Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 5269236, 12 pages
http://dx.doi.org/10.1155/2016/5269236

Research Article

Sparsity Preserving Discriminant Projections with Applications to Face Recognition

Yingchun Ren,[1,2] Zhicheng Wang,[1] Yufei Chen,[1] and Weidong Zhao[1]

[1]*Research Center of CAD, Tongji University, Shanghai 201804, China*
[2]*College of Mathematics, Physics and Information Engineering, Jiaxing University, Jiaxing 314001, China*

Correspondence should be addressed to Zhicheng Wang; zhichengwang@tongji.edu.cn

Received 22 July 2015; Accepted 10 December 2015

Academic Editor: Joaquim Joao Judice

Dimensionality reduction is extremely important for understanding the intrinsic structure hidden in high-dimensional data. In recent years, sparse representation models have been widely used in dimensionality reduction. In this paper, a novel supervised learning method, called Sparsity Preserving Discriminant Projections (SPDP), is proposed. SPDP, which attempts to preserve the sparse representation structure of the data and maximize the between-class separability simultaneously, can be regarded as a combiner of manifold learning and sparse representation. Specifically, SPDP first creates a concatenated dictionary by classwise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary using the least square method. Secondly, a local between-class separability function is defined to characterize the scatter of the samples in the different submanifolds. Then, SPDP integrates the learned sparse representation information with the local between-class relationship to construct a discriminant function. Finally, the proposed method is transformed into a generalized eigenvalue problem. Extensive experimental results on several popular face databases demonstrate the feasibility and effectiveness of the proposed approach.

1. Introduction

In many fields such as object recognition [1, 2], text categorization [3], and information retrieval [4], the data are usually provided in high-dimensional form; this makes it difficult to describe, understand, and recognize these data. As an effective method, dimensionality reduction has been widely used in practice to handle these problems [5–8]. Up to now, a variety of dimensionality reduction algorithms have been designed. Based on the data structure they utilize, these methods fall into three categories: global structure-based methods, local neighborhood-based methods, and sparse representation-based methods.

Principal Component Analysis (PCA) [9], Linear Discriminant Analysis (LDA) [10], and their kernelized versions are typical global structure-based methods [11, 12]. Owing to its simplicity and effectiveness, PCA, which aims at maximizing the variance of the projected data, has extensive applications in the fields of science and engineering. PCA is a good dimensionality reduction method; however, it does not employ the label information of the samples, leading to inefficiency of the classification. Unlike PCA, LDA is a supervised method that attempts to identify an optimal projection by maximizing the between-class scatter and as such minimizing the within-class scatter. Because the label information is fully exploited, LDA has been proven more efficient than PCA in classification [13]. However, LDA can extract at best $K - 1$ features (K is the number of categories), which is unacceptable in many situations. Moreover, both PCA and LDA are based on the hypothesis that samples from each class lie on a linear subspace [14, 15]; that is, neither of them can identify the local submanifold structure hidden in high-dimensional data.

Recently, manifold learning methods, which are especially useful for the analysis of the data that lie on a submanifold of the original space, have been proposed [16–26]. Representative manifold learning methods include Isomap [16], Laplacian Eigenmaps (LE) [17], and Locally Linear Embedding (LLE) [18]. All these nonlinear methods are able to discover the optimal feature subspace by solving an

optimization problem based on the weight graph question; however, none of them can overcome the "out-of-sample" problem [19]. That is, they yield maps that are characterized only on the training data points but how to evaluate the maps on new test data points is still unclear. In order to address this problem, Cai et al., respectively, developed the linear visions of the above manifold learning methods such as isometric projection [20], Locality Preserving Projections (LPP) [21], and Neighborhood Preserving Embedding (NPE) [22]. However, these methods suffer from a limitation that they do not encode discriminant information, which is very important for recognition tasks. Recently, Gui et al. proposed a new supervised learning algorithm called Locality Preserving Discriminant Projections (LPDP) to improve the classification performance of LPP and applied it to face recognition [26]. Experimental results show that LPDP is more suitable for recognition tasks than LPP.

Sparse representation, as a new branch of the state-of-the-art techniques for signal representation, has attracted considerable research interests [27–38]. It attempts to preserve the sparse representation structure of the samples in a low-dimensional embedding subspace. The representative dimensionality reduction algorithms based on sparse representation include Sparsity Preserving Projections (SPP) [39], Sparsity Preserving Discriminant Analysis (SPDA) [40], Discriminative Learning by Sparse Representation Projections (DLSP) [41], Sparse Tensor Discriminant Analysis (STDA) [42], and sparse nonnegative matrix factorization [43]. It is worthwhile to note that a sparse model also depends on the subspace assumption: each sample can be linearly expressed by other samples from the same class; that is, each sample can be sparsely recovered by samples from all classes. In general, these sparse learning algorithms provide superior recognition accuracy compared with the conditional methods. However, all these dimensionality reduction methods based on sparse coding mentioned above are required to solve the ℓ_1 norm minimization problem to construct the sparse weight matrix. Therefore, they are computationally prohibitive for large-scale problems. For example, SPP attempts to preserve the sparse reconstructive relationship of the data [39], which is an effective and powerful technique for dimensionality reduction. However, the computational complexity of SPP is excessively high and hence, it cannot be used extensively for large-scale data processing (in fact, the time cost for constructing the sparse weight graph is $O(n^4)$, where n indicates the total number of training samples). Moreover, SPP does not absorb the label information. Thus, the algorithm is unsupervised.

Motivated by the above works, a novel supervised learning method, called Sparsity Preserving Discriminant Projection (SPDP), is proposed in this paper. By integrating SPP with local discriminant information for dimensionality reduction, SPDP can be viewed as a combiner of sparse representation and manifold learning. Because sparse representation can implicitly discover the local structure of the data owing to the sparsity prior, this property can be used to describe the local structure. However, differing from the existing SPP, which is time-consuming in sparse reconstruction for each test sample, SPDP first creates a concatenated dictionary using classwise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary quickly with the least square method. Then, a local between-class separability function is defined to characterize the scatter of the samples in the different submanifolds. Subsequently, by integrating the sparse representation information with the local between-class relationship, SPDP attempts to preserve the sparse representation structure of the data and maximize the local between-class separability simultaneously. Finally, the proposed method is converted into a generalized eigenvalue problem.

It is worth emphasizing some merits of SPDP and the main contributions of this paper:

(1) SPDP is a supervised dimensionality reduction method that attempts to identify a discriminating subspace where the sparse representation structure of the data and the label information are maintained. Meanwhile, the separability of different submanifolds is maximized; that is, different submanifolds can be distinguished more clearly.

