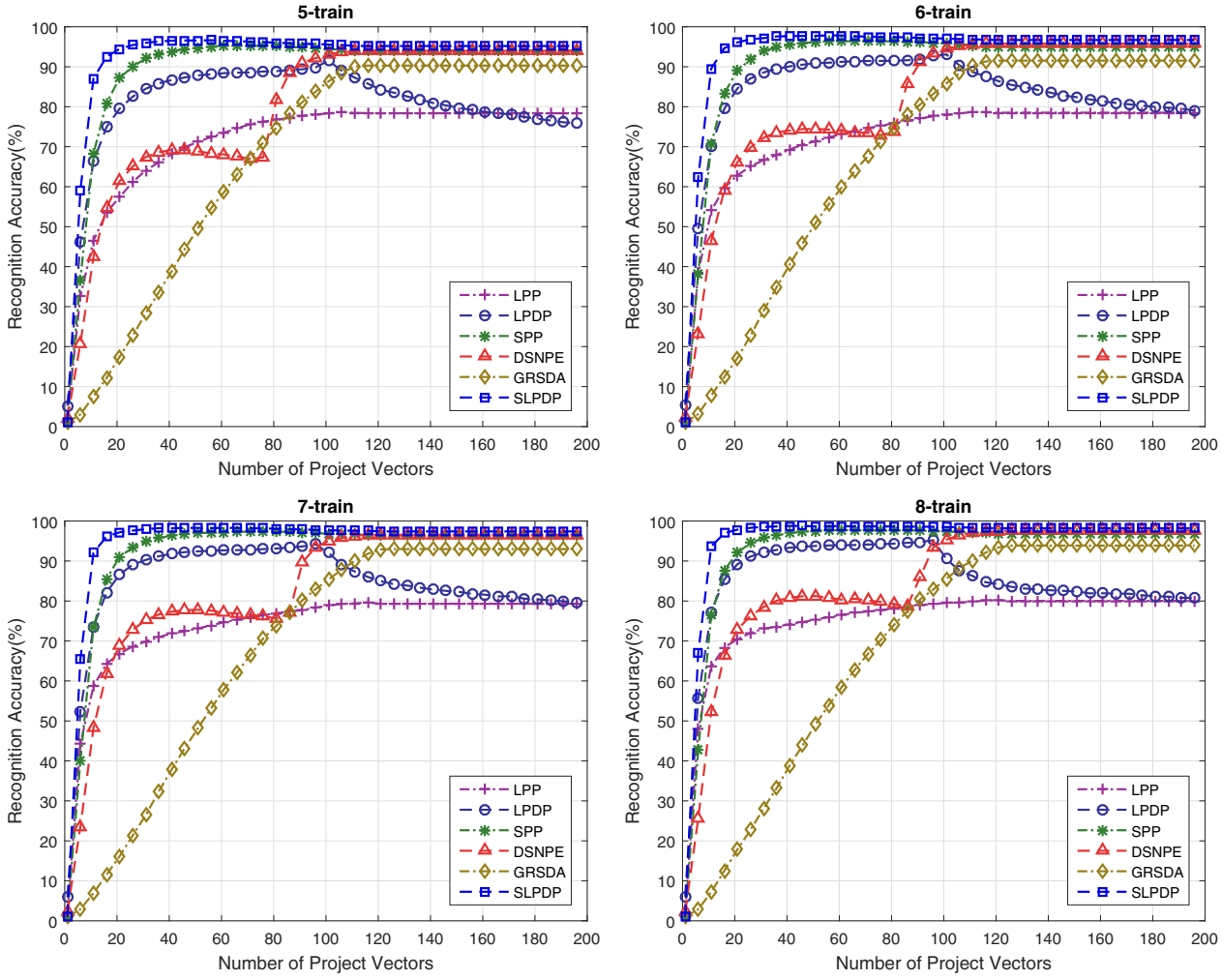


Fig. 4. Recognition accuracy vs. number of project vectors on AR database with 5, 6, 7, 8 images for each individual randomly selected for training.

