

# Sparsity Regularization Discriminant Projection for Feature Extraction

Sen Yuan<sup>1</sup> · Xia Mao<sup>1</sup> · Lijiang Chen<sup>1</sup>

Published online: 24 April 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

**Abstract** Recently, sparse representation models have attracted considerable interests in the field of feature extraction. In this paper, we propose a novel supervised feature extraction method called sparsity regularization discriminant projection (SRDP), which aims to preserve the sparse representation structure of the data and simultaneously maximize the ratio of nonlocal scatter to local scatter. More specifically, SRDP first constructs a concatenated dictionary through the class-wise principal component analysis decompositions. Second, the sparse representation structure of each sample is quickly learned with the constructed dictionary by matrix–vector multiplications. Then SRDP regards the learned sparse representation structure as an additional regularization term of unsupervised discriminant projection so as to construct a new discriminant function. Finally, SRDP is transformed into a generalized eigenvalue problem. Experimental results on five representative image databases demonstrate the effectiveness of our proposed method.

**Keywords** Sparse representation · Feature extraction · Manifold learning · Unsupervised discriminant projection · Face recognition

## 1 Introduction

Feature extraction is a fundamental and challenging problem in the area of computer vision and pattern recognition [1–3]. In many existing feature extraction algorithms, principal components analysis (PCA) [4] and linear discriminant analysis (LDA) [5] are the two classical linear feature extraction methods and have been widely used in many practical applications. PCA attempts to project the data along an optimal direction by maximizing the variance matrix of data. Unlike PCA, LDA is a supervised method that seeks to find a projection direction by maximizing the inter-class scatter when minimizing the inner-class scatter. Because the label

---

✉ Lijiang Chen  
chenlijiang@buaa.edu.cn

<sup>1</sup> School of Electronic and Information Engineering, Beihang University, Beijing 100191, China

information is fully exploited, LDA has been proven more efficient than PCA for the classification tasks [5]. To further improve the discriminant ability of feature extraction, some LDA variants have also been proposed, such as Enhanced Fisher Discriminant Criterion (EFDC) [6], Maximum Margin Criterion (MMC) [7], EMKFDA [8], and Orthogonal LDA (OLDA) [9].

To our best knowledge, linear feature extraction methods may fail to discover the underlying nonlinear structure hidden in high-dimensional data. To remedy this deficiency, a large number of manifold learning algorithms have been proposed. The representative manifold learning algorithms include locally linear embedding (LLE) [10], isometric feature mapping (ISOMAP) [11], and Laplacian eigenmaps (LE) [12]. Unfortunately, these manifold learning methods usually suffer from the out-of-sample problem [13]. This is because they fail to construct explicit maps over new measurements. To address this problem, locality preserving projection (LPP) [14] tries to seek a linear approximation to the eigen-functions of Laplace—Beltrami operator on the manifold derived from LE. Neighborhood preserving embedding (NPE) [15] tries to find a linear subspace which preserves the local structure based on the same principle of LLE. To construct an optimal graph for the later feature extraction, discriminative unsupervised dimensionality reduction (DUDR) [16] was proposed. Yan et al. [17] introduced a general framework for feature extraction, called graph embedding. A large number of methods, e.g., LPP, DUDR and OLMGMP [3] can all be considered as the special cases within this framework. Recently, many other methods have been explored in [18–20], and as expected, they have achieved good performance in the classification tasks.

However, a common problem with current subspace learning methods is that they only character the locality of samples such that they cannot guarantee a good projection for the classification purposes. To address this problem, Yang et al. [21] proposed unsupervised discriminant projection (UDP). UDP introduces the concept of non-locality and learns the low-dimensional representation of data by maximizing the ratio of nonlocal scatter to local scatter. Nie et al. [22] proposed neighborhood min–max projections (NMMP) by introducing discriminant information into the local structure. Zhang et al. [23] proposed complete global–local LDA (CGLDA) to incorporate three kinds of local information into LDA. Gao et al. [24] proposed joint global and local structure discriminant analysis (JGLDA), which used two quadratic functions to characterize the geometric properties of similarity and diversity of data. In the literature [25], the authors proposed elastic preserving projections (EPP) which considers both the local structure and the global information of data. Luo et al. [26] added the discriminant information and orthogonal constraint into EPP, and proposed discriminant orthogonal elastic preserving projections (DOEPP). To overcome the singular problem of EPP, exponential EPP (EEPP) [27] was proposed. These methods usually share underlying commonality that they integrate both the nonlocal (global) and local structure into the objective function of feature extraction.

Sparse representation has received considerable interest in recent years, especially in image recognition [28–30]. The main idea of sparse representation is that a given sample can be represented as a linear combination of the others. The coefficients obtained by sparse representation reflect the contributions of the samples to reconstruct the given sample. The most popular feature extraction methods based on sparse representation include sparsity preserving projection (SPP) [31], discriminant sparse neighborhood preserving embedding (DSNPE) [32], and discriminant sparse local spline embedding (D-SLSE) [33]. In general, they have achieved better performance than the conditional methods. But, all of them must solve the L1-norm minimization problem to construct the sparse weight matrix, so that their computational complexity is excessively high [34]. Recently, objective functions based on complex-norm have been explored, and they are widely used in different fields [35–38].

Motivated by the above works, we propose a novel supervised feature extraction algorithm called sparsity regularization discriminant projection (SRDP). Specially, SRDP first constructs a concatenated dictionary through the class-wise PCA decompositions and learns the sparse representation structure of each sample. Then SRDP utilizes the learned sparse representation structure as an additional regularization term of UDP so as to construct a new discriminant function. Finally, SRDP is transformed into a generalized eigenvalue problem. Our primary contributions can be summarized as follows:

1. We proposed a novel feature extraction algorithm, called SRDP, to learn the discriminant features of data. SRDP considers both the nonlocal and local structure of data, and simultaneously preserves the sparse representation structure.
2. Compared to UDP, SRDP is a supervised method and can alleviate the small sample size problem (SSS) by introducing the sparse regularization term.
3. Under the concatenated dictionary constructed by class-wise PCA decompositions, the sparse coefficient in SRDP can be obtained quickly via matrix–vector multiplication. So, its computational complexity is significantly less than other algorithms based on sparse representation via L1-norm optimization, such as SPP and DSNPE.
4. Unlike LPP, EPP, and DOEPP, SRDP considers the local structure twice. The first time is in learning the sparse representation. This is because the sparse representation can implicitly discover the local structure of data. The second time is in constructing the adjacency graph by  $K$  nearest neighbors. This adjacency graph also characterizes the locality of samples.

The rest of this paper is organized as follows. We briefly review the SPP and UDP algorithm in Sect. 2. In Sect. 3, we introduce the proposed SRDP algorithm in detail. In Sect. 4, extensive experiments are carried out to demonstrate the effectiveness of the proposed method. Finally, Sect. 5 concludes the paper.

## 2 Review of the Related Work

### 2.1 Sparsity Preserving Projections

SPP [31] attempts to preserve the sparse reconstruction relationship of samples. Given a sample set  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^{D \times N}$ , where  $D$  is the number of features and  $N$  is the number of samples. SPP first learns the sparse coefficient vector  $\mathbf{s}_i$  for each sample  $\mathbf{x}_i$  by solving the following L1-norm minimization problem:

$$\begin{aligned} \min \|\mathbf{s}_i\|_1 \\ \text{s.t. } \mathbf{x}_i = \mathbf{X}\mathbf{s}_i, \mathbf{1} = \mathbf{1}^T \mathbf{s}_i \end{aligned} \quad (1)$$

where  $\|\cdot\|_1$  is the L1-norm and  $\mathbf{1} \in \mathbb{R}^N$  is a vector of all ones. Once all the coefficient vectors  $\mathbf{s}_i (i = 1, 2, \dots, N)$  are computed, the sparse reconstruction weight matrix  $\mathbf{S}$  can be obtained by

$$\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N] \quad (2)$$

Finally, based on the weight matrix  $\mathbf{S}$ , the objective function of SPP can be represented as:

$$\min_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{X} (\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T \mathbf{S}) \mathbf{X}^T \mathbf{w}}{\mathbf{w}^T \mathbf{X} \mathbf{X}^T \mathbf{w}} \tag{3}$$

The optimal projection vector  $\mathbf{w}$  can be obtained by solving for the eigenvector corresponding to the smallest eigenvalue in the generalized eigenvalue equation:

$$\mathbf{X} (\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T \mathbf{S}) \mathbf{X}^T \mathbf{w} = \lambda \mathbf{X} \mathbf{X}^T \mathbf{w} \tag{4}$$

Seen from Eqs. (1) and (2), SPP has to resolve  $N$  time-consuming L1-norm minimization problems to obtain the sparse weight matrix  $\mathbf{S}$ , such that its computational complexity reaches up to  $O(N^4)$ , which is excessively high in real applications. Besides, the matrix  $\mathbf{X} \mathbf{X}^T$  is always singular since the training sample size is much smaller than the feature dimensions.

### 2.2 Unsupervised Discriminant Projection

UDP [21] incorporates the advantage of both the locality and nonlocality of samples. A concise criterion for feature extraction can be obtained by maximizing the ratio of nonlocal scatter to local scatter. The local scatter matrix is defined by

$$\mathbf{S}_L = \frac{1}{2} \sum_{i,j} (\mathbf{x}_i - \mathbf{x}_j) (\mathbf{x}_i - \mathbf{x}_j)^T H_{ij} \tag{5}$$

where  $H_{ij}$  is defined as

$$H_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in O(K, \mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where  $O(K, \mathbf{x}_i)$  denotes the set of  $K$  nearest neighbors of  $\mathbf{x}_i$ .

Similarly, the nonlocal scatter matrix can be defined by

$$\mathbf{S}_N = \frac{1}{2} \sum_{i,j} (\mathbf{x}_i - \mathbf{x}_j) (\mathbf{x}_i - \mathbf{x}_j)^T (1 - H_{ij}) \tag{7}$$

UDP then optimizes:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_N \mathbf{w}}{\mathbf{w}^T \mathbf{S}_L \mathbf{w}} \tag{8}$$

Generally, the number of training samples is always less than their features. This results in that UDP suffers from the SSS problem. In addition, UDP does not exploit the sparse representation structure of data, which is important for improving classification tasks.

### 3 Sparsity Regularization Discriminant Projection

In this section, we introduce the proposed SRDP in detail. SRDP can be regarded as the combiner of sparse representation and UDP. But differing from UDP, it considers the sparse representation structure of data, and avoids the SSS problem by introducing the sparse regularization term. And differing from sparse representation based methods, such as SPP and DSNPE, SRDP significantly reduces the computational complexity of learning the sparse representation structure via a concatenated dictionary rather than solving the L1-norm optimization.

### 3.1 Constructing the Concatenated Dictionary and Learning the Sparse Representation Structure

Given a set of training samples  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ , where  $\mathbf{x}_i \in \mathbb{R}^D$ . Now, we categorize the samples as  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_C]$ , where  $C$  is the number of classes, and  $\mathbf{X}_i \in \mathbb{R}^{D \times N_i}$  contains the samples from class  $i$ . For the convenience of relevant calculations, we first center the samples from each class at the origin,  $\tilde{\mathbf{X}}_i = [\mathbf{x}_1 - \mu_i, \mathbf{x}_2 - \mu_i, \dots, \mathbf{x}_{N_i} - \mu_i]$  ( $i=1, 2, \dots, C$ ), where  $\mu_i$  is mean of samples belonging to class  $i$ . Then we conduct PCA decomposition for each class  $\mathbf{X}_i$ , whose objective function is

$$\max_{\|\mathbf{d}\|=1} \mathbf{d}^T \Phi_i \mathbf{d} \tag{9}$$

where  $\Phi_i$  is the covariance matrix of  $\tilde{\mathbf{X}}_i$ . For each class  $i$ ,  $N_i$  principal components are selected to construct  $\mathbf{D}_i = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{N_i}]$ . Thus, a sample  $\mathbf{x}$  from class  $i$  can be represented by

$$\mathbf{x} = \mathbf{D}_i \mathbf{s}_i = [\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_{i-1}, \mathbf{D}_i, \mathbf{D}_{i+1}, \dots, \mathbf{D}_C] \tilde{\mathbf{s}} = \mathbf{D} \tilde{\mathbf{s}} \tag{10}$$

where  $\mathbf{D}$  is the concatenated dictionary constructed by PCA decompositions, and it consists of all  $\mathbf{D}_i$  ( $i = 1, 2, \dots, C$ ).  $\tilde{\mathbf{s}} = [0^T, 0^T, \dots, 0^T, \mathbf{s}_i^T, 0^T, \dots, 0^T]^T$  is the sparse co-efficient vector under the concatenated dictionary  $\mathbf{D}$ .

According to the previous procedure, each sample corresponds to a sparse coefficient vector. From Eq. (10), we find that the computation of  $\tilde{\mathbf{s}}$  involves only  $\mathbf{D}_i$ , which is also column orthogonal, so that the sparse coefficient vector  $\tilde{\mathbf{s}}$  of any training sample from class  $i$  can be quickly learned by matrix–vector multiplication, i.e.

$$\tilde{\mathbf{s}} = \left[ 0^T, 0^T, \dots, 0^T, \left( \mathbf{D}_i^T \mathbf{x} \right)^T, 0^T, \dots, 0^T \right]^T \tag{11}$$

### 3.2 Preserving Sparse Representation Structure

As can be seen from Sect. 3.1, the sparse representation structure well encodes the local information of the training samples. It is expected that the sparse representation structure in the original high-dimensional space can be preserved in the low-dimensional projective subspace. Thus, the following objective function is defined to seek a projection that preserves the sparse representation structure:

$$J_s(\mathbf{w}) = \min_{\mathbf{w}} \sum_i \mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{D} \tilde{\mathbf{s}} \|_{\text{IF}}^2 \tag{12}$$

where  $\tilde{\mathbf{s}}$  is the sparse coefficient vector corresponding to  $\mathbf{x}_i$ . With some algebraic operations, Eq. (12) can be rewritten as

$$\begin{aligned} J_s(\mathbf{w}) &= \min_{\mathbf{w}} \sum_i \mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{D} \tilde{\mathbf{s}} \|_{\text{IF}}^2 \\ &= \min_{\mathbf{w}} \mathbf{w}^T \left( \sum_{i=1}^N (\mathbf{x}_i - \mathbf{D} \tilde{\mathbf{s}}_i) (\mathbf{x}_i - \mathbf{D} \tilde{\mathbf{s}}_i)^T \right) \mathbf{w} \\ &= \min_{\mathbf{w}} \mathbf{w}^T \left( \mathbf{X} \mathbf{X}^T - \mathbf{X} \mathbf{S}^T \mathbf{D}^T - \mathbf{D} \mathbf{S} \mathbf{X}^T + \mathbf{D} \mathbf{S} \mathbf{S}^T \mathbf{D}^T \right) \mathbf{w} \end{aligned} \tag{13}$$

where  $\mathbf{S} = [\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \dots, \tilde{\mathbf{s}}_N]$ .

### 3.3 Sparsity Regularization Discriminant Projection

The goal of proposed SRDP aims to find the optimal projections that can, on the one hand, preserve the sparse representation structure, on the other hand, maximize the ratio of nonlocal scatter to local scatter. To this end, we choose to maximize the following criterion:

$$\max_w \frac{\mathbf{w}^T \mathbf{S}_N \mathbf{w}}{\alpha \mathbf{w}^T \mathbf{S}_L \mathbf{w} + (1 - \alpha) J_s(\mathbf{w})} \tag{14}$$

where  $\alpha(0 < \alpha < 1)$  is a parameter that controls the trade-off between the two terms in numerator, and it can be adjusted if balancing is needed.  $\mathbf{S}_N$  and  $\mathbf{S}_L$  are the nonlocal scatter matrix and the local scatter matrix, respectively.

Substituting Eq. (13) into (14), we have

$$\begin{aligned} & \max_w \frac{\mathbf{w}^T \mathbf{S}_N \mathbf{w}}{\alpha \mathbf{w}^T \mathbf{S}_L \mathbf{w} + (1 - \alpha) J_s(\mathbf{w})} \\ &= \max_w \frac{\mathbf{w}^T \mathbf{S}_N \mathbf{w}}{\alpha \mathbf{w}^T \mathbf{S}_L \mathbf{w} + (1 - \alpha) \mathbf{w}^T (\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T \mathbf{D}^T - \mathbf{D}\mathbf{S}\mathbf{X}^T + \mathbf{D}\mathbf{S}\mathbf{S}^T \mathbf{D}^T) \mathbf{w}} \\ &= \max_w \frac{\mathbf{w}^T \mathbf{X}\mathbf{S}_N \mathbf{X}^T \mathbf{w}}{\mathbf{w}^T [\alpha \mathbf{S}_L + (1 - \alpha) (\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T \mathbf{D}^T - \mathbf{D}\mathbf{S}\mathbf{X}^T + \mathbf{D}\mathbf{S}\mathbf{S}^T \mathbf{D}^T)] \mathbf{w}} \end{aligned} \tag{15}$$

Let

$$\Psi = \alpha \mathbf{S}_L + (1 - \alpha) (\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T \mathbf{D}^T - \mathbf{D}\mathbf{S}\mathbf{X}^T + \mathbf{D}\mathbf{S}\mathbf{S}^T \mathbf{D}^T)$$

Then the optimization problem in Eq. (15) can be converted to the following generalized eigenvalue problem:

$$\mathbf{S}_N \mathbf{w} = \lambda \Psi \mathbf{w} \tag{16}$$

The projection matrix  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d]$  consists of the eigenvectors corresponding to the  $d$  largest eigenvalues.

The algorithmic procedure of SRDP can be formally summarized as follows:

**Input:** Training set  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ ,  $\mathbf{x}_i \in \mathbb{R}^D$  ( $i = 1, 2, \dots, N$ ) and the balance parameter  $\alpha$ .

**Output:** The projection matrix  $\mathbf{W}$

Step 1: Calculate  $\mathbf{S}_L$  and  $\mathbf{S}_N$  by Eq. (5) and Eq. (7), respectively;

Step 2: Conduct PCA decompositions to construct the concatenated dictionary  $\mathbf{D}$  by Eq. (9);

Step 3: Learn the sparse coefficient vector  $\bar{\mathbf{s}}$  for every sample by Eq. (11), and then calculate the sparse reconstruction weight matrix  $\mathbf{S}$ ;

Step 4: Calculate  $\Psi$ , and solve the generalized eigenvalue problem in Eq. (16).

### 3.4 Computational Complexity Analysis

The computational complexity of SRDP includes four parts: the scatter matrices, concatenated dictionary, sparse coefficient vector, and generalized eigenvalue. The computation of scatter matrices requires  $O(N^2 * D + D^2 * N)$ . The complexity to construct concatenated dictionary via PCA decompositions is  $O(D^2 * \sum_i^C m_i)$ . The complexity of calculating the sparse coefficient vectors for all samples is  $O(D * \sum_i^C m_i N_i)$ , where  $N_i$  denotes the number of samples in class  $i$ . In general,  $C \ll N, m_i \ll N, N_i \ll N$ , so  $O(D * \sum_i^C m_i N_i) \ll$



**Fig. 1** Ten sample images of one person in ORL database

$O(N^4)$ . This means that the computational complexity of learning the sparse weight matrix in SRDP is much less than that in the algorithms based on sparse representation via L1-norm optimization, such as SPP, DSNPE, and D-SLSE. The complexity for solving the generalized eigenvalue is  $O(D^3)$ . Finally, we can conclude that the computational complexity of SRDP is  $O(N^2 * D + D^2 * N + D^2 * \sum_i^C m_i + D * \sum_i^C m_i N_i + D^3)$ .

## 4 Experimental Results

In this section, we make a set of experiments to evaluate the effectiveness of the proposed SRDP for feature extraction, and compare it with several popular algorithms, including LPP, UDP, EPP, DOEPP, and SPP. Five benchmark databases are used in our experiments: ORL database [39], Yale database, CMU PIE database [40], FERET database [41] and LFW database [42]. Note that, to overcome the SSS problem, LPP, UDP, EPP, and SPP all involve a PCA phase. In this phase, we keep 98% data energy. For LPP, UDP, EPP, DOEPP, and SRDP, the  $K$ -nearest neighborhood parameter  $K$  is set to  $l - 1$ , where  $l$  denotes the number of training samples per class. In DOEPP, we set  $\alpha = 0.5$  and  $\beta = 1$ , respectively. The value of  $\alpha$  in SRDP is empirically set to be 0.1 in all experiments. After all the methods have been adopted to extract low-dimensional feature, the nearest neighbor classifier with Euclidean matrix is employed to perform classification task.

### 4.1 Experiments on ORL Database

The ORL database contains 400 images of 40 individuals, and each individual has ten images. The face images are captured at different times and have different variations including illumination, expressions (open or closed eyes, smiling or non-smiling) and facial details (glasses or no glasses). In our experiments, each image was resized to  $32 \times 32$ . Some samples from this database are shown in Fig. 1.

First, in order to demonstrate the performance of each method with varying number of training samples, we randomly selected  $l$  ( $l = 4, 5, 6, 7$ ) images per person for training, and the rest images for testing. For each giving  $l$ , the experiments are independently performed 20 times. Table 1 presents the best average recognition results and the corresponding dimensions for each method. As can be seen from Table 1, SRDP can get the best recognition rates in all experiments.

Second, to report the computational time (C-T) of different methods on ORL database, the first five images per person are selected for training and the remaining images are used for testing. The C-T is represented by the whole operation time for training and classification. The results are shown in Table 2. From the results, we can see three main points. First, LPP is the fastest in all methods. Second, SPP is slower than all other methods. This is probably because it needs to learn the sparse weight matrix by L1-norm optimization, which is time-consuming. Third, SRDP and UDP are comparable in C-T.

**Table 1** The best average recognition accuracy (%) and the corresponding dimension on ORL database

Train	LPP	EPP	DOEPP	SPP	UDP	SRDP
4	86.00 (55)	85.42 (60)	86.75 (57)	80.71 (79)	83.96 (60)	88.67 (60)
5	88.75 (42)	89.35 (50)	89.85 (52)	87.20 (100)	86.55 (51)	91.00 (59)
6	91.81 (46)	90.00 (48)	91.06 (59)	90.13 (119)	89.31 (47)	92.69 (39)
7	92.50 (50)	93.33 (58)	93.50 (53)	93.08 (140)	91.83 (53)	93.67 (48)

**Table 2** Comparison of computational time (s) on ORL database

	LPP	EPP	DOEPP	SPP	UDP	SRDP
C-T	7.9687	8.4294	12.4046	15.8103	10.1901	10.2271

**Fig. 2** Sample images of one individual in Yale database

## 4.2 Experiments on Yale Database

The Yale database contains 165 gray scale face images from 15 individuals. The images demonstrate variations in lighting condition and facial expression (happy, normal, sad, sleepy, surprised, and wink). In the experiments, each image was resized to  $32 \times 32$ . Figure 2 shows the sample images of one individual.

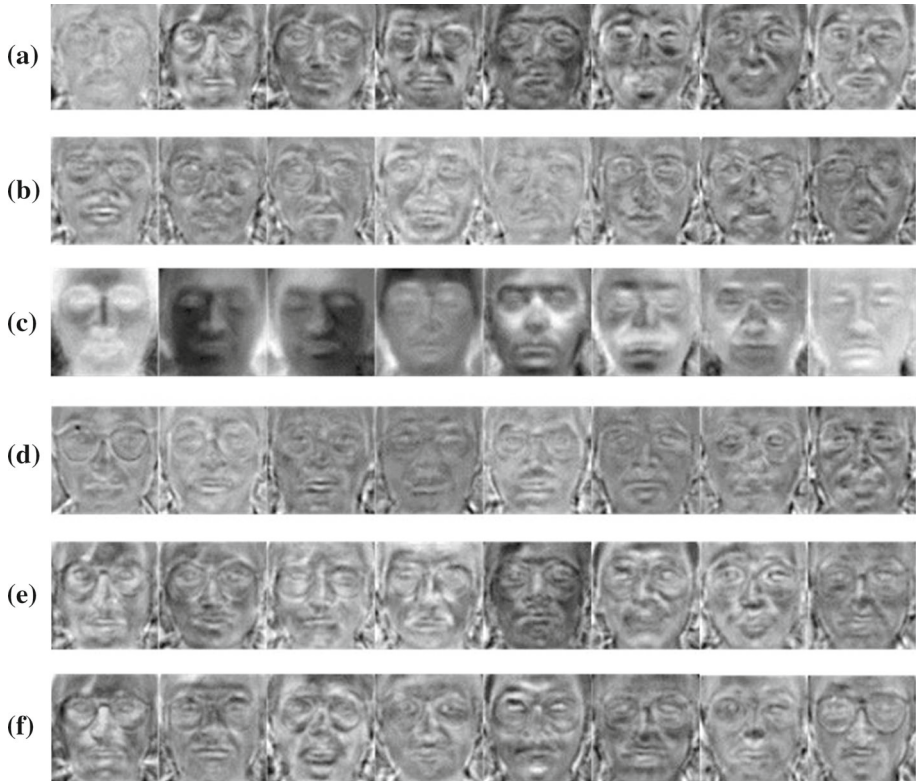
First, the projected subspaces learned by LPP, EPP, DOEPP, SPP, UDP, and SRDP are different. Thus, all images in the Yale database are used to learn such spaces spanned by the eigenvectors of the corresponding algorithms. The first eight basis vectors of different algorithms are presented in Fig. 3. It can be seen that SRDP learns a set of basis images that are different from those of the other algorithms.

In the next experiments, we randomly select  $l$  ( $l = 5, 8$ ) images per person to form the training set, and the rest are used for testing. Note that there is no overlap between the training and test sets. For each  $l$ , we average the results over 20 random splits. The performances of each method are shown in Table 3. The recognition curves versus the dimension of reduced space for each algorithm are shown in Fig. 4. From Fig. 4, we find that, when the dimension is very low, DOEPP performs better than SRDP. This is because forcing an orthogonal relationship between the projection vectors is useful for preserving the structure of data. But with the increase of dimension, SRDP becomes superior to DOEPP. This is probably due to the fact that when the number of projection vector turns to be larger, sparsity representation has an apparent advantage over orthogonal relationship. In general, SRDP can obtain the best recognition rates in the recognition tasks.

## 4.3 Experiments on CMU PIE Database

The CMU PIE face database contains more than 40,000 face images of 68 individuals. The face images were captured under varying pose, illumination and expression. In our experiments, a subset (C27) which contains about 3329 images of 68 individuals was used.





**Fig. 3** First eight basis images of different methods: **a** LPP, **b** EPP, **c** DOEPP, **d** SPP, **e** UDP, and **f** SRDP

**Table 3** The optimal recognition accuracy (%) and the corresponding dimension on Yale database

Methods	5 train		8 train	
	Accuracy	Dimension	Accuracy	Dimension
LPP	82.33	36	84.22	56
EPP	82.78	47	83.33	59
DOEPP	84.11	30	89.78	45
SPP	83.33	36	84.44	40
UDP	77.22	39	83.33	38
SRDP	85.56	39	91.33	28

Similarly, all the images were cropped to the solution of  $32 \times 32$  in our experiments. Some image samples are shown in Fig. 5.

In the experiments,  $l$  ( $l = 5, 10$ ) images per person are selected for training and the rest are used for testing. For each  $l$ , we average the results over 20 random splits. To be fair, the reduced feature dimension is searched from 1 to 40. The best average recognition results and the corresponding dimensions for each method are shown in Table 4. The recognition curves versus the dimension of reduced space for each algorithm are shown in Fig. 6. It can be found that, SRDP and SPP perform better than other methods. This is because sparse representation can improve the robustness to illumination. On the other hand, with the increase of dimension,

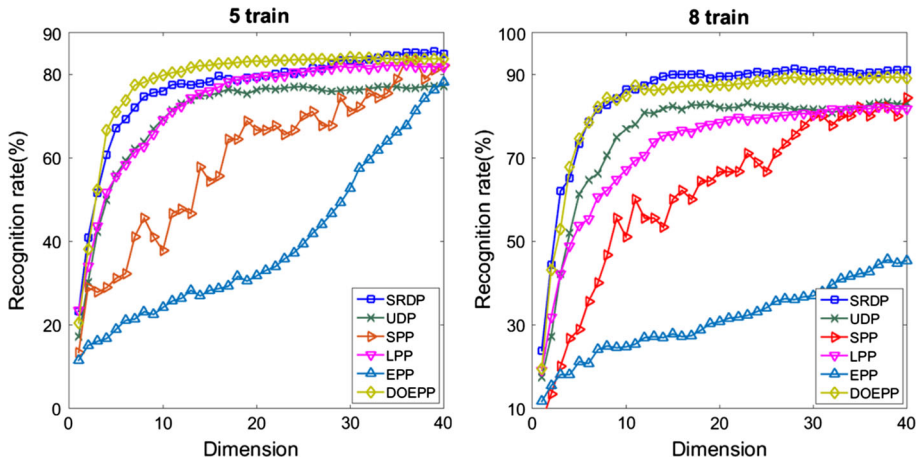


Fig. 4 Recognition accuracy versus feature dimension on Yale database for (left) 5 train and (right) 8 train



Fig. 5 Sample images of one person in CMU PIE database

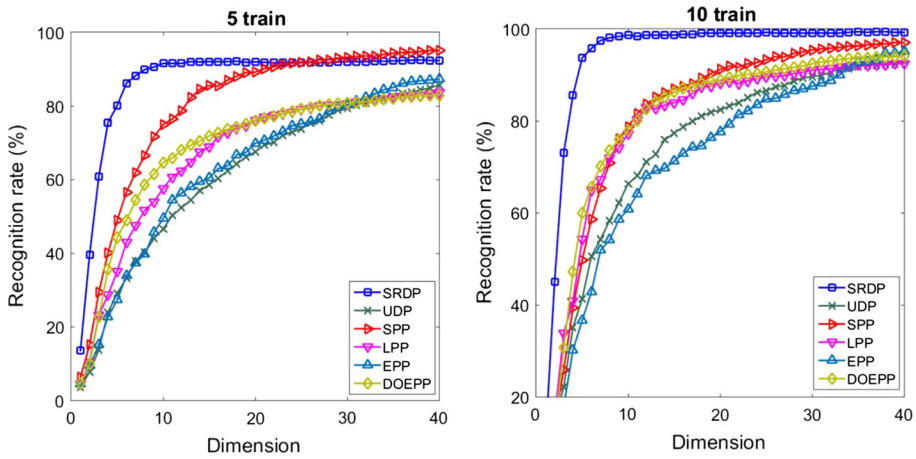
Table 4 The optimal recognition accuracy (%) and the corresponding dimension on CMU PIE database

Methods	5 train		10 train	
	Accuracy	Dimension	Accuracy	Dimension
LPP	83.73	40	92.50	39
EPP	87.25	40	95.14	40
DOEPP	82.84	39	94.12	40
SPP	95.10	39	97.06	39
UDP	85.69	40	93.09	40
SRDP	92.54	38	99.41	35

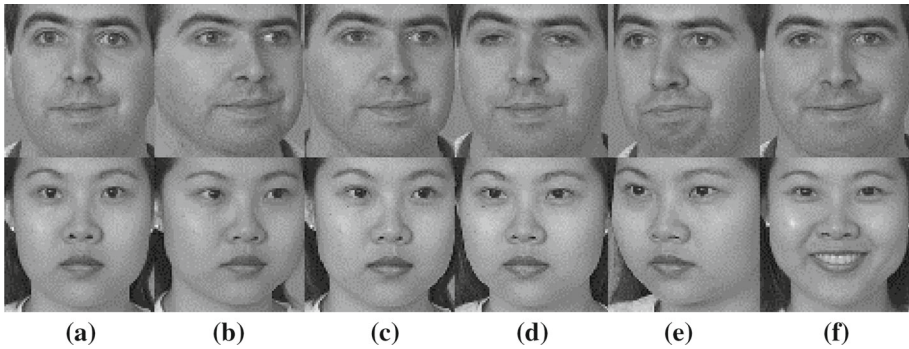
SRDP tends to be more stable than SPP. This is induced by considering the local and non-local structure of data.

### 4.4 Experiments on FERET Database

The FERET database comprises a total of 11,338 facial images of 994 distinct individuals. The size of the images is of  $512 \times 768$  pixels. In our experiments, we select a subset which includes 1200 images of 200 different subjects from the FERET face database. Figure 7 shows the samples of two subjects in the FERET subset. The images named “a” are frontal view. The images named “b”, “c”, “d”, and “e”, denote the variation of face in view. The images named “f” are smiling and frontal view. For the sake of efficient computation, each image was manually cropped and scaled to  $40 \times 40$  pixels.



**Fig. 6** Average recognition rate versus the dimension of reduced space on CMU PIE database for (left) 5 train and (right) 10 train



**Fig. 7** Sample images of two subjects in the FERET subset

**Table 5** The optimal recognition accuracy (%) and the corresponding dimension on FERET database

Methods	Test #1		Test #2	
	Accuracy	Dimension	Accuracy	Dimension
LPP	65.33	76	47.50	63
EPP	63.33	73	54.75	79
DOEPP	65.83	65	58.25	76
SPP	46.33	80	41.00	78
UDP	60.00	66	41.75	49
SRDP	67.50	38	60.25	38

To evaluate the performance of different methods, we design two tests called **Test #1** and **Test #2**. In **Test #1**, the training set contains “a”, “b”, and “d”, and the test set contains “c”, “e”, and “f”. In **Test #2**, the training set contains “b”, “c”, “d”, and “e”, and the test set contains “a” and “f”. Table 5 gives the optimal correct recognition rates of each method in two tests.



**Fig. 8** Sample images in LFW database and the face area is in the red bounding box

**Table 6** Performance of different methods on LFW database

	LPP	EPP	DOEPP	SPP	UDP	SRDP
Accuracy (%)	61.4	63.4	64.8	66.1	63.6	66.4
Dimension	80	45	110	120	88	76

## 4.5 Experiments on LFW Database

The LFW (Labeled faces in the Wild) database is designed for studying the problem of the unconstrained face verification task. It contains more than 13,000 images from 5749 people. All the images are collected from web, and the image number of each subject is different. Our experiments are made on a set of 600 images referring to 60 subjects (10 images for each subject). For each image, we first extract the face area, and then resize it to  $32 \times 32$  pixels. Figure 8 shows some sample images and the extracted face area from LFW database. In the experiments, 50% of the images of each subject are used for training and the rest for testing. Table 6 gives the recognition accuracy and the corresponding dimension of different methods. As shown in Table 6, the recognition accuracies of all methods are relatively low. This is because the images have great variations in age, pose, lighting and expression. But SRDP has an apparent advantage.

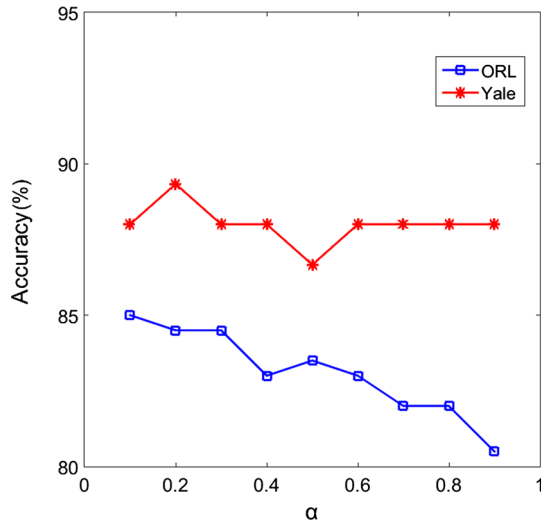
## 4.6 Discussion

To further explore the influence of  $\alpha$  in SRDP, we conduct experiments on ORL and Yale databases. To be fair, the reduced dimension is set to 40.  $\alpha$  is searched from 0.1 to 0.9. Figure 9 presents the recognition accuracy of SRDP under different values of  $\alpha$ . As seen from Fig. 9, the value of  $\alpha$  takes influence on the performance of SRDP. Moreover, it can be seen that the results in the case of  $\alpha = 0.1$  is usually comparable to the best result among other values.

From the experiments above, the following observations are obtained:

1. Seen from Table 2, SRDP is faster than SPP. This is because SRDP learns the sparse structure through matrix–vector multiplication rather than solving the L1-norm optimization.
2. From Table 4, it can be found that, sparse representation based feature extraction methods, i.e. SPP and SRDP, performs better than other methods in the experiments on CMU PIE database, where there are variations in illumination. This is because the sparse representation structure can improve the robustness to illumination.
3. SRDP can obtain the best results in most experiments. This is probably because SRDP well encodes the local, non-local, and sparse representation structure.

**Fig. 9** Recognition accuracy of SRDP for varying  $\alpha$  on ORL and Yale databases



## 5 Conclusions

In this paper, we propose a novel feature extraction algorithm, called sparsity regularization discriminant projection (SRDP). SRDP first uses the class-wise PCA decompositions to construct a concatenated dictionary under which the sparse representation structure of each sample can be learned. Then SRDP regards the sparse representation structure as an additional regularization term so as to construct a new discriminant function. Finally, SRDP is transformed into a generalized eigenvalue problem. The experimental results on five real datasets have demonstrated the effectiveness of the proposed method.

**Acknowledgements** The authors would like to thank the anonymous reviewers for their valuable and constructive criticisms that are very helpful to improve the quality of this paper. This work was supported by the National Science Foundation of China (Grant No. 61603013).

## References

1. Wang Z, Yang W, Shen F (2016) Face recognition using a low rank representation based projections method. *Adv Space Res* 43(3):823–835
2. Yuan S, Mao X, Chen L (2017) Multilinear spatial discriminant analysis for dimensionality reduction. *IEEE Trans Image Process* 26(6):2669–2681
3. Nie F, Xiang S, Song Y, Zhang C (2009) Orthogonal locality minimizing globality maximizing projections for feature extraction. *Opt Eng* 48(1):017202
4. Turk Matthew, Pentland Alex (1991) Eigenfaces for recognition. *J Cogn Neurosci* 3(1):71–86
5. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces versus fisherfaces: recognition using class specific linear projection. *IEEE Trans Pattern Anal Mach Intell* 19(7):711–720
6. Gao Q, Liu J, Zhang H et al (2012) Enhanced fisher discriminant criterion for image recognition. *Pattern Recogn* 45(10):3717–3724
7. Li H, Jiang T, Zhang K (2006) Efficient and robust feature extraction by maximum margin criterion. *IEEE Trans Neural Netw* 17(1):157–165
8. Wang G, Shi N, Shu Y et al (2016) Embedded manifold-based kernel Fisher discriminant analysis for face recognition. *Neural Process Lett* 43(1):1–16
9. Lu GF, Zou J, Wang Y (2016) A new and fast implementation of orthogonal LDA algorithm and its incremental extension. *Neural Process Lett* 43(3):687–707

10. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326
11. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323
12. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15(6):1373–1396
13. Bengio Y, Païement JF, Vincent P et al (2004) Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Adv Neural Inf Process Syst* 16:177–184
14. He X, Yan S, Hu Y et al (2005) Face recognition using Laplacianfaces. *IEEE Trans Pattern Anal Mach Intell* 27(3):328–340
15. He X, Cai D, Yan S et al (2005) Neighborhood preserving embedding[C]//Computer Vision, 2005. In: ICCV 2005. Tenth IEEE international conference on IEEE, vol 2, pp 1208–1213
16. Wang X, Liu Y, Nie F, Huang H (2015) Discriminative unsupervised dimensionality reduction. In: *IJCAI*, pp 3925–3931
17. Yan S, Xu D, Zhang B et al (2007) Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell* 29(1):40–51
18. Nie F, Xu D, Tsang IWH, Zhang C (2010) Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction. *IEEE Trans Image Process* 19(7):1921–1932
19. Wang R, Nie F, Hong R, Chang X, Yang X, Yu W (2017) Fast and orthogonal locality preserving projections for dimensionality reduction. *IEEE Trans Image Process* 26(10):5019–5030
20. Wang R, Nie F, Yang X, Gao F, Yao M (2015) Robust 2DPCA with non-greedy L1-norm maximization for image analysis. *IEEE Trans Cybern* 45(5):1108–1112
21. Yang J, Zhang D, Yang J et al (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
22. Nie F, Shiming X, Changshui Z (2007) Neighborhood MinMax Projections. In: *IJCAI*
23. Zhang D, He J, Zhao Y et al (2014) Global plus local: a complete framework for feature extraction and recognition. *Pattern Recogn* 47(3):1433–1442
24. Gao Q, Liu J, Zhang H et al (2013) Joint global and local structure discriminant analysis. *IEEE Trans Inf Forensics Secur* 8(4):626–635
25. Zang F, Zhang J, Pan J (2012) Face recognition using elasticfaces. *Pattern Recogn* 45(11):3866–3876
26. Luo T, Hou C, Yi D et al (2016) Discriminative orthogonal elastic preserving projections for classification. *Neurocomputing* 179:54–68
27. Yuan S, Mao X (2018) Exponential elastic preserving projections for facial expression recognition. *Neurocomputing* 275:711–724
28. Shojaeilangari S, Yau WY, Nandakumar K et al (2015) Robust representation and recognition of facial emotions using extreme sparse learning. *IEEE Trans Image Process* 24(7):2140–2152
29. Zhang X, Pham DS, Venkatesh S et al (2015) Mixed-norm sparse representation for multi view face recognition. *Pattern Recogn* 48(9):2935–2946
30. Liu Z, Pu J, Xu M et al (2015) Face recognition via weighted two phase test sample sparse representation. *Neural Process Lett* 41(1):43–53
31. Qiao L, Chen S, Tan X (2010) Sparsity preserving projections with applications to face recognition. *Pattern Recogn* 43(1):331–341
32. Gui J, Sun Z, Jia W et al (2012) Discriminant sparse neighborhood preserving embedding for face recognition. *Pattern Recogn* 45(8):2884–2893
33. Lei YK, Han H, Hao X (2015) Discriminant sparse local spline embedding with application to face recognition. *Knowl Based Syst* 89:47–55
34. Yin F, Jiao LC, Shang F et al (2014) Sparse regularization discriminant analysis for face recognition. *Neurocomputing* 128:341–362
35. Nie F, Huang H, Cai X, Ding CH (2010) Efficient and robust feature selection via joint L2, 1-norms minimization. In: *Advances in neural information processing systems*, pp 1813–1821
36. Nie F, Wang H, Deng C, Gao X, Li X, Huang H (2016) New L1-norm relaxations and optimizations for graph clustering. In: *AAAI*, pp 1962–1968
37. Deng C, Lv Z, Liu W, Huang J, Tao D, Gao X (2015) Multi-view matrix decomposition: a new scheme for exploring discriminative information. In: *IJCAI*, pp 3438–3444
38. Nie F, Wang X, Huang H (2014) Clustering and projected clustering with adaptive neighbors. In: *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp 977–986
39. The Olivetti & Oracle Research Laboratory Face Database of Faces. <http://www.cam-orl.co.uk/facedatabase.html>



40. Sim T, Baker S, Bsat M (2002) The CMU pose, illumination, and expression (PIE) database[C]//Automatic Face and Gesture Recognition, 2002. In: Proceedings of the fifth IEEE international conference on IEEE, pp 46–51
41. Phillips PJ, Wechsler H, Huang J et al (1998) The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis Comput* 16(5):295–306
42. Huang GB et al (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Vol. 1. No. 2. Technical Report 07-49, University of Massachusetts, Amherst

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.