(2) SPDP is able to explore the local submanifold structure hidden in high-dimensional data because the manifold learning is employed to characterize the local between-class separability.

(3) The time required for extracting discriminant vectors in SPDP is significantly less than many algorithms based on sparse representation. Therefore, the proposed method can be widely applied for large-scale problems.

(4) Label information is employed twice in SPDP. First, it is absorbed in constructing the dictionary for sparse representation and calculating the sparse coefficient vector, which may contribute to a more discriminating sparse representation structure. Further, it is utilized in computing the local between-class separability, which is more conducive for classification.

The rest of this paper is organized as follows: Section 2 briefly reviews the existing SPP algorithm. The SPDP algorithm is described in detail in Section 3. The experimental results and analysis are presented in Section 4 and the paper ends with concluding remarks in Section 5.

2. Brief Review of Sparsity Preserving Projections (SPP)

SPP aims to preserve the sparse reconstruction relationship of the samples [39]. Given a set of training samples $\{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbf{R}^m$ and n is the number of training samples, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbf{R}^{m\times n}$ be the data matrix consisting of all the training samples. SPP first seeks the sparse reconstruction coefficient vector $\mathbf{s}_i$ for each sample $\mathbf{x}_i$ through the following modified ℓ_1 minimization problem:

$$\begin{aligned} \min_{\mathbf{s}_i} \quad & \|\mathbf{s}_i\|_1, \\ \text{s.t.} \quad & \mathbf{x}_i = \mathbf{X}\mathbf{s}_i, \\ & \mathbf{1} = \mathbf{1}^T\mathbf{s}_i, \end{aligned} \tag{1}$$

where $\mathbf{s}_i = [\mathbf{s}_{i1}, \ldots, \mathbf{s}_{i,i-1}, 0, \mathbf{s}_{i,i+1}, \ldots, \mathbf{s}_{i,n}]^T$ is an n-dimensional column vector in which the ith element is equal to zero, implying $\mathbf{x}_i$ is removed from $\mathbf{X}$, and the element $\mathbf{s}_{ij}$, $j \neq i$, denotes the contribution of $\mathbf{x}_j$ for reconstructing $\mathbf{x}_i$. Then, the sparse reconstructive weight matrix $\mathbf{S}$ is given as follows:

$$\mathbf{S} = \left[\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\right], \tag{2}$$

where $\mathbf{s}_i$ is the optimal solution of (1). The final optimal projection vector $\mathbf{w}$ is obtained through the following maximization problem:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{X} \mathbf{S}_\beta \mathbf{X}^T \mathbf{w}}{\mathbf{w}^T \mathbf{X} \mathbf{X}^T \mathbf{w}}, \tag{3}$$

with $\mathbf{S}_\beta = \mathbf{S} + \mathbf{S}^T - \mathbf{S}^T \mathbf{S}$. This problem transforms to a generalized eigenvalue problem.

It follows that SPP must resolve n time-consuming ℓ_1 norm minimization problems to obtain the sparse weight matrix $\mathbf{S}$. Thus, the computational complexity of SPP is excessively high and therefore not widely applicable to large-scale data processing. Moreover, SPP does not exploit the prior knowledge of class information, which is valuable for classification and recognition problems such as face recognition.

3. Sparsity Preserving Discriminative Learning

In this section, the proposed SPDP algorithm is described in more detail. To reduce the disadvantage that is inevitable for SPP to resolve n time-consuming ℓ_1 norm minimization problems to obtain the sparse weight matrix $\mathbf{S}$, SPDP first constructs a concatenated dictionary through classwise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary quickly using the least square method. To enhance the discriminant performance, it defines a local between-class separability function to characterize the scatter of the samples in the different submanifolds. Then, by integrating the sparse representation information with the local interclass relationship, SPDP aims to maximize the separation between the submanifolds (or intrinsic clusters) without destroying localities and meanwhile preserve the sparse representation structure of the data. Hence, the proposed algorithm is expected to preserve the intrinsic geometry structure and have superior discriminant abilities.

3.1. Constructing the Concatenated Dictionary. For convenience, we first provide some notations used in this paper. Assume that $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a set of training samples, where $\mathbf{x}_i \in \mathbf{R}^m$. We can categorize the training samples as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$, where $\mathbf{X}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}] \in \mathbf{R}^{m \times n_i}$ $(i = 1, 2, \ldots, K)$ consists of samples from class i. Suppose that samples from a single class lie on a linear subspace. Thus, each sample can be sparse linearly represented by samples from all classes. The subspace model is a powerful tool to capture the underlying information in real data sets [44]. For the convenience of PCA decomposition and relevant calculations, we first center the samples from each class at the origin, $\widetilde{\mathbf{X}}_i = [\mathbf{x}_{i1} - \boldsymbol{\mu}_i, \mathbf{x}_{i2} - \boldsymbol{\mu}_i, \ldots, \mathbf{x}_{in_i} - \boldsymbol{\mu}_i]$ $(i = 1, 2, \ldots, K)$, where $\boldsymbol{\mu}_i$ denotes the mean of class i; that is, $\boldsymbol{\mu}_i = \sum_{i=1}^{n_i} \mathbf{x}_i / n_i$. Therefore, the training sample can be recast as $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2, \ldots, \widetilde{\mathbf{X}}_K]$. Afterwards, PCA decomposition is conducted for every $\widetilde{\mathbf{X}}_i$ $(i = 1, 2, \ldots, K)$, whose objective function is

$$\max_{\|\mathbf{d}\|=1} \mathbf{d}^T \sum_i \mathbf{d}, \tag{4}$$

where $\sum_i$ is the sample covariance matrix of $\widetilde{\mathbf{X}}_i$. For every class i, the first l_i principal components are selected to construct $\mathbf{D}_i = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{l_i}]$ (in fact, l_i is automatically selected by the value of the PCA ratio from the system). Thus, a sample $\mathbf{x}$ from class i can be simply represented as

$$\begin{aligned} \mathbf{x} &= \mathbf{D}_i \widetilde{\mathbf{s}}_i = \left[\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{i-1}, \mathbf{D}_i, \mathbf{D}_{i+1}, \ldots, \mathbf{D}_K\right] \mathbf{s} \\ &= \mathbf{D}\mathbf{s}, \end{aligned} \tag{5}$$

with $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K]$ and $\mathbf{s} = [\mathbf{0}^T, \mathbf{0}^T, \ldots, \mathbf{0}^T, \widetilde{\mathbf{s}}_i^T, \mathbf{0}^T, \ldots, \mathbf{0}^T]^T$. $\mathbf{D}_i$ is the dictionary of class i by the PCA decomposition above, $\mathbf{D}$ is the concatenated dictionary composed of all $\mathbf{D}_i$ $(i = 1, 2, \ldots, K)$, $\mathbf{s}$ is the sparse representation of a sample $\mathbf{x}$ under the concatenated dictionary $\mathbf{D}$, and $\widetilde{\mathbf{s}}_i$ is the coefficient vector under the dictionary $\mathbf{D}_i$. In fact, $\widetilde{\mathbf{s}}_i$ can be quickly computed from the least square method as

$$\widetilde{\mathbf{s}}_i = \left(\mathbf{D}_i^T \mathbf{D}_i\right)^{-1} \mathbf{D}_i^T \mathbf{x} = \mathbf{D}_i^T \mathbf{x}. \tag{6}$$

The orthogonality of each principal component of PCA decomposition of the same class is utilized in the reduction of the above formula. The process of constructing the concatenated dictionary is presented in Figure 1.