Table 2
The maximal average recognition accuracy rates (%) of six approaches across 20 runs on the AR database and the corresponding standard deviations (std) and dimensionality (shown in parentheses).

Methods	5-train	6-train	7-train	8-train
LPP	78.72 ± 0.034(109)	78.88 ± 0.034(114)	79.66 ± 0.023(117)	80.21 ± 0.020(118)
LPDP	92.74 ± 0.011(99)	94.66 ± 0.008(99)	95.64 ± 0.006(99)	96.19 ± 0.009(99)
SPP	95.37 ± 0.006(64)	96.49 ± 0.005(68)	97.43 ± 0.006(73)	97.81 ± 0.006(63)
DSNPE	93.97 ± 0.009(114)	95.81 ± 0.008(116)	96.59 ± 0.009(122)	97.60 ± 0.005(125)
GRSDA	90.30 ± 0.010(114)	91.53 ± 0.014(118)	92.99 ± 0.012(123)	93.87 ± 0.015(126)
SLPDP	96.70 ± 0.007(54)	97.85 ± 0.006(57)	98.36 ± 0.005(54)	98.78 ± 0.005(46)

4.4. Experiment using the Yale database

The Yale face database (http://cvc.yale.edu/projects/~yalefaces/yalefaces.html) was constructed at the Yale Center for Computation Vision and Control. There are 165 images of 15 individuals (each person providing 11 different images). The images demonstrate variations in lighting condition (left-light, center-light and right-light), facial expression (normal, happy, sad, sleepy, surprised, and wink), and with or without glasses.In our experiments, all images are cropped and resized to the resolution of 32 × 32 pixels, and the gray values of all images are rescaled to [0, 1]. For each individual, we randomly select p (from 3 to 6) images for training and the remaining images for testing, and repeat the experiment for 20 times. Fig. 7 shows all images of the first individual. Fig. 8 plots the recognition accuracy vs. number of project vectors for one experiment. Table 4 lists the average recognition accuracy of 8 approaches and the corresponding standard deviation and dimensionality number on the Yale database. In experiments, we set parameter γ is 0.1.

From Tables 1 to 4 and Figs. 2, 4, 6 and 8, we can see that

(1) LDPD is better than LPP in our experiments. This is probably because that LPDP preserves both global discriminant and local structure of data while LPP only preserves the local structure of the samples.
(2) SPP is superior to LPDP. This is probably because that, compare with the graph constructed by k-nearest neighbor or ε-ball, l_1-graph can well characterize the intrinsic structure of images and impair the very important local discriminant structure, which is embedded in among nearby data having different class label.

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

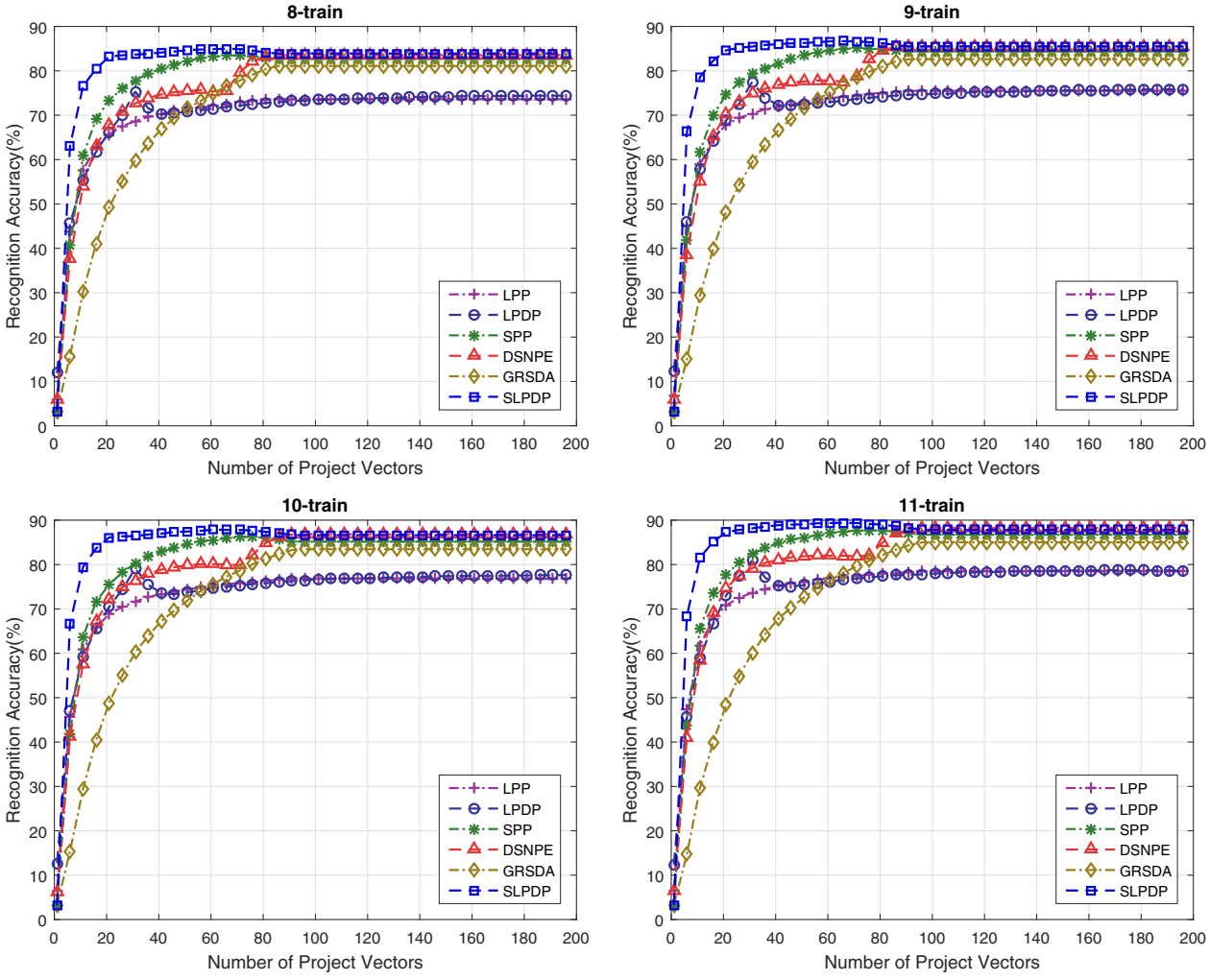


Fig. 6. Recognition accuracy vs. number of project vectors on Extended Yale B database with 8, 9, 10, 11 images for each individual randomly selected for training.

Table 3
The maximal average recognition accuracy rates (%) of six approaches across 20 runs on the Extended Yale B database and the corresponding standard deviations (std) and dimensionality (shown in parentheses).

Methods	8-train	9-train	10-train	11-train
LPP	73.61 ± 0.010(85)	75.47 ± 0.012(90)	76.80 ± 0.008(91)	78.45 ± 0.014(94)
LPDP	75.33 ± 0.015(31)	77.50 ± 0.017(31)	79.13 ± 0.015(31)	80.91 ± 0.015(31)
SPP	83.54 ± 0.014(67)	85.17 ± 0.011(70)	86.27 ± 0.010(73)	87.78 ± 0.017(75)
DSNPE	83.55 ± 0.014(85)	85.55 ± 0.012(88)	86.84 ± 0.010(94)	88.28 ± 0.014(97)
GRSDA	81.12 ± 0.015(85)	82.61 ± 0.012(90)	83.45 ± 0.010(94)	85.02 ± 0.018(98)
SLPDP	84.98 ± 0.012(66)	86.72 ± 0.012(67)	87.89 ± 0.010(62)	89.41 ± 0.016(67)

Fig. 7. Sample images of one person in the Yale database.

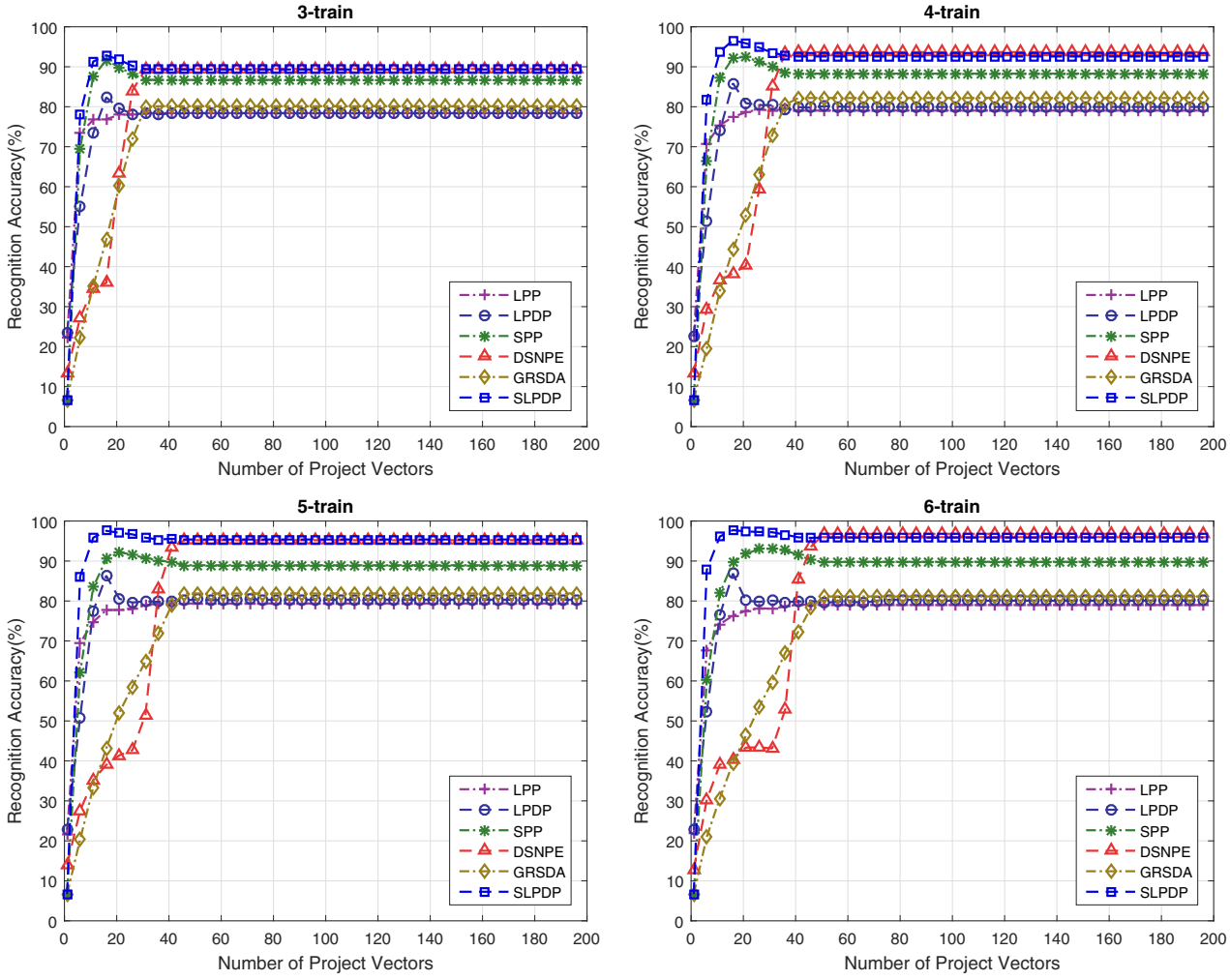


Fig. 8. Recognition accuracy vs. number of project vectors on Yale database with 3, 4, 5, 6 images for each individual randomly selected for training.

Table 4
The maximal average recognition accuracy rates (%) of six approaches across 20 runs on the Yale database and the corresponding standard deviations (std) and dimensionality (shown in parentheses).

Methods	3-train	4-train	5-train	6-train
LPP	78.46 ± 0.035(25)	79.62 ± 0.037(29)	79.50 ± 0.046(37)	79.13 ± 0.044(41)
LPDP	85.36 ± 0.024(14)	86.86 ± 0.035(14)	88.22 ± 0.031(14)	88.00 ± 0.039(14)
SPP	91.88 ± 0.029(15)	92.90 ± 0.027(17)	92.17 ± 0.023(20)	93.13 ± 0.027(34)
DSNPE	89.46 ± 0.028(32)	93.86 ± 0.021(37)	95.22 ± 0.027(46)	96.80 ± 0.015(50)
GRSDA	80.13 ± 0.034(32)	82.19 ± 0.040(38)	81.89 ± 0.048(46)	81.20 ± 0.046(50)
SLPDP	93.21 ± 0.031(14)	96.62 ± 0.018(15)	97.83 ± 0.016(15)	97.87 ± 0.013(15)

(3) DSNPE is overall superior to SPP in our experiments. This is probably because DSNPE combines unsupervised l_1-graph and global structure together to project the input high-dimensional image into a low-dimensional feature vector and well preserves the discriminant information.
(4) As a variant of LPP, GRSDA is overall superior to LPP. This is probably because GRSDA preserves the intra-class sparse locality relationships like LPP, at the same time, maximizes the inter-class sparse scatter.
(5) Our proposed approach SLPDP is superior to other approaches. This is probably because SLPDP preserves the intra-class sparse locality relationships and maximizes the margin between classes simultaneously, which well preserves the discriminant information of the data.

4.5. Discussion of parameter

In this section, we discuss the effect of parameter γ for classification performance. In experiments, we randomly select 10 images per person for training and the remaining images for testing in PIE database. Similarly, we also randomly select 10 and 5 images per person for training and the remaining images for testing in Extended Yale B database and Yale database, respectively. Figs. 9, 10 and 11 plot the curve of recognition accuracy of SLPDP vs. parameter γ on PIE, Extended Yale B and Yale database, respectively. We can see that, if γ is 0, the recognition accuracy of SLPDP is not the best. It indicates that (14) is important for image classification, i.e. the global geometric structure of samples is important for image classification. Moreover, Figs. 9–11 indicate that γ is different

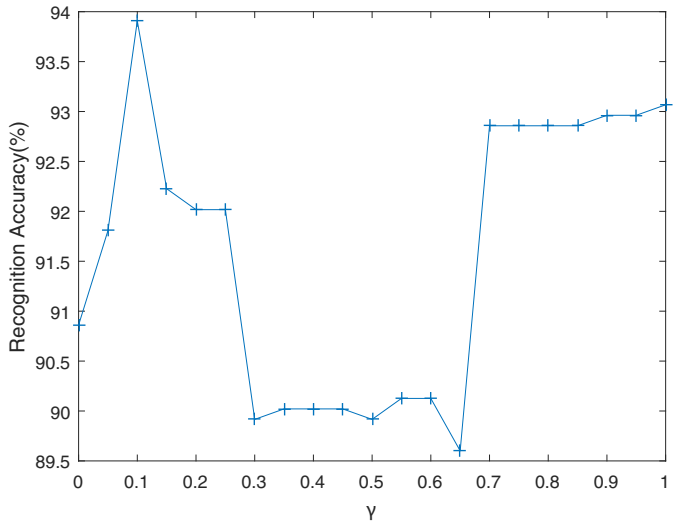


Fig. 9. Recognition accuracy vs. γ on PIE database with 10 images for each individual randomly selected for training.

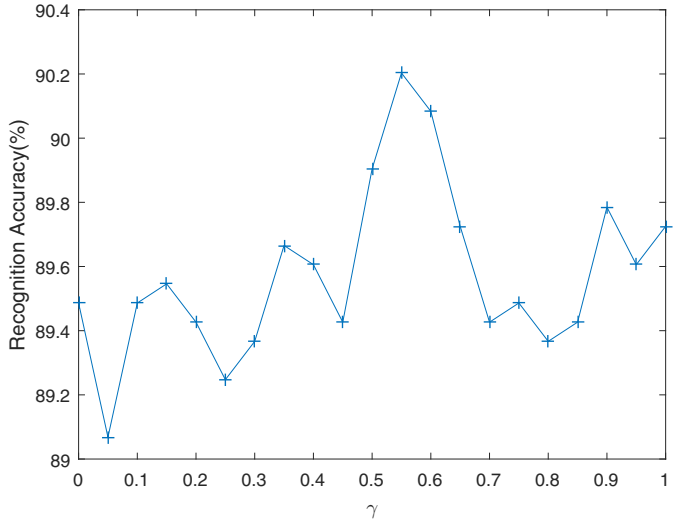


Fig. 10. Recognition accuracy vs. γ on Extended Yale B database with 10 images for each individual randomly selected for training.

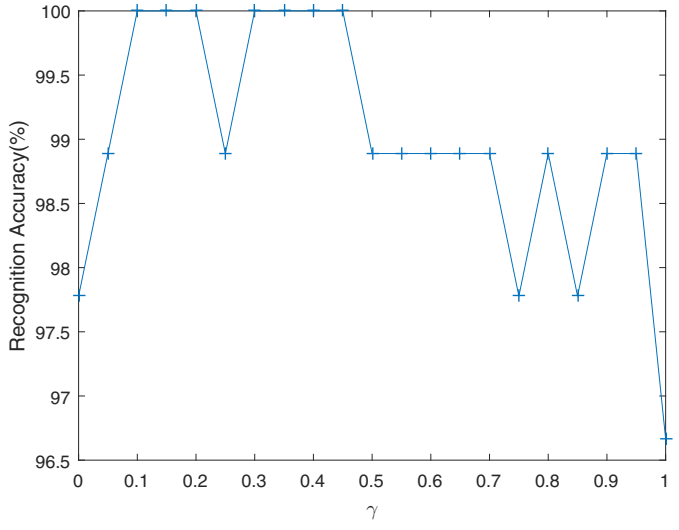


Fig. 11. Recognition accuracy vs. γ on Yale database with 5 images for each individual randomly selected for training.

for different databases when SLPDP obtains the better recognition accuracy.

5. Conclusion and future work

In this paper, based on sparse representation, a new sparse subspace learning algorithm called sparse locality preserving discriminative projections is proposed for supervised learning. Different from the existing subspace learning methods, which manually construct an adjacency graph like manifold learning or fully emphasize the sparse structure of the samples like sparse subspace learning, SLPDP preserves the intra-class sparse locality relationships via sparse representation model to adaptively build intrinsic adjacency graph and maximizes the average margin between classes. Experiments were carried out on face recognition, and the results confirmed that the proposed algorithm has superior performance compared with related algorithms. Moreover, we only conduct our experiments on face images. Since sparse representation has been applied to other pattern recognition problem, e.g. handwritten numeral [37], we will conduct some experiments on these data sets in our future work.

Conflict of interest

The authors declare that there is no conflict of interest.

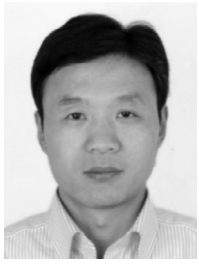
Acknowledgements

This work was supported by the National Natural Science Foundation of China [Grant numbers 61374097 and 61601108].

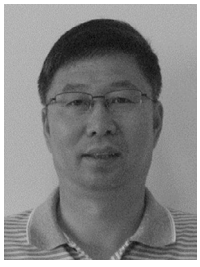
References

[1] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, Proceedings of the IEEE Computer Visual and Pattern Recognition (CVPR), 1994, pp. 84–91, doi:10.1109/CVPR.1994.323814.
[2] H. Murase, S.K. Nayar, Visual learning and recognition of 3-D objects from appearance, Int. J. Comput. Vis. 14 (1) (1995) 5–24, doi:10.1007/BF01421486.
[3] M. Turk, A. Pentland, Face recognition using eigenfaces, Proceedings of the IEEE Computer Visual and Pattern Recognition (CVPR), 1991, pp. 586–591, doi:10.1109/CVPR.1991.139758.
[4] P.N. Bellhumer, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720, doi:10.1109/34.598228.
[5] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233, doi:10.1109/34.908974.
[6] R. Kuhn, P. Nguyen, J. Junqua, Eigenfaces and eigenvoices: dimensionality reduction for specialized pattern recognition, Proceedings of the IEEE 2nd Workshop on Multimedia Signal Processing (MMSP), 1998, pp. 71–76, doi:10.1109/MMSP.1998.738915.
[7] H.F. Li, T. Jiang, K.S. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Netw. 17 (1) (2006) 157–165, doi:10.1109/TNN.2005.860852.
[8] J. Liu, S.C. Chen, X.Y. Tan, D.Q. Zhang, Comments on efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Netw. 18 (6) (2007) 1862–1864, doi:10.1109/TNN.2007.900813.
[9] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326, doi:10.1126/science.290.5500.2323.
[10] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323, doi:10.1126/science.290.5500.2319.
[11] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396, doi:10.1162/089976603321780317.
[12] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005a) 328–340, doi:10.1109/TPAMI.2005.55.
[13] X. He, D. Cai, S. Yan, H. Zhang, Neighborhood preserving embedding, Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV), 2005b, pp. 1208–1213, doi:10.1109/ICCV.2005.167.
[14] H. Chen, H. Chang, T. Liu, Local discriminant embedding and its variants, Proceedings of the IEEE Computer Visual and Pattern Recognition (CVPR), 2005, pp. 846–853, doi:10.1109/CVPR.2005.216.
[15] G. Lu, Z. Lin, Z. Jin, Face recognition using discriminant locality preserving projections based on maximum margin criterion, Pattern Recognit. 43 (10) (2010) 3572–3579, doi:10.1016/j.patcog.2010.04.007.

[16] J. Gui, W. Jia, L. Zhu, S. Wang, D. Huang, Locality preserving discriminant projections for face and palmprint recognition, Neurocomputing 73 (13–15) (2010) 2696–2707, doi:10.1016/j.neucom.2010.04.017.
[17] D. Cai, X. He, J. Han, Spectral regression: a unified approach for sparse subspace learning, in: Proceedings of the International Conference on Data Mining (ICDM), 2007, pp. 73–82, doi:10.1109/ICDM.2007.89.
[18] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227, doi:10.1109/TPAMI.2008.79.
[19] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341, doi:10.1016/j.patcog.2009.05.005.
[20] B. Cheng, J.C. Yang, S.C. Yan, Y. Fu, T.S. Huang, Learning with l(1)-graph for image analysis, IEEE Trans. Image Process. 19 (4) (2010) 858–866, doi:10.1109/TIP.2009.2038764.
[21] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process. 20 (5) (2011) 1327–1336, doi:10.1109/TIP.2010.2090535.
[22] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, S. Ji, Discriminant sparse neighborhood preserving embedding for face recognition, Pattern Recognit. 45 (8) (2012) 2884–2893, doi:10.1016/j.patcog.2012.02.005.
[23] Q. Gao, Y. Huang, H. Zhang, X. Hong, Y.W. K. Li, Discriminative sparsity preserving projections for image recognition, Pattern Recognit. 48 (8) (2015) 2543–2553, doi:10.1016/j.patcog.2015.02.015.
[24] G.F. Lu, Z. Jin, J. Zou, Face recognition using discriminant sparsity neighborhood preserving embedding, Knowl.-Based Syst. 31 (8) (2012) 119–127, doi:10.1016/j.patcog.2012.02.005.
[25] S. Yan, H. Wang, Semi-supervised learning by sparse representation, SIAM International Conference on Data Mining (SDM), 2009, pp. 792–801.
[26] S. Lou, X. Zhao, Y. Chuang, H. Yu, S. Zhang, Graph regularized sparsity discriminant analysis for face recognition, Neurocomputing 173 (2) (2016) 290–297, doi:10.1016/j.neucom.2015.04.116.
[27] Q. Zhang, K. Deng, T. Chu, Sparsity induced locality preserving projection approaches for dimensionality reduction, Neurocomputing 200 (2016) 35–46, doi:10.1016/j.neucom.2016.03.019.
[28] D.L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (4) (2006) 1289–1306, doi:10.1109/TIT.2006.871582.
[29] S. Zhang, R. Jing, Dimension reduction based on modified maximum margin criterion for tumor classification, Proceedings of the 4th International Conference on Information and Computing (ICIC), 2011, pp. 552–554, doi:10.1109/ICIC.2011.148.
[30] S.C. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Ling, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51, doi:10.1109/TPAMI.2007.250598.
[31] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 14 (6) (2002) 585–591.
[32] W. Liu, S.F. Chang, Robust multi-class transductive learning with graphs, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 381–388, doi:10.1109/CVPRW.2009.5206871.
[33] J. Yang, D. Zhang, J. Yang, B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007) 650–664, doi:10.1109/TPAMI.2007.1008.
[34] T. Sim, S. Baker, M. Bsa, The CMU pose, illumination, and expression (pie) database, Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (AFGR), 2002, pp. 46–51, doi:10.1109/AFGR.2002.1004130.
[35] A.M. Martinez, R. Benavente, The AR Face Database, CVC Technical Report, 1998 http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html.
[36] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698, doi:10.1109/TPAMI.2005.92.
[37] J. Yang, L. Zhang, Y. Xu, J. Yang, Beyond sparsity: the role of l1-optimizer in pattern classification, Pattern Recognit. 45 (3) (2012) 1104–1118, doi:10.1016/j.patcog.2011.08.022.



Jianbo Zhang received his B.S. degree in Computing Mathematics from Hebei University, Baoding, China, in 2001, and the M.S. degree in Computer Application and Technology from Northeastern University, Shenyang, China, in 2007. He is currently working toward the Ph.D. degree in the School of Information Science and Engineering, Northeastern University. His research interests are machine learning, pattern recognition and image processing.



Jinkuan Wang received his M.S. degree from Northeastern University, Shenyang, China, in 1985 and the Ph.D. degree from University of Electro-Communication, Tokyo, Japan, in 1993. As a special member, he joined the Institute of Space Astronautical Science, Japan, in 1990. And he worked as an Engineer in the Research Department, COSEL, Japan, in 1994. He is currently a Professor in the Institute of Information and Engineering at Northeastern University, China, since 1998. His main interests are in the area of adaptive signal processing, mobile communication and intelligent control.

Xi Cai received her B.S. and Ph.D. degrees from the School of Electronic and Information Engineering, Beihang University, China, in 2005 and 2011, respectively. Now she is a teacher of engineering optimization at the Smart Antenna Institute, Northeastern University in Qinhuangdao, China. Her research interests include image processing and video analysis.