Computing and Informatics, Vol. 36, 2017, 815–836, doi: 10.4149/cai_2017_4_815

View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Computing and Informatics (E-Journal - Institute of Informatics,...

SUPERVISED SPARSITY PRESERVING PROJECTIONS FOR FACE RECOGNITION

Yingchun Ren

College of Mathematics Physics and Information Engineering
Jiaxing University
Jiaxing, 314001, P.R. China
e-mail: renyingchun2008@163.com

Yufei Chen

Research Center of CAD
Tongji University
Shanghai, 201804, P.R. China
e-mail: april337@163.com

Xiaodong Yue

School of Computer Engineering and Science
Shanghai University
Shanghai, 200444, P.R. China
e-mail: yswantfly@shu.edu.cn

Abstract. Recently feature extraction methods have commonly been used as a principled approach to understand the intrinsic structure hidden in high-dimensional data. In this paper, a novel supervised learning method, called Supervised Sparsity Preserving Projections (SSPP), is proposed. SSPP attempts to preserve the sparse representation structure of the data when identifying an efficient discriminant subspace. First, SSPP creates a concatenated dictionary by class-wise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary using the least squares method. Second, by maximizing the ratio of non-local scatter to local scatter, a Laplacian discriminant function is defined to characterize the separability of the samples in the different sub-manifolds. Then, to achieve improved recognition results, SSPP integrates the learned sparse representation structure as a regular term into the Laplacian discriminant function. Finally, the proposed method is converted into a generalized eigenvalue problem. The extensive and promising experimental results on several popular face databases validate the feasibility and effectiveness of the proposed approach.

Keywords: Feature extraction, sparse representation, manifold learning, Laplacian discriminant function

Mathematics Subject Classification 2010: 68T10

1 INTRODUCTION

In many application domains, such as speech recognition [1, 2], protein function prediction [3, 4], and time series analysis [5, 6], data are typically provided in a high-dimensional form because of which they are difficult to describe, interpret, and classify. In practice, feature extraction has become a widely used approach to handle the aforementioned problem [7, 8]. Thus far, a variety of feature extraction methods have been designed. Based on the data structure these methods utilize they are divided 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 to maximize the variance of the projected data, has extensive applications in the fields of science and engineering. Although PCA is an effective feature extraction method, it does not employ the label information of the samples, and this leads to an inefficient classification. Unlike PCA, LDA is a supervised method that attempts to identify an optimal projection by maximizing the between-class scatter while minimizing the within-class scatter. Because the label information is fully exploited in LDA, it has been proven more efficient than PCA with regards to classification [13]. However, LDA is limited in that it can extract $K-1$ features at best (K is the number of categories), what is unacceptable in many situations. Furthermore, LDA often encounters the small sample size (SSS) problem when high-dimensional data are involved. Baudat et al. proposed the Regularized LDA to address these problem [14, 15]; however, all the methods mentioned above are based on the hypothesis that samples from each class lie on a linear subspace [16, 17], and thus neither of them can identify the local sub-manifold structure hidden in high-dimensional data.

Recently, manifold learning methods, which are especially useful for the analysis of the data that lie on a sub-manifold of the original space, have been proposed [18, 19, 20, 21, 22]. Representative manifold learning methods include Laplacian Eigenmaps (LE) [19], Locally Linear Embedding (LLE) [20], Locality Preserving Projections (LPP) [21], and Neighborhood Preserving Embedding (NPE) [22]. All of the above manifold learning methods can determine the optimal feature subspace by solving an optimization problem based on the weight graph question; however, all of them consider only the local structure of the data, ignoring the non-local property of the samples: they do not consider the projected relationship of two distant samples in the original space. Furthermore, they are not supervised and do not utilize the label information when training data.

Sparse representation, a new state-of-the-art technique for signal representation, has been successfully applied to object identification [23, 24, 25, 26], medical image segmentation [27, 28, 29], visual tracking [30, 31], and image super-resolution reconstruction [32, 33, 34]. It models a signal as a sparse linear combination of the elementary signals from a dictionary and attempts to preserve the sparse representation structure of the samples in a low-dimensional embedding subspace. The representative feature extraction algorithms based on sparse representation include Sparsity Preserving Projections (SPP) [35], Sparsity Regularization Discriminant Analysis (SRDA) [36], Sparse Tensor Discriminant Analysis (STDA) [37], and Sparse Nonnegative Matrix Factorization [38]. It is noteworthy that a sparse model also depends on the subspace assumption that each sample can be linearly expressed by other samples from the same class, i.e., each sample can be sparsely recovered by samples from all classes. In general, these sparse learning algorithms provide a superior recognition accuracy compared with the conditional methods. However, all the feature extraction methods based on sparse coding mentioned earlier must 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 [35], which is an effective and powerful technique for feature extraction. However, the computational complexity of SPP is excessively high; 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 utilize the label information. Thus, the algorithm is unsupervised. Recently, Lu et al. and Zang et al. proposed Discriminant Sparsity Neighborhood Preserving Embedding (DSNPE) [39] and Discriminative Learning by Sparse Representation Projections (DLSP) [40], respectively, to improve the classification performance of SPP and applied them to face recognition. Experimental results show that DSNPE and DLSP are more suitable for recognition tasks than SPP. However, DSNPE and DLSP also require resolving the time-consuming ℓ_1-norm minimization problems to obtain the sparse weight graph; consequently, the computational costs of learning the sparse representation structure for DSNPE and DLSP are still very high.