According to the preceding procedure, each training sample corresponds to a sparse representation under the concatenated dictionary $\mathbf{D}$ and the sparse coefficient vector $\mathbf{s}$ of any training sample from class i can be quickly computed from the least square method (in fact, it is the primary reason that the proposed approach is significantly faster than SPP, which will be explained in detail in Section 4.4) because the computational process of $\mathbf{s}$ involves only $\mathbf{D}_i$, which is column orthogonal in view of (5) and (6).

3.2. Preserving Sparse Representation Structure. As can be seen in Section 3.1, to some extent, the dictionary $\mathbf{D}$ describes the intrinsic geometric properties of the data and the sparse coefficient vectors explicitly encode the discriminant information of the training samples. Thus, it is hoped that this valued property in the original high-dimensional space can be preserved in the low-dimensional embedding subspace. Therefore, the objective function is expected to look for an optimal projection that can best preserve the sparse representation structure:

$$J_s(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^{n} \left\| \mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{D} \mathbf{s}_i \right\|_2^2, \tag{7}$$

where $\mathbf{s}_i$ is the sparse reconstruction vector corresponding to $\mathbf{x}_i$.

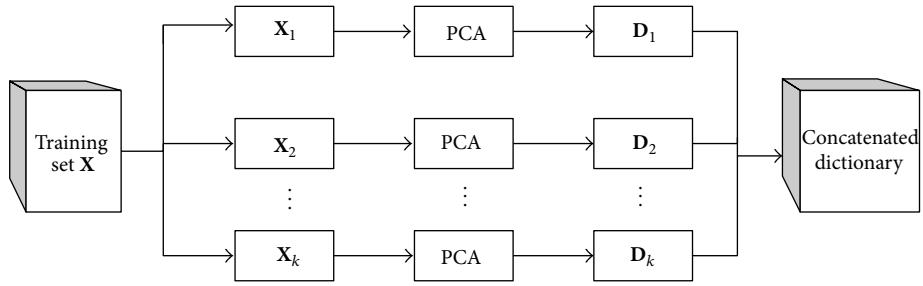

Figure 1: The process of constructing the concatenated dictionary.

Using algebraic operations, (7) can be arranged as

$$\begin{aligned}
&\sum_{i=1}^{n} \left\| \mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{D} \mathbf{s}_i \right\|_2^2 \\
&= \mathbf{w}^T \left(\sum_{i=1}^{n} \left(\mathbf{x}_i - \mathbf{D}\mathbf{s}_i \right) \left(\mathbf{x}_i - \mathbf{D}\mathbf{s}_i \right)^T \right) \mathbf{w} \\
&= \mathbf{w}^T \sum_{i=1}^{n} \left(\mathbf{x}_i \mathbf{x}_i{}^T - \mathbf{x}_i \mathbf{s}_i{}^T \mathbf{D}^T - \mathbf{D}\mathbf{s}_i \mathbf{x}_i{}^T + \mathbf{D}\mathbf{s}_i \left(\mathbf{D}\mathbf{s}_i \right)^T \right) \mathbf{w} \\
&= \mathbf{w}^T \left(\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T \mathbf{D}^T - \mathbf{D}\mathbf{S}\mathbf{X}^T + \mathbf{D}\mathbf{S}\mathbf{S}^T \mathbf{D}^T \right) \mathbf{w},
\end{aligned} \tag{8}$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n]$, and therefore, (7) can be simply recast as

$$\begin{aligned}
&J_s(\mathbf{w}) \\
&= \min_{\mathbf{w}} \mathbf{w}^T \left(\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T \mathbf{D}^T - \mathbf{D}\mathbf{S}\mathbf{X}^T + \mathbf{D}\mathbf{S}\mathbf{S}^T \mathbf{D}^T \right) \mathbf{w}.
\end{aligned} \tag{9}$$

3.3. Characterization of the Local Interclass Separability. To effectively discover the discriminant structure embedded in high-dimensional data and improve the classification performance, in this subsection, we construct a local interclass weight graph. Because data in the same class lie on one or more submanifolds and data belonging to different classes are distributed on different submanifolds, it is important for classification problems to distinguish one submanifold from another. Therefore, a local between-class separability function is defined in this section to characterize the separability of the samples in different submanifolds. The aim of SPDP is that different submanifolds can be distinguished more clearly after being projected; hence, the local between-class separability of different submanifolds should be maximized. Thus, we can construct a label matrix $\mathbf{B}$ to describe the local and interclass relationships of each point as follows:

$$\begin{aligned}
&\mathbf{B}_{ij} \\
&= \begin{cases} 1 + \exp\left(-\dfrac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2^2}{\sigma} \right), & \text{if } i \in N_K^-(j) \text{ or } j \in N_K^-(i); \\ 0, & \text{otherwise,} \end{cases}
\end{aligned} \tag{10}$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ denotes the geodesic distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$, σ is a parameter which is often set to be as the standard deviation of the samples, $N_K^-(i)$ denotes the index in the K nearest neighbors of the sample $\mathbf{x}_i$, however with a different class label, and $\mathbf{B}$ is called the local between-class weight matrix (or local interclass weight graph). As can be seen in the above definition, if two distant points $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different submanifolds, the scatter of them is big and vice versa. That is, the points belonging to different submanifolds should be located farther after projection. Therefore, the local interclass separability can be characterized as the following equation:

$$J_b(\mathbf{w}) = \frac{1}{2} \sum_i \sum_j \left\| \mathbf{y}_i - \mathbf{y}_j \right\|_2^2 \mathbf{B}_{ij}, \tag{11}$$

where $\mathbf{y}_i = \mathbf{w}^T \mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ is the low-dimensional representation of the original data, which can be obtained by projecting each $\mathbf{x}_i$ onto the direction vector $\mathbf{w} \in \mathbf{R}^m$. With algebraic simplifications, (11) can be rewritten as

$$\begin{aligned}
J_b &= \frac{1}{2} \sum_{i,j}^{n} \mathbf{B}_{ij} \left\| \mathbf{y}_i - \mathbf{y}_j \right\|_2^2 = \frac{1}{2} \\
&\cdot \mathbf{w}^T \left(\sum_{i,j}^{n} \mathbf{B}_{ij} \left(\mathbf{x}_i - \mathbf{x}_j \right)^T \left(\mathbf{x}_i - \mathbf{x}_j \right) \right) \mathbf{w} = \mathbf{w}^T \left(\frac{1}{2} \right. \\
&\left. \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}_{ij} \left(\mathbf{x}_i - \mathbf{x}_j \right)^T \left(\mathbf{x}_i - \mathbf{x}_j \right) \right) \mathbf{w} \\
&= \mathbf{w}^T \left(\frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}_{ij} \mathbf{x}_i \mathbf{x}_i{}^T - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}_{ij} \mathbf{x}_i \mathbf{x}_j{}^T \right. \right. \\
&\left. \left. + \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}_{ij} \mathbf{x}_j \mathbf{x}_j{}^T \right) \right) \mathbf{w} = \mathbf{w}^T \left(\sum_{i=1}^{n} \mathbf{D}'_{ii} \mathbf{x}_i \mathbf{x}_i{}^T \right. \\
&\left. - \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}_{ij} \mathbf{x}_i \mathbf{x}_j{}^T \right) \mathbf{w} = \mathbf{w}^T \left(\mathbf{X} \left(\mathbf{D}' - \mathbf{B} \right) \mathbf{X}^T \right) \mathbf{w} \\
&= \mathbf{w}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{w},
\end{aligned} \tag{12}$$

where $\mathbf{L}$ is Laplacian matrix with definition $\mathbf{L} = \mathbf{D}' - \mathbf{B}$ and $\mathbf{D}'$ is a diagonal matrix [45]; that is, $\mathbf{D}'_{ii} = \sum_j \mathbf{B}_{ij}$. Equation (12) characterizes the separability (or scatter) of the data set in different submanifolds. Therefore, each manifold can be

separated clearly, as long as the optimal projection $\mathbf{w}^*$ is adopted.

3.4. Sparsity Preserving Discriminant Projections. To achieve improved recognition results, we explicitly integrate the sparsity preserving constraint as indicated in (7) with the local between-class separability as illustrated in (12). The novel supervised algorithm SPDP, which not only preserves the sparse representation structure but also separates each submanifold as distant as possible, is defined as

$$\begin{aligned}\max_{\mathbf{w}} J(\mathbf{w}) &= \frac{J_b(\mathbf{w})}{J_s(\mathbf{w})} \\ &= \frac{\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\left(\mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T\mathbf{D}^T - \mathbf{D}\mathbf{S}\mathbf{X}^T + \mathbf{D}\mathbf{S}\mathbf{S}^T\mathbf{D}^T\right)\mathbf{w}},\end{aligned} \tag{13}$$

where the denominator term $J_s(\mathbf{w})$ measures the quality of preserving the sparse representation structure and the numerator term $J_b(\mathbf{w})$ measures the separability of different submanifolds. It is well known that the criterion of LDA is to maximize the between-class scatter and, meanwhile, minimize the within-class scatter. Similar to LDA, the aim of SPDP is to maximize the ratio of the local between-class separability to the sparse representation weight scatter.

Letting

$$\mathbf{M} = \mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{S}^T\mathbf{D}^T - \mathbf{D}\mathbf{S}\mathbf{X}^T + \mathbf{D}\mathbf{S}\mathbf{S}^T\mathbf{D}^T, \tag{14}$$

the objective function can be recast as the following optimization problem:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}}{\mathbf{w}^T\mathbf{M}\mathbf{w}}. \tag{15}$$

Then, the optimal $\mathbf{w}$'s are the eigenvectors corresponding to the largest d eigenvalues of the following generalized eigenvalue problem:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{M}\mathbf{w}. \tag{16}$$

It is worth noting that since the training sample size is much smaller than the feature dimensions for those high-dimensional data, $\mathbf{M}$ might be singular. This problem can be tackled by projecting the training set $\mathbf{X}$ onto a PCA subspace spanned by the leading eigenvectors to get $\mathbf{X}'$ and replacing $\mathbf{X}$ by $\mathbf{X}'$.

Based on the above discussion, the proposed SPDP is summarized in Algorithm 1.

Algorithm 1 (Sparsity Preserving Discriminant Projections (SPDP)). We have the following steps.

Step 1. Execute PCA decomposition for each $\mathbf{X}_i$ ($i = 1, 2, \ldots, K$) using (4) to obtain the concatenated dictionary $\mathbf{D}$.

Step 2. Calculate the coefficient vector $\tilde{\mathbf{s}}_i$ under the dictionary $\mathbf{D}_i$ for each sample based on (6) to obtain the sparse coefficient vector $\mathbf{s}$ and then calculate $\mathbf{S}$.

Step 3. Calculate $\mathbf{B}$ and $\mathbf{L}$ by (10) and (12), respectively.

Step 4. Calculate the projecting vectors by the generalized eigenvalue problem in (16).

Figure 2: Some face samples from the Yale database.

4. Experiments

In this section, the proposed SPDP algorithm is tested on three publicly available face databases (Yale [13], ORL [46], and CMU PIE [47]) and compared with six popular dimensionality reduction methods—PCA, LDA, LPP, NPE, LPDP, and SPP. For PCA, the only model parameter is the subspace dimension and for LDA, the performance is directly influenced by the energy of the eigenvalues kept in the PCA preprocessing phase. For LPP and NPE, the supervised versions are adopted. In particular, the neighbor mode in LPP and NPE is set to be "supervised"; the weight mode in LPP is set to be "Cosine." The empirically determined parameter α in LPDP is taken to be 1 [26], ε in SPP is set to be 0.05 as indicated in [39], and σ in SPDP is set to be the standard deviation of the samples. The nearest neighbor classifier ($1-NN$) is employed to predict the classes of the test data. All experiments are accomplished with MATLAB R2013a on a personal computer with Intel(R) Core i7-4770 K 3.50 GHz CPU, 16.0 GB main memory, and the Windows 7 operating system.

4.1. Experiment on Yale Face Database. The Yale face database contains 165 face images of 15 individuals. There are 11 images per individual. These images were collected under different facial expressions (normal, happy, sad, surprised, sleepy, and wink) and configurations (left-light, center-light, and right-light) and with or without glasses. All the images are cropped to a size of 32×32 and then normalized to have a unit norm. Some samples from this database are presented in Figure 2. For each person, k (k varies from 2 to 8) images are randomly selected as the training samples and the remaining $11 - k$ for the test. For each k, the results are averaged over 50 random splits. Table 1 presents the best recognition rate and the associated standard deviation of the seven algorithms under the different sizes of the training set. Figure 3(a) presents the best recognition rate versus the variation of the size of the training set. Figure 3(b) is the variation rules of the recognition rates of the seven algorithms under different reduced dimensions when the size of the training samples from each class is fixed as six. The fact that the upper bound for the dimensionality of LDA is $K - 1$ (K is the number of categories) because there are at most $K - 1$ generalized nonzero eigenvalues [13] deserves to be noted; similar situations will occur in other experiments in this paper. Hence, one can see that the SPDP algorithm significantly outperforms the other methods.

4.2. Experiment on ORL Face Database. There are 400 images of 40 people in the ORL face data set, where each one has 10 different pictures. The images were collected at different time

Table 1: The best recognition rate and the corresponding standard deviation of the seven algorithms under the different size of the training set on Yale (k is the training sample size).

Methods	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
SPDP	0.7137 (±0.031)	0.8333 (±0.036)	0.9125 (±0.027)	0.9458 (±0.031)	0.9619 (±0.025)	0.9786 (±0.029)	0.9916 (±0.033)
LPDP	0.5674 (±0.057)	0.7175 (±0.045)	0.7807 (±0.037)	0.8186 (±0.039)	0.8679 (±0.035)	0.8816 (±0.029)	0.9066 (±0.036)
PCA	0.4389 (±0.027)	0.4895 (±0.035)	0.5514 (±0.037)	0.5838 (±0.048)	0.6241 (±0.038)	0.6561 (±0.043)	0.6727 (±0.046)
LDA	0.5354 (±0.061)	0.6486 (±0.052)	0.7222 (±0.036)	0.7792 (±0.047)	0.8132 (±0.037)	0.8375 (±0.040)	0.8613 (±0.044)
LPP	0.5783 (±0.041)	0.6814 (±0.044)	0.7469 (±0.036)	0.8025 (±0.035)	0.8139 (±0.027)	0.8244 (±0.014)	0.8392 (±0.018)
NPE	0.5635 (±0.025)	0.6811 (±0.019)	0.7455 (±0.027)	0.7593 (±0.023)	0.8112 (±0.017)	0.8284 (±0.025)	0.8463 (±0.023)
SPP	0.5202 (±0.038)	0.6425 (±0.027)	0.7098 (±0.033)	0.7471 (±0.033)	0.7653 (±0.026)	0.7827 (±0.032)	0.8037 (±0.035)

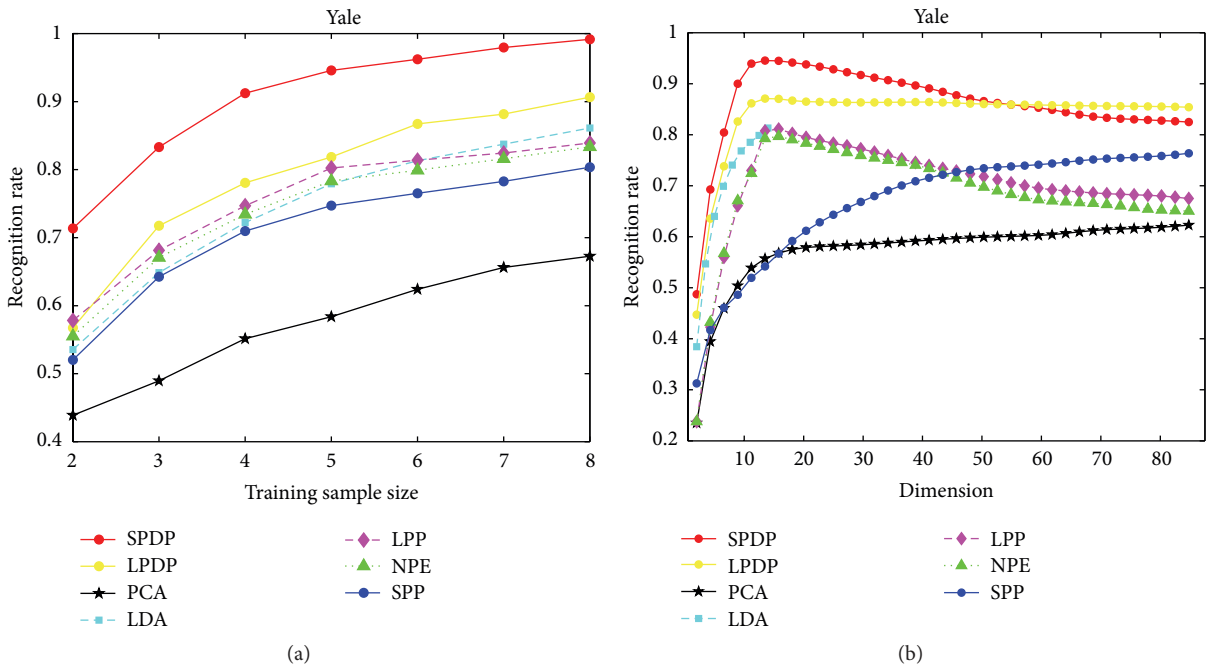

Figure 3: Recognition rates of the seven algorithms on the Yale database: (a) the best recognition rates versus the different size of the training set and (b) the average recognition rates versus the variation of dimensions when the size per class is fixed as six.

points, under different lighting conditions, varying facial expressions. In our experiment, each image is cropped to a resolution of 32×32 as shown in Figure 4. We randomly select k (k varies from 2 to 8) pictures from each person for training; the remainder are used for testing. We repeat these splits 50 times and report the average results. Table 2 displays the best classification accuracy of the seven algorithms under the different sizes of the training set; the number in parentheses is the corresponding standard deviation. Figure 5(a) presents the best recognition rate versus the variation of the size of the training set. Figure 5(b) is the variation rules of the recognition rates of the seven algorithms under different reduced dimensions when the size of the training samples from each class is fixed as five. It can be seen that SPDP and LPDP are superior to other compared methods (their performances on the ORL database are quite similar), especially when the size of the training set is small. The reason may be that both SPDP and LPDP consider the discriminant information and local structure of the data.

4.3. Experiment on CMU PIE Face Database. In this subsection, it is verified that the proposed algorithm achieves higher classification accuracy than the other dimensionality reduction methods under varying illumination, pose, and expression. The CMU PIE face database contains over 41,368 face images of 68 subjects that were captured by 13 synchronized cameras and 21 flashes under varying poses, illumination, and expression. In our experiments, we choose the five frontal poses (C05, C07, C09, C27, and C29). This leaves 170 face images per subject; all the images are cropped to 32 × 32. Figure 6 shows some pictures of one subject. A

Figure 4: Some face samples from the ORL database.

random subset with k(=5, 10, 15, 20) pictures per subject is selected with labels to form the training set; the remainder are used for testing. For each given k, we average the classification accuracies over 50 random splits. Table 3 presents the best recognition rate and the associated standard deviation in brackets of the seven algorithms under the different size of the training set. Figure 7(a) presents the best recognition rate versus the variation of the size of the training set. Figure 7(b) is the variation rules of the recognition rates of the seven algorithms under different reduced dimensions when the size of the training samples from each class is fixed as ten. We can observe that the proposed SPDP outperforms the other dimensionality reduction methods such as PCA, LDA, LPP, NPE, LPDP, and SPP about pose, illumination, and expression variations.

4.4. Comparison of Time Cost for Acquiring the Discriminant Vectors of SPP with SPDP. In this subsection, the time cost for acquiring the discriminant vectors of SPDP is compared with that of SPP. Tables 4, 5, and 6 list the average time costs for acquiring the discriminant vectors of SPP and SPDP versus the different sizes of the training set on the three face data sets. It is demonstrated that SPDP is significantly faster than SPP in acquiring the embedding functions in our experiments, especially in the large-scale problems such as CMU PIE.

Table 2: The best recognition rate and the corresponding standard deviation of the seven algorithms under the different size of the training set on ORL (k is the training sample size).

Methods	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
SPDP	0.8283 (±0.023)	0.8939 (±0.019)	0.9266 (±0.022)	0.9523 (±0.015)	0.9686 (±0.021)	0.9802 (±0.015)	0.9839 (±0.017)
LPDP	0.8109 (±0.032)	0.9006 (±0.017)	0.9305 (±0.015)	0.9579 (±0.010)	0.9772 (±0.009)	0.9863 (±0.013)	0.9906 (±0.015)
PCA	0.6709 (±0.026)	0.7576 (±0.035)	0.8204 (±0.036)	0.8626 (±0.023)	0.8866 (±0.027)	0.9045 (±0.033)	0.9116 (±0.026)
LDA	0.7241 (±0.021)	0.8276 (±0.022)	0.8958 (±0.031)	0.9231 (±0.027)	0.9360 (±0.033)	0.9465 (±0.041)	0.9563 (±0.044)
LPP	0.7833 (±0.023)	0.8657 (±0.019)	0.9060 (±0.016)	0.9289 (±0.022)	0.9432 (±0.027)	0.9527 (±0.025)	0.9546 (±0.026)
NPE	0.7869 (±0.015)	0.8689 (±0.017)	0.9047 (±0.022)	0.9331 (±0.023)	0.9469 (±0.021)	0.9565 (±0.028)	0.9584 (±0.023)
SPP	0.7324 (±0.028)	0.8055 (±0.022)	0.8442 (±0.025)	0.8704 (±0.031)	0.8935 (±0.026)	0.9162 (±0.035)	0.9397 (±0.034)

Table 3: The best recognition rate and the corresponding standard deviation of the seven algorithms under the different size of the training set on CMU PIE (k is the training sample size).

Methods	$k = 5$	$k = 10$	$k = 15$	$k = 20$
SPDP	0.7882 (±0.012)	0.9015 (±0.009)	0.9356 (±0.017)	0.9517 (±0.022)
LPDP	0.7653 (±0.012)	0.8759 (±0.015)	0.9127 (±0.009)	0.9405 (±0.011)
PCA	0.2817 (±0.024)	0.4260 (±0.032)	0.5345 (±0.021)	0.6028 (±0.028)
LDA	0.7250 (±0.011)	0.8625 (±0.016)	0.9175 (±0.008)	0.9337 (±0.015)
LPP	0.7253 (±0.033)	0.8659 (±0.028)	0.9005 (±0.039)	0.9342 (±0.023)
NPE	0.7148 (±0.035)	0.8601 (±0.029)	0.8905 (±0.021)	0.9231 (±0.028)
SPP	0.6391 (±0.026)	0.7720 (±0.031)	0.8285 (±0.018)	0.8607 (±0.034)

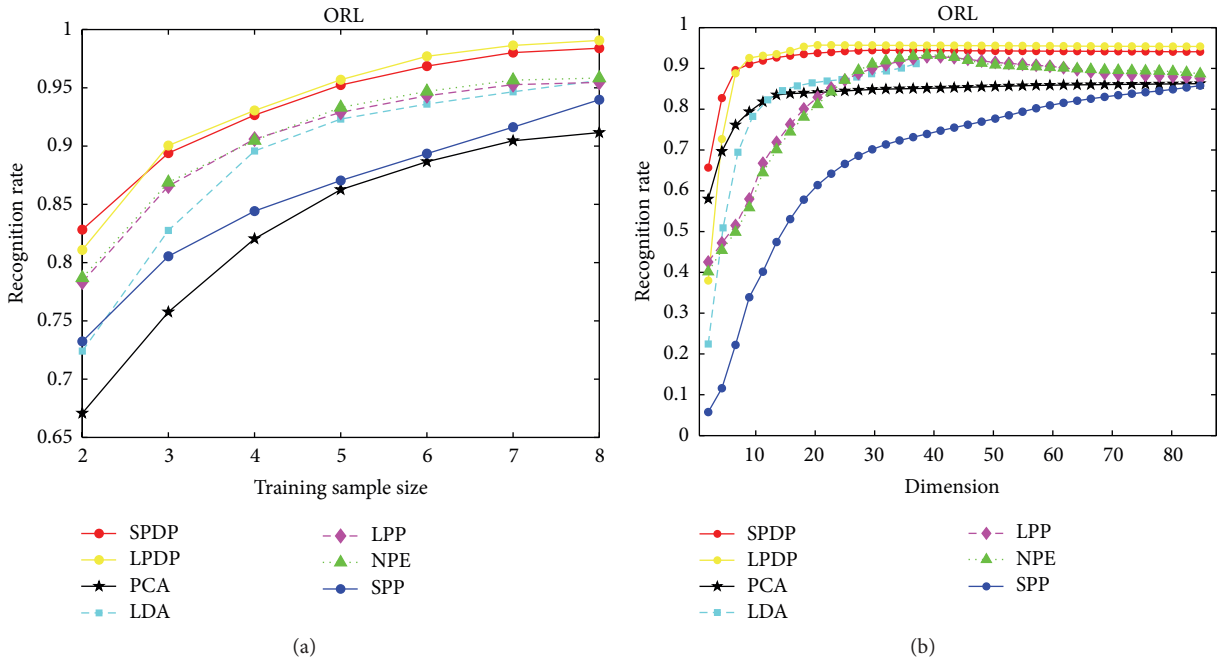

Figure 5: Recognition rates of the seven algorithms on the ORL database: (a) the best recognition rates versus the different size of the training set and (b) the average recognition rates versus the variation of dimensions when the size per class is fixed as five.

Figure 6: Some face samples from the CMU PIE database.

The critical factor of the above phenomenon is that the approaches of SPP and SPDP to obtain the sparse representation structure are entirely different. In SPP, n time-consuming ℓ_1 norm minimization problems are required to be solved to construct the sparse weight matrix, whose computational cost is $O(n^4)$ [48, 49], whereas SPDP can achieve this significantly faster through only K PCA decompositions and n least square methods. Because K PCA decompositions can be completed in $O(m^2 \sum_{i=1}^{K} l_i)$ according to the more efficient algorithm [50], the time cost for learning the sparse coefficient vector of each sample, which only involves the least square method, is $O(ml_i)$ and the sparse weight matrix $\mathbf{S}$ can be calculated with $O(m \sum_{i=1}^{K} n_i l_i)$; the computational complexity of SPDP to learn the sparse representation structure is $O(m^2 \sum_{i=1}^{K} l_i + m \sum_{i=1}^{K} n_i l_i)$. In general, $n_i \ll n$, $l_i \ll n$, and $K \ll n$; hence, SPDP performs considerably faster than SPP as indicated in Tables 4, 5, and 6.

4.5. Overall Observations and Discussions. Several observations and analysis can be achieved from the above experimental results.

(1) From Tables 1, 2, and 3 and Figures 3(a), 5(a), and 7(a), we can draw a conclusion that the proposed algorithm consistently outperforms the other compared methods, especially when the number of the training data is particularly small. The reason is that SPDP simultaneously considers both the sparse representation structure and the separability of the different submanifolds. Further, this indicates that SPDP can capture more inherent information that is hidden in the data compared to the other compared methods.

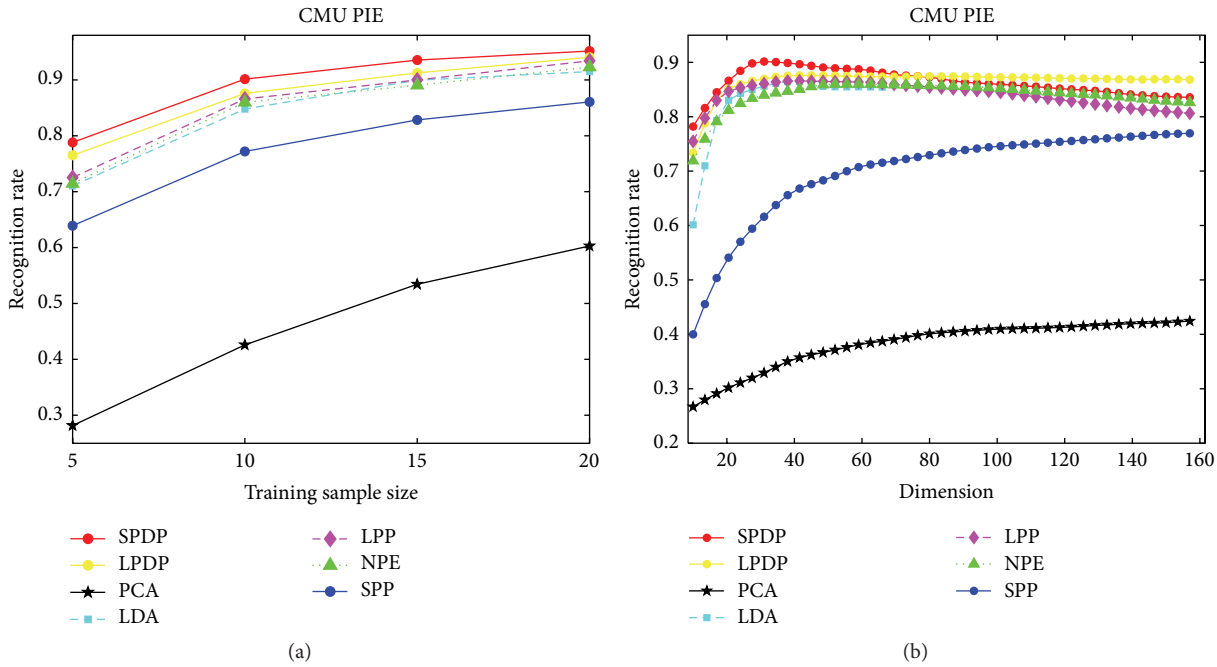

Figure 7: Recognition rates of the seven algorithms on the CMU PIE database: (a) the best recognition rates versus the different size of the training set and (b) the average recognition rates versus the variation of dimensions when the size per class is fixed as ten.

Table 4: Time (s) for acquiring the discriminant vectors of SPP and SPDP on Yale (k is the training sample size).

Methods	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
SPP	0.3729	0.6387	1.0335	1.5506	2.1609	2.9087	4.0471
SPDP	0.2475	0.4016	0.4835	0.5352	0.6827	0.7036	0.8263

Table 5: Time (s) for acquiring the discriminant vectors of SPP and SPDP on ORL (k is the training sample size).

Methods	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
SPP	1.1933	2.5641	5.1679	8.4467	13.0688	19.7787	29.4638
SPDP	0.2672	0.3526	0.5128	0.5933	0.7032	0.8158	1.2875

Table 6: Time (s) for acquiring the discriminant vectors of SPP and SPDP on CMU PIE (k is the training sample size).

Methods	$k = 5$	$k = 10$	$k = 15$	$k = 20$
SPP	40.9725	67.6737	104.9756	178.6327
SPDP	0.2736	0.8135	1.9037	2.7218

(2) From Figures 3(b), 5(b), and 7(b), it can be observed that the reduction dimensions for SPDP to achieve the best recognition rate are less than those of the other compared algorithms. This saves a considerable amount of time and storage space after obtaining the optimal embedding functions.

(3) From Tables 4, 5, and 6, it is indicated that SPDP is considerably faster than SPP in obtaining the discriminant vectors. This is because the method SPDP uses to learn the sparse representation structure which is more effective than that of SPP as analyzed in Section 4.4.

5. Conclusions

This paper proposed a new supervised learning method, called Sparsity Preserving Discriminative Projections (SPDP), by combining manifold learning and sparse representation. Specifically, SPDP first constructs a concatenated dictionary by means of classwise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary quickly using the least square method. Then, it defines a local between-class separability function to characterize the separability of the samples in different submanifolds. Subsequently, SPDP integrates the sparse representation information with the local between-class relationship. Thus, SPDP preserves the sparse representation structure of the data and maximizes the local between-class separability simultaneously. Finally,

the proposed method is transformed into a generalized eigenvalue problem. Extensive experiments on three publicly available face data sets confirmed the promising performance of the proposed SPDP approach.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61103070, 11301226); Zhejiang Provincial Natural Science Foundation of China (LQ13A010017); and the Program for Young Excellent Talents in Tongji University (2013KJ008).

References

[1] S. Gupta, R. Girshick, P. Arbelez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Computer Vision—ECCV 2014*, pp. 345–360, Springer, 2014.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[3] W. Zhang, X. Tang, and T. Yoshida, "TESC: an approach to TExt classification using semi-supervised Clustering," *Knowledge-Based Systems*, vol. 75, pp. 152–160, 2015.

[4] X. Zhao, X. Li, and Z. Zhang, "Multimedia retrieval via deep learning to rank," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1487–1491, 2015.

[5] C.-H. Li, H.-H. Ho, B.-C. Kuo, J.-S. Taur, H.-S. Chu, and M.-S. Wang, "A semi-supervised feature extraction based on supervised and fuzzy-based linear discriminant analysis for hyperspectral image classification," *Applied Mathematics & Information Sciences*, vol. 9, no. 1, pp. 81–87, 2015.

[6] D. Zhang, D. Ding, J. Li, and Q. Liu, "Pca based extracting feature using fast fourier transform for facial expression recognition," in *Transactions on Engineering Technologies*, pp. 413–424, Springer, Amsterdam, The Netherlands, 2015.

[7] J. Kalina, "Classification methods for high-dimensional genetic data," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 10–18, 2014.

[8] F. Shang, L. C. Jiao, J. Shi, and J. Chai, "Robust positive semi-definite l-isomap ensemble," *Pattern Recognition Letters*, vol. 32, no. 4, pp. 640–649, 2011.

[9] I. T. Jolliffe, *Principal Component Analysis*, Wiley Online Library, 2002.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 2013.

[11] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[12] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[15] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[16] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[18] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[19] S. Yan, D. Xu, B. Zhang, and H.-J. Zhang, "Graph embedding: a general framework for dimensionality reduction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 830–837, IEEE, June 2005.

[20] D. Cai, X. He, and J. Han, "Isometric projection," in *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI '07)*, vol. 1, pp. 528–533, AAAI Press, Vancouver, Canada, July 2007.

[21] X. Niyogi, "Locality preserving projections," in *Neural Information Processing Systems*, vol. 16, pp. 153–160, MIT, 2004.

[22] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1208–1213, IEEE, Beijing, China, October 2005.

[23] T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *Computer Vision—ECCV 2008*, pp. 725–738, Springer, 2008.

[24] Y. Fu, L. Cao, G. Guo, and T. S. Huang, "Multiple feature fusion by subspace learning," in *Proceedings of the International Conference on Content-Based Image and Video Retrieval (CIVR '08)*, pp. 127–134, ACM, Niagara Falls, Canada, July 2008.

[25] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *International Journal of Computer Vision*, vol. 109, no. 1-2, pp. 74–93, 2014.

[26] J. Gui, W. Jia, L. Zhu, S.-L. Wang, and D.-S. Huang, "Locality preserving discriminant projections for face and palmprint recognition," *Neurocomputing*, vol. 73, no. 13, pp. 2696–2707, 2010.

[27] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: which helps face recognition?" in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 471–478, IEEE, Barcelona, Spain, November 2011.

[28] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 625–632, IEEE, Providence, RI, USA, June 2011.

[29] X. Zhang, D. Pham, S. Venkatesh, W. Liu, and D. Phung, "Mixed-norm sparse representation for multi view face recognition," *Pattern Recognition*, vol. 48, no. 9, pp. 2935–2946, 2015.

[30] J. Yang, D. Chu, L. Zhang, Y. Xu, and J. Yang, "Sparse representation classifier steered discriminative projection with applications to face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 7, pp. 1023–1035, 2013.

[31] A. Shrivastava, V. M. Patel, and R. Chellappa, "Multiple kernel learning for sparse representation-based classification," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3013–3024, 2014.

[32] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, and Y. Y. Tang, "Group sparse multiview patch alignment framework with view consistency for image classification," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3126–3137, 2014.

[33] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1794–1801, IEEE, Miami, Fla, USA, June 2009.

[34] S. Zhang, H. Zhou, F. Jiang, and X. Li, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1749–1760, 2015.

[35] S. Zhang, H. Yao, X. Sun, and X. Lu, "Sparse coding based visual tracking: review and experimental comparison," *Pattern Recognition*, vol. 46, no. 7, pp. 1772–1788, 2013.

[36] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, 2013.

[37] W. Dong, G. Shi, and X. Li, "Nonlocal image restoration with bilateral variance estimation: a low-rank approach," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 700–711, 2013.

[38] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang, "Sparse unsupervised dimensionality reduction for multiple view data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 10, pp. 1485–1496, 2012.

[39] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.

[40] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving discriminant analysis for single training image face recognition," *Pattern Recognition Letters*, vol. 31, no. 5, pp. 422–429, 2010.

[41] F. Zang and J. Zhang, "Discriminative learning by sparse representation for classification," *Neurocomputing*, vol. 74, no. 12-13, pp. 2176–2183, 2011.

[42] Z. Lai, Y. Xu, J. Yang, J. Tang, and D. Zhang, "Sparse tensor discriminant analysis," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3904–3915, 2013.

[43] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, "MahNMF: Manhattan non-negative matrix factorization," http://arxiv.org/abs/1207.3438.

[44] Y. Fu, M. Liu, and T. S. Huang, "Conformal embedding analysis with local graph modeling on the unit hypersphere," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–6, Minneapolis, Minn, USA, June 2007.

[45] S. Lou, G. Zhang, H. Pan, and Q. Wang, "Supervised Laplacian discriminant analysis for small sample size problem with its application to face recognition," *Journal of Computer Research and Development*, vol. 49, no. 8, article 020, pp. 1730–1737, 2012.

[46] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision*, pp. 138–142, IEEE, December 1994.

[47] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46–51, IEEE, Washington, DC, USA, May 2002.

[48] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[49] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[50] G. H. Golub and C. F. Van Loan, *Matrix Computations*, vol. 3, JHU Press, 2012.