Motivated by the above works, a novel supervised learning method, called Supervised Sparsity Preserving Projections (SSPP), is proposed in this paper. By integrating SPP with local discriminant information for dimensionality reduction, SSPP can be regarded as a combination of sparse representation and manifold learning. More specifically, SSPP first creates a concatenated dictionary using class-wise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary using the least squares method. Then, by maximizing the ratio of non-local scatter to local scatter, a Laplacian discriminant function is defined to characterize the separability of the samples in the different sub-manifolds. Subsequently, by integrating the sparse representation information as a regular term into the Laplacian discriminant function, SSPP attempts to preserve the sparse representation structure of the data and maximize the the separability of the different manifolds simultaneously. Finally, the proposed method is transformed into a generalized eigenvalue problem.

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

1. SSPP is a supervised feature extraction method that aims to identify a discriminating subspace where the sparse representation structure of the data and the label information are maintained. Meanwhile, the separability of different sub-manifolds is maximized and the separability of each sub-manifold is minimized; consequently, samples belonging to different classes can be distinguished more clearly.
2. The time required for extracting discriminant vectors in SSPP is significantly less than that in the SPP algorithm. Therefore, the proposed method can be applied to solve large-scale problems more time-efficiently.
3. Label information is employed twice in SSPP. First, it is adopted to construct the dictionary for sparse representation and to calculate the sparse coefficient vector, which may contribute to a more discriminating sparse representation structure. Second, it is utilized in computing the local scatter and non-local scatter, which is more conducive for classification.
4. The small sample size problem is effectively avoided in SSPP because the singular problem of the local scatter matrix is circumvented owing to the sparse representation and the Tikhonov [41] regularized term in the SSPP formulation.

The rest of this paper is organized as follows: Section 2 briefly reviews the existing SPP algorithm. The SSPP algorithm is introduced in detail in Section 3. The experimental results and analysis are presented in Section 4 and lastly, the concluding remarks are given in Section 5.

2 BRIEF REVIEW OF SPARSITY PRESERVING PROJECTIONS (SPP)

SPP aims to preserve the sparse reconstruction relationship of the samples [35]. Given a set of training samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ where $\boldsymbol{x}_i \in \boldsymbol{R}^m$ and n is the number of training samples. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \boldsymbol{R}^{m\times n}$ be the data matrix consisting of all the training samples. SPP first seeks the sparse reconstruction coefficient vector $\boldsymbol{s}_i$ for each sample $\boldsymbol{x}_i$ through the following modified ℓ_1-minimization problem:

$$\min_{\boldsymbol{s}_i} \|\boldsymbol{s}_i\|_1, s.t.\boldsymbol{x}_i{=}\boldsymbol{X}\boldsymbol{s}_i, 1{=}\mathbf{1}^T\boldsymbol{s}_i \tag{1}$$

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

$$\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_n] \tag{2}$$

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

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

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

It follows that SPP need resolve n time-consuming ℓ_1-norm minimization problems to obtain the sparse weight matrix $\boldsymbol{S}$. 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 SUPERVISED SPARSITY PRESERVING PROJECTIONS

In order to minimize the disadvantage caused in the case of SPP because of the requirement of resolving n time-consuming ℓ_1-norm minimization problems to obtain the sparse weight matrix $\boldsymbol{S}$, SSPP first constructs a concatenated dictionary through class-wise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary quickly using the least squares method. To enhance the classification performance, it defines a non-local scatter matrix and local scatter matrix to characterize the separability of the samples in the sub-manifolds and then constructs a Laplacian discriminant function by maximizing the ratio of non-local scatter to local scatter. Subsequently, by integrating the sparse representation information into the Laplacian discriminant function, SSPP aims to maximize the separation between the sub-manifolds (or intrinsic clusters) without destroying localities while preserving the sparse representation structure of the data. Hence, the proposed algorithm is expected to not only preserve the intrinsic geometry structure, but also to have superior discriminant abilities.

3.1 Constructing the Concatenated Dictionary

For convenience, we first provide some notations used in this paper. Assume that $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ is a set of training samples, where $\boldsymbol{x}_i \in \boldsymbol{R}^m$. We can categorize the training samples as $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_K]$, where $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \dots, \boldsymbol{x}_{in_i}] \in \boldsymbol{R}^{m \times n_i} (i = 1, 2, \dots, K)$ 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 [36]. The subspace model is a powerful tool to capture the underlying information in real data sets [42, 43]. For the convenience of PCA decomposition and relevant calculations, we first center the samples from each class at the origin, $\tilde{\boldsymbol{X}}_i = [\boldsymbol{x}_{i1} - \boldsymbol{\mu}_i, \boldsymbol{x}_{i2} - \boldsymbol{\mu}_i, \dots, \boldsymbol{x}_{in_i} - \boldsymbol{\mu}_i]$ $(i = 1, 2, \dots, K)$, where $\boldsymbol{\mu}_i$ denotes the mean of class i; that is, $\boldsymbol{\mu}_i = \sum_{i=1}^{n_i} \boldsymbol{x}_i / \mathrm{n}_i$. Therefore, the training sample can be recast as $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2, \dots, \tilde{\boldsymbol{X}}_K]$. Afterwards, PCA decomposition is conducted for every $\tilde{\boldsymbol{X}}_i (i = 1, 2, \dots, K)$, whose objective function is:

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

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

$$\boldsymbol{x} = \boldsymbol{D}_i \tilde{\boldsymbol{s}}_i = [\boldsymbol{D}_1, \boldsymbol{D}_2, \dots, \boldsymbol{D}_{i-1}, \boldsymbol{D}_i, \boldsymbol{D}_{i+1}, \dots, \boldsymbol{D}_K] \boldsymbol{s} = \boldsymbol{D} \boldsymbol{s}, \tag{5}$$

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

$$\tilde{\boldsymbol{s}}_i = (\boldsymbol{D}_i{}^T \boldsymbol{D}_i)^{-1} \boldsymbol{D}_i{}^T \boldsymbol{x} = \boldsymbol{D}_i{}^T \boldsymbol{x}. \tag{6}$$

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 $\boldsymbol{D}$ and the sparse coefficient vector $\boldsymbol{s}$ of any training sample from class i can be quickly computed from the least squares 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 $\boldsymbol{s}$ involves only $\boldsymbol{D}_i$, which is column orthogonal in view of Equations (5) and (6).

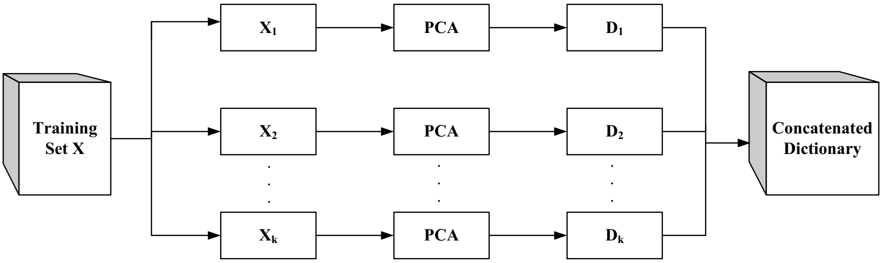


Figure 1. The process of constructing the concatenated dictionary

3.2 Preserving Sparse Representation Structure

As can be seen in Section 3.1, to some extent, the dictionary $\boldsymbol{D}$ 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:

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

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

Using algebraic operations, Equation (7) can be arranged as:

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

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

$$J_s(\boldsymbol{w}) = \min_{\boldsymbol{w}} \boldsymbol{w}^T \left(\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{S}^T\boldsymbol{D}^T - \boldsymbol{D}\boldsymbol{S}\boldsymbol{X}^T + \boldsymbol{D}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{D}^T \right) \boldsymbol{w}. \tag{9}$$

3.3 Characterization of the Laplacian Discriminant Function

To effectively discover the discriminant structure embedded in high-dimensional data and improve the classification performance, we construct a Laplacian discriminant function. Because data belonging to the same class lie on one or more sub-manifolds and data belonging to different classes are distributed on different sub-manifolds, it is important for classification problems to distinguish one sub-manifold from another. Therefore, a non-local scatter matrix and a local scatter matrix are defined to characterize the separability of the samples in the sub-manifolds. The aim of SSPP is to distinguish between different sub-manifolds more clearly after they are projected; hence, the non-local scatter of different sub-manifolds should be maximized and the local scatter of the data belonging to the same manifold should be minimized simultaneously. Thus, we can construct the similarity matrix $\boldsymbol{\Omega} = [\Omega_{ij}]$ and diversity matrix $\boldsymbol{B} = [B_{ij}]$, to describe the local and non-local relationships, respectively of each point as follows:

$$\Omega_{ij} = \begin{cases} exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2}{\sigma}\right), & \text{if both } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are } k \text{ nearest neighbors} \\ & \text{each other and have the same label;} \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

$$B_{ij} = \begin{cases} 1 - exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2}{\sigma}\right), & \text{if both } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are } k \text{ nearest neighbors} \\ & \text{each other and have different labels;} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ denotes the geodesic distance between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, σ is a parameter, which is often set as the standard deviation value of the samples. As it is evident in the above definition, if two close points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same manifold, the similarity between them is considerable and in contrast, if two distant points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to different sub-manifolds, the diversity between them is considerable. In summary, the points belonging to the same sub-manifold should be located closer while the points belonging to different sub-manifolds should be farther apart after projection. Therefore, the local scatter and non-local scatter (or separability) can be characterized by Equations (12) and (13), respectively:

$$J_l(\boldsymbol{w}) = \frac{1}{2nn} \sum_i \sum_j \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2 \Omega_{ij}, \tag{12}$$

$$J_n(\boldsymbol{w}) = \frac{1}{2nn} \sum_i \sum_j \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2 B_{ij} \tag{13}$$

where $\boldsymbol{y}_i = \boldsymbol{w}^T\boldsymbol{x}_i$ $(i = 1, 2, \ldots, n)$ is the low-dimensional representation of the original data, which can be obtained by projecting each $\boldsymbol{x}_i$ onto the direction vector $\boldsymbol{w} \in \boldsymbol{R}^m$. With algebraic simplifications, Equation (12) can be rewritten as:

$$\begin{aligned}
J_l &= \frac{1}{2nn}\sum_{i,j}^{n}\Omega_{ij}\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|_2^2 \\
&= \frac{1}{2nn}\boldsymbol{w}^T\left(\sum_{i,j}^{n}\Omega_{ij}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{x}_i-\boldsymbol{x}_j)\right)\boldsymbol{w} \\
&= \frac{1}{nn}\boldsymbol{w}^T\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{ij}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{x}_i-\boldsymbol{x}_j)\right)\boldsymbol{w} \\
&= \frac{1}{nn}\boldsymbol{w}^T\left(\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{ij}\boldsymbol{x}_i{\boldsymbol{x}_i}^T-2\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{ij}\boldsymbol{x}_i{\boldsymbol{x}_j}^T+\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{ij}\boldsymbol{x}_j{\boldsymbol{x}_j}^T\right)\right)\boldsymbol{w} \\
&= \frac{1}{nn}\boldsymbol{w}^T\left(\sum_{i=1}^{n}\boldsymbol{D'}_{ii}\boldsymbol{x}_i{\boldsymbol{x}_i}^T-\sum_{i=1}^{n}\sum_{j=1}^{n}\Omega_{ij}\boldsymbol{x}_i{x_j}^T\right)\boldsymbol{w} \\
&= \frac{1}{nn}\boldsymbol{w}^T\left(\boldsymbol{X}(\boldsymbol{D'}-\boldsymbol{\Omega})\boldsymbol{X}^T\right)\boldsymbol{w} \\
&= \frac{1}{nn}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}_\Omega\boldsymbol{X}^T\boldsymbol{w} \qquad (14)
\end{aligned}$$

where $\boldsymbol{L}_\Omega$ is the intra-class Laplacian matrix with definition $\boldsymbol{L}_\Omega = \boldsymbol{D'} - \boldsymbol{\Omega}$ and $\boldsymbol{D'}$ is a diagonal matrix [44], i.e., $\boldsymbol{D'}_{ii} = \sum_j \Omega_{ij}$. Equation (14) characterizes the separability of the data set in the same sub-manifold. Similarly, the non-local scatter (or separability) can be expressed as:

$$\begin{aligned}
J_n &= \frac{1}{2nn}\sum_{i,j}^{n}B_{ij}\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|_2^2 \qquad (15)\\
&= \frac{1}{nn}\boldsymbol{w}^T\left(\boldsymbol{X}(\boldsymbol{D''}-\boldsymbol{B})\boldsymbol{X}^T\right)\boldsymbol{w} \qquad (16)\\
&= \frac{1}{nn}\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}_B\boldsymbol{X}^T\boldsymbol{w} \qquad (17)
\end{aligned}$$

where $\boldsymbol{L}_B$ is the inter-class Laplacian matrix with definition $\boldsymbol{L}_B = \boldsymbol{D''}-\boldsymbol{B}$ and $\boldsymbol{D''}$ is a diagonal matrix, i.e., $\boldsymbol{D''}_{ii} = \sum_j B_{ij}$. Equation (17) characterizes the diversity (or scatter) of the data set in the different sub-manifolds. Therefore, each manifold can be separated clearly, as long as the optimal projection $\boldsymbol{w}^*$ is adopted. To ensure that the projected samples of different sub-manifolds remain distant from each other while samples from the same sub-manifold remain close, we can construct the Laplacian discriminant function as follows:

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{J_n(\boldsymbol{w})}{J_l(\boldsymbol{w})} = \frac{\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}_B\boldsymbol{X}^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}_\Omega\boldsymbol{X}^T\boldsymbol{w}}. \qquad (18)$$

3.4 Algorithm of Supervised Sparsity Preserving Projections

To achieve improved recognition results, we explicitly integrate the sparsity preserving constraint as indicated in Equation (7) into the Laplacian discriminant function, as illustrated in Equation (18). According to Section 3.2, minimizing the sparsity preserving regular term can preserve the intrinsic sparse structure; therefore, the novel supervised algorithm SSPP, which not only identifies an efficient discriminating subspace but also preserves the sparse representation structure, is defined as:

$$\max_{\boldsymbol{w}} \frac{J_n(\boldsymbol{w})}{J_l(\boldsymbol{w}) + \lambda_1 \boldsymbol{w}^T \boldsymbol{w} + \lambda_2 J_s(\boldsymbol{w})} \tag{19}$$

where λ_1 and λ_2 are two parameters that control the tradeoff among the three terms in the denominator; $J_n(\boldsymbol{w})$ and $J_l(\boldsymbol{w})$ are the non-local scatter and local scatter in Section 3.3, respectively; $J_s(\boldsymbol{w})$ is the sparsity preserving term in Section 3.2. To avoid the small sample size problem, the Tikhonov regular term $\boldsymbol{w}^T\boldsymbol{w}$ is employed. Substituting Equation (8) into (19) and making some simple algebraic manipulations, we obtain

$$\begin{aligned}
&\frac{J_n(\boldsymbol{w})}{J_l(\boldsymbol{w}) + \lambda_1 \boldsymbol{w}^T \boldsymbol{w} + \lambda_2 J_s(\boldsymbol{w})} \\
&= \frac{\boldsymbol{w}^T \boldsymbol{X} \boldsymbol{L}_B \boldsymbol{X}^T \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{X} \boldsymbol{L}_\Omega \boldsymbol{X}^T \boldsymbol{w} + \lambda_1 \boldsymbol{w}^T \boldsymbol{w} + \lambda_2 \boldsymbol{w}^T (\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{S}^T\boldsymbol{D}^T - \boldsymbol{D}\boldsymbol{S}\boldsymbol{X}^T + \boldsymbol{D}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{D}^T)\boldsymbol{w}} \\
&= \frac{\boldsymbol{w}^T \boldsymbol{X} \boldsymbol{L}_B \boldsymbol{X}^T \boldsymbol{w}}{\boldsymbol{w}^T (\boldsymbol{X} \boldsymbol{L}_\Omega \boldsymbol{X}^T + \lambda_1 \boldsymbol{I} + \lambda_2 \boldsymbol{M}) \boldsymbol{w}}
\end{aligned} \tag{20}$$

where $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{S}^T\boldsymbol{D}^T - \boldsymbol{D}\boldsymbol{S}\boldsymbol{X}^T + \boldsymbol{D}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{D}^T$ and $\boldsymbol{I}$ is an identity matrix. In fact, $\boldsymbol{M}$ corresponds to the sparsity preserving regular term and $\boldsymbol{I}$ is associated with the Tikhonov regular term. Therefore, it can be deduced that the projecting matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d]$ is composed of eigenvectors associated with the largest d eigenvalues of the following generalized eigenvalue problem:

$$\boldsymbol{X} \boldsymbol{L}_B \boldsymbol{X}^T \boldsymbol{w} = \eta (\boldsymbol{X} \boldsymbol{L}_\Omega \boldsymbol{X}^T + \lambda_1 \boldsymbol{I} + \lambda_2 \boldsymbol{M}) \boldsymbol{w}. \tag{21}$$

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

4 EXPERIMENTS

In this section, the proposed SSPP algorithm is tested on three publicly available face databases (Yale [13], ORL [45], and Extended Yale B [46]) and compared with seven popular dimensionality reduction methods—PCA, LDA, LPP, NPE, SPP, DLSP, and DSNPE. Furthermore, 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

Algorithm 1 Supervised Sparsity Preserving Projections (SSPP)

```
Step 1: Execute PCA decomposition for each $\boldsymbol{X}_i (i = 1, 2, \ldots, K)$ using Equation (4) to
        obtain the concatenated dictionary $\boldsymbol{D}$;
Step 2: Calculate the coefficient vector $\tilde{\boldsymbol{s}}_i$ under the dictionary $\boldsymbol{D}_i$ for each sample based
        on Equation (6) to obtain the sparse coefficient vector $\boldsymbol{s}$ and then calculate $\boldsymbol{S}$;
Step 3: Calculate $\boldsymbol{L}_\Omega$ and $\boldsymbol{L}_B$ by Equations (14) and (15), respectively;
Step 4: Calculate the projecting vectors by the generalized eigenvalue problem in Equa-
        tion (19).
```

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". ε in SPP is set to be 0.05 as indicated in [35] and μ in DSNPE is empirically set to be 10 as shown in [39]. The trade-off parameter α in DLSP is set to be 0.01 as indicated in [40]. σ in SSPP is set as the standard deviation value of the samples and the trade-off parameters (λ_1, λ_2) are set to be $(0.94, 0.25)$, $(0.72, 0.36)$, and $(0.90, 0.20)$ for Yale, ORL, and Extended Yale B, respectively, by the tenfold cross-validation where λ_1 and λ_2 are selected from $\{0.01, 0.02, \ldots, 0.99\}$. Since the dimensionality of the face vector space is considerably larger than the number of training samples, LPP, NPE, SPP, DLSP, and DSNPE all include a PCA preprocessing phase; that is, projecting the training set X onto a PCA subspace spanned by the leading eigenvectors. For Yale and ORL, which are relatively small databases, 100 % energy is kept in the PCA preprocessing phase; for Extended Yale B, which are relatively large-scale databases, to obtain the experimental results in a reasonable time, 98 % energy is maintained in the PCA preprocessing phase. 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 ® Core$^{\mathrm{TM}}$ i7-4770K 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 winky); configurations (left-right, 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. 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 eight 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) shows the variation rules of the recognition rates of the eight algorithms under different reduced dimensions when the size of the training samples from each class is fixed as six. It deserves to be noted 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 non-zero eigenvalues [13]; similar situations will occur in other experiments in this paper. Figures 3 c) and 3 d) describe the relationship of the classification accuracy versus the trade-off parameters λ_1 and λ_2 corresponding to Figure 3 b). Hence, one can see that the SSPP algorithm significantly outperforms the other methods and it is robust to the parameters λ_1 and λ_2; that is to say, SSPP is not sensitive to the facial expression and configuration changes.

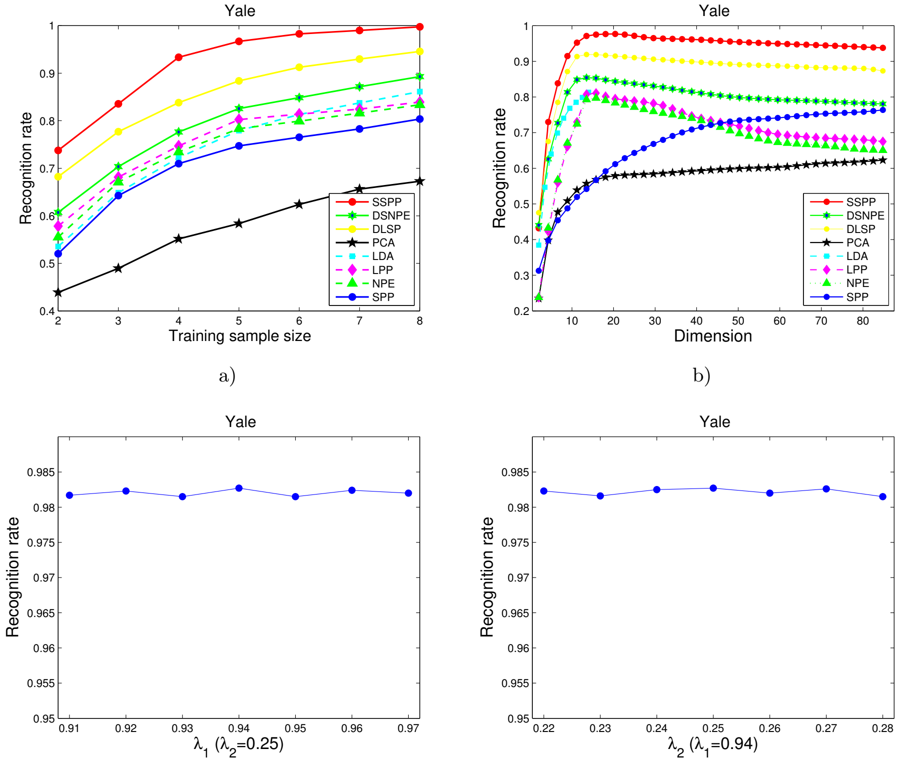
Figure 2. Some face samples from the Yale database

Method	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
SSPP	0.7376 (±0.027)	0.8356 (±0.031)	0.9335 (±0.029)	0.9669 (±0.026)	0.9827 (±0.033)	0.9899 (±0.035)	0.9976 (±0.029)
DSNPE	0.6072 (±0.033)	0.7035 (±0.029)	0.7763 (±0.037)	0.8256 (±0.035)	0.8485 (±0.033)	0.8715 (±0.031)	0.8927 (±0.028)
DLSP	0.6823 (±0.022)	0.7769 (±0.018)	0.8381 (±0.021)	0.8838 (±0.026)	0.9125 (±0.022)	0.9297 (±0.025)	0.9458 (±0.021)
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)

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

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 points, under different lighting conditions, varying facial expressions. In our experiment, each image is cropped to a resolution of 32×32 as show in Figure 4. We randomly select k (k varies from 2 to 8) pictures from each person for training; the remaining ones are used for testing. We repeat these splits 50 times and report the average results. Table 2 displays the best classification accuracy of the eight 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 eight algorithms under different reduced dimensions when the size of the training samples from each class is fixed as five. The relationship of the classification accuracy versus the trade-off parameters λ_1 and λ_2 corresponding to Figure 5 b) are



a) b)

c) d)

Figure 3. Recognition rates of the eight algorithms on the Yale database: a) the best recognition rates versus the different size of the training set, b) the average recognition rates versus the variation of dimensions when the size of per class is fixed as six, c) influence of λ_1 on the performance of SSPP on Yale, and d) influence of λ_2 on the performance of SSPP on Yale

described in Figures 5 c) and 5 d). It can be seen that SSPP is superior to other compared methods, especially when the size of the training set is small.

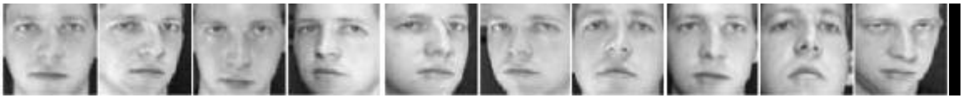


Figure 4. Some face samples from the ORL database

Method	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
SSPP	0.8415 (±0.031)	0.8990 (±0.019)	0.9431 (±0.027)	0.9646 (±0.019)	0.9823 (±0.023)	0.9921 (±0.022)	0.9975 (±0.018)
DSNPE	0.8272 (±0.023)	0.8935 (±0.027)	0.9363 (±0.022)	0.9596 (±0.018)	0.9755 (±0.026)	0.9890 (±0.024)	0.9935 (±0.020)
DLSP	0.8192 (±0.035)	0.8898 (±0.031)	0.9331 (±0.029)	0.9543 (±0.033)	0.9758 (±0.030)	0.9822 (±0.028)	0.9889 (±0.031)
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 2. The best recognition rate and the corresponding standard deviation of the eight algorithms under the different size of the training set on ORL (k is the training sample size)

4.3 Experiment on Extended Yale B Face Database

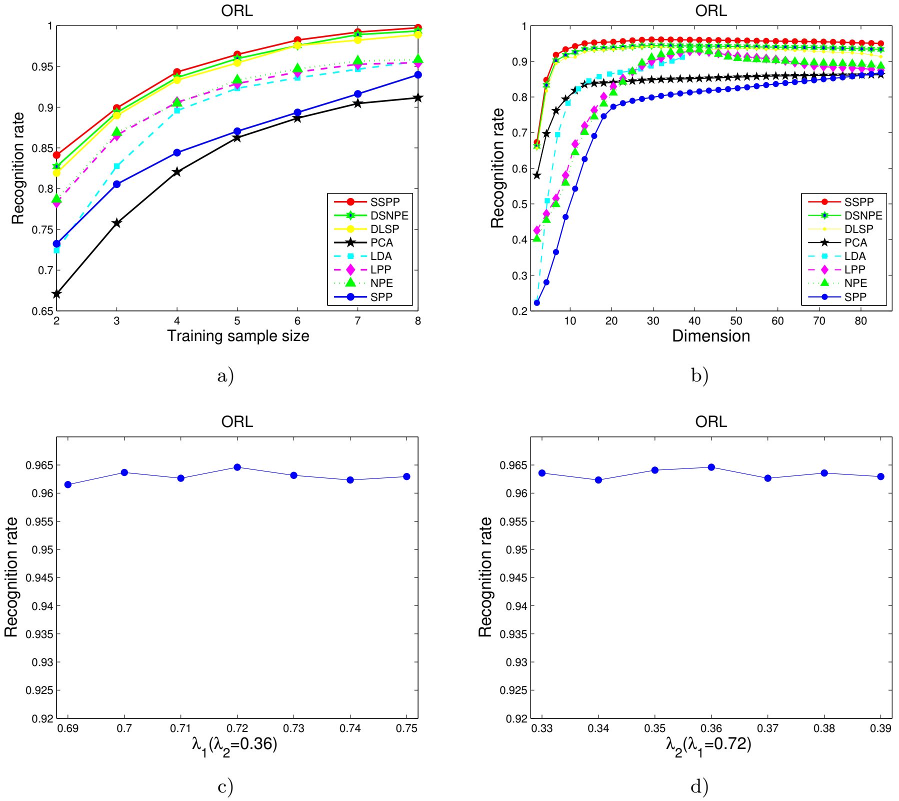
The Extended Yale B consists of over 2 414 front-view face images of 38 subjects, with approximately 64 pictures under various laboratory-controlled lighting conditions for each subject. We crop the images to 32×32; Figure 6 presents some pictures of one subject. A random subset with k $(= 10, 20, 30, 40, 50)$ pictures per subject is selected with labels to form the training set; the remaining pictures 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 eight 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 eight algorithms under different reduced dimensions when the size of the training samples from each class is fixed as 30. Figures 7 c) and 7 d) indicate the influence of the classification accuracy versus the trade-off parameter λ_1 and λ_2 corresponding to Figure 7 b). It can be observed that the proposed SSPP provides superior recognition performance compared to the other feature extraction methods such as PCA, LDA, LPP, NPE, SPP, DSNPE,and DLSP regarding lighting changes.

Method	$k=10$	$k=20$	$k=30$	$k=40$	$k=50$
SSPP	0.8965 (±0.029)	0.9655 (±0.027)	0.9814 (±0.019)	0.9895 (±0.025)	0.9979 (±0.027)
DSNPE	0.8815 (±0.033)	0.9532 (±0.031)	0.9698 (±0.029)	0.9805 (±0.034)	0.9859 (±0.037)
DLSP	0.8906 (±0.025)	0.9587 (±0.023)	0.9755 (±0.022)	0.9833 (±0.026)	0.9916 (±0.027)
PCA	0.7496 (±0.012)	0.7877 (±0.016)	0.8041 (±0.021)	0.8248 (±0.023)	0.8406 (±0.014)
LDA	0.8642 (±0.021)	0.9475 (±0.032)	0.9647±0.026)	0.9816 (±0.028)	0.9871 (±0.035)
LPP	0.8780 (±0.043)	0.9642 (±0.039)	0.9591 (±0.037)	0.9729 (±0.038)	0.9818 (±0.041)
NPE	0.8306 (±0.022)	0.9309 (±0.019)	0.9471 (±0.026)	0.9686 (±0.024)	0.9752 (±0.021)
SPP	0.8379 (±0.034)	0.9197 (±0.028)	0.9473 (±0.025)	0.9583 (±0.030)	0.9639 (±0.029)

Table 3. The best recognition rate and the corresponding standard deviation of the eight algorithms under the different size of the training set on Extended Yale B (k is the training sample size)

4.4 Comparison of Time Cost for Acquiring the Discriminant Vectors of SPP with SSPP

In this subsection, the time cost for acquiring the discriminant vectors of SSPP is compared with that of SPP. Table 4, Table 5, and Table 6 list the average time costs



a)

b)

c)

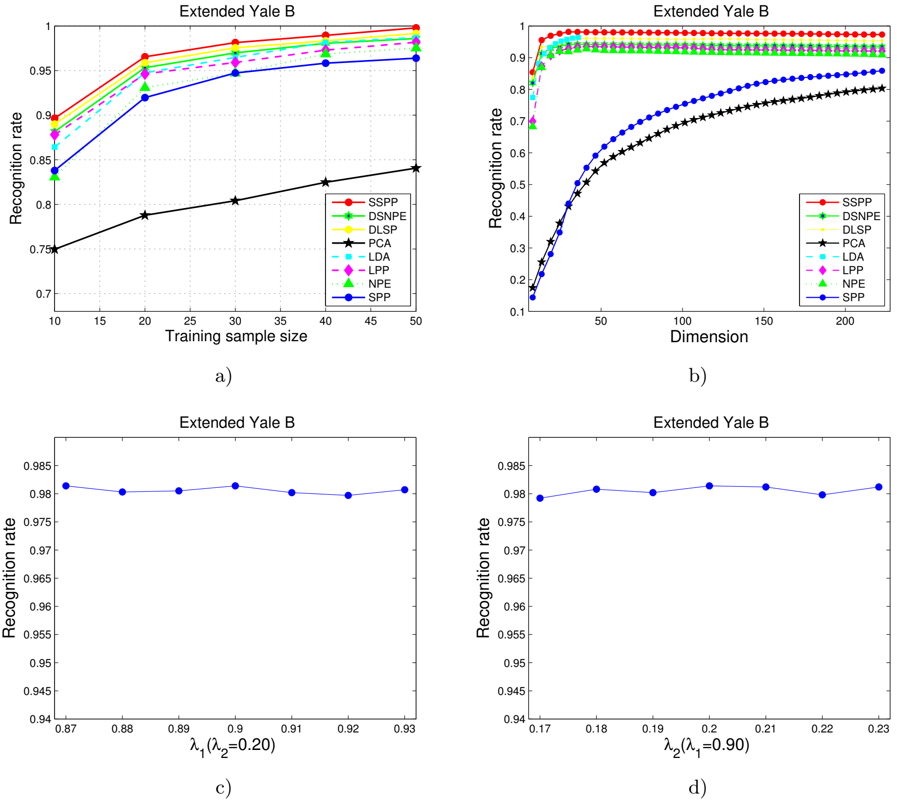
d)

Figure 5. Recognition rates of the eight algorithms on the ORL database: a) the best recognition rates versus the different size of the training set, b) the average recognition rates versus the variation of dimensions when the size of per class is fixed as five, c) influence of λ_1 on the performance of SSPP on ORL, and d) influence of λ_2 on the performance of SSPP on ORL

for acquiring the discriminant vectors of SPP and SSPP versus the different sizes of the training set on the three face data sets. It is demonstrated that SSPP is significantly faster than SPP in acquiring the embedding functions in our experiments, especially in the large-scale problems such as Extended Yale B.

The critical factor of the above phenomenon is that the approaches of SPP and SSPP to obtain the sparse representation structure are entirely different. In SPP,

Figure 6. Some face samples from the Extended Yale B database



a)

b)

c)

d)

Figure 7. Recognition rates of the eight algorithms on the Extended Yale B database: a) the best recognition rates versus the different size of the training set, b) the average recognition rates versus the variation of dimensions when the size of per class is fixed as thirty, c) influence of λ_1 on the performance of SSPP on Extended Yale B, and d) influence of λ_2 on the performance of SSPP on Extended Yale B

Method	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
SPP	0.3729	0.6387	1.0335	1.5506	2.1609	2.9087	4.0471
SSPP	0.3245	0.6098	0.6672	0.7367	0.7895	0.8758	0.9569

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

Method	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
SPP	1.1933	2.5641	5.1679	8.4467	13.0688	19.7787	29.4638
SSPP	0.3013	0.3796	0.4931	0.6062	0.7905	0.9536	1.1375

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

Method	$k = 10$	$k = 20$	$k = 30$	$k = 40$	$k = 50$
SPP	33.4459	42.9977	190.0017	418.6019	602.6975
SSPP	0.5286	1.8159	3.9985	6.9359	10.7869

Table 6. Time (s) for acquiring the discriminant vectors of SPP and SSPP on Extended Yale B (k is the training sample size)

n time consuming ℓ_1 norm minimization problems are required to be solved to construct the sparse weight matrix, whose computational cost is $O(n^4)$ [47]; whereas, SSPP 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 [48], 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 $\boldsymbol{S}$ can be calculated with $O(m \sum_{i=1}^{K} n_i l_i)$, the computational complexity of SSPP 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, SSPP 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 concluded from the above experimental results.

1. From Tables 1, 2, 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 SSPP simultaneously considers both the sparse representation structure and the separability of the different sub-manifolds. Further, this indicates that SSPP can capture more inherent information that is hidden in the data compared to the other compared methods.
2. From Figures 3 b), 5 b), and 7 b), it can be observed that the reduction dimensions for SSPP 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 SSPP is considerably faster than SPP in obtaining the discriminant vectors. This is because the method SSPP uses to learn the sparse representation structure is more effective than that of SPP as analyzed in Section 4.4.
4. According to the experimental results in Figures 3 c) d), 5 c) d), and 7 c) d), the performance of SSPP does not fluctuate significantly based on the variation of λ_1 and λ_2 on all three tested data sets; therefore, it is robust to the regular parameters λ_1 and λ_2.

5 CONCLUSIONS

This paper proposed a new supervised learning method called SSPP, by combining manifold learning and sparse representation. First, SSPP constructs a concatenated dictionary using class-wise PCA decompositions and learns the sparse representation structure of each sample under the constructed dictionary quickly using the least squares method. Then, it defines a Laplacian discriminant function to characterize the separability of the samples in different sub-manifolds. Subsequently, SSPP integrates the sparse representation information into the Laplacian discriminant function. Thus, SSPP preserves the sparse representation structure of the data and simultaneously maximizes the separability of different sub-manifolds. 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 SSPP approach.

Acknowledgement

The authors would like to express their gratitude to the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. This work is supported by the National Natural Science Foundation of China (61573235); and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Graves, A.—Mohamed, A.-R.—Hinton, G.: Speech Recognition with Deep Recurrent Neural Networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.

[2] Su, J.—Srivastava, A.—de Souza, F. D. M.—Sarkar, S.: Rate-Invariant Analysis of Trajectories on Riemannian Manifolds with Application in Visual Speech Recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, doi: 10.1109/CVPR.2014.86.

[3] Radivojac, P.—Clark, W. T.—Oron, T. R.—Schnoes, A. M.—Wittkop, T.—Sokolov, A.—Graim, K.—Funk, C.—Verspoor, K.—Ben-Hur, A. et al.: A Large-Scale Evaluation of Computational Protein Function Prediction. Nature Methods, Vol. 10, 2013, No. 3, pp. 221–227, doi: 10.1038/nmeth.2340.

[4] Yang, J.—Rangwala, H.—Domeniconi, C.—Zhang, G.—Yu, Z.: Protein Function Prediction with Incomplete Annotations. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), Vol. 11, 2014, No. 3, pp. 579–591.

[5] Frank, J.—Mannor, S.—Pineau, J.—Precup, D.: Time Series Analysis Using Geometric Template Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, 2013, No. 3, pp. 740–754, doi: 10.1109/TPAMI.2012.121.

[6] Wang, X.—Mueen, A.—Ding, H.—Trajcevski, G.—Scheuermann, P.—Keogh, E.: Experimental Comparison of Representation Methods and Distance Measures for Time Series Data. Data Mining and Knowledge Discovery, Vol. 26, 2013, No. 2, pp. 275–309, doi: 10.1007/s10618-012-0250-5.

[7] Kantorov, V.—Laptev, I.: Efficient Feature Extraction, Encoding and Classification for Action Recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, doi: 10.1109/CVPR.2014.332.

[8] Sun, L.—Jia, K.—Chan, T.-H.—Fang, Y.—Wang, G.—Yan, S.: DL-SFA: Deeply-Learned Slow Feature Analysis for Action Recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, doi: 10.1109/CVPR.2014.336.

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

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

[11] Schölkopf, B.—Smola, A.—Müller, K. R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation, Vol. 10, 1998, No. 5, pp. 1299–1319, doi: 10.1162/089976698300017467.

[12] Yang, J.—Frangi, A. F.—Yang, J. Y.—Zhang, D.—Jin, Z.: KPCA plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, 2005, No. 2, pp. 230–244.

[13] Belhumeur, P. N.—Hespanha, J. P.—Kriegman, D. J.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, 1997, No. 7, pp. 711–720.

[14] Friedman, J. H.: Regularized Discriminant Analysis. Journal of the American Statistical Association, Vol. 84, 1989, No. 405, pp. 165–175, doi: 10.1080/01621459.1989.10478752.

[15] Baudat, G.—Anouar, F.: Generalized Discriminant Analysis Using a Kernel Approach. Neural Computation, Vol. 12, 2000, No. 10, pp. 2385–2404, doi: 10.1162/089976600300014980.

[16] Nikitidis, S.—Zafeiriou, S.—Pantic, M.: Merging SVMs with Linear Discriminant Analysis: A Combined Model. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, doi: 10.1109/CVPR.2014.140.

[17] Ji, S.—Ye, J.: Linear Dimensionality Reduction for Multi-Label Classification. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), 2009, pp. 1077–1082.

[18] Tenenbaum, J. B.: Mapping a Manifold of Perceptual Observations. Advances in Neural Information Processing Systems, 1998, pp. 682–688.

[19] Belkin, M.—Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation, Vol. 15, 2003, No. 6, pp. 1373–1396, doi: 10.1162/089976603321780317.

[20] Roweis, S. T.—Saul, L. K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, Vol. 290, 2000, No. 5500, pp. 2323–2326.

[21] He, X.—Niyogi, P.: Locality Preserving Projections. Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS '03). MIT Press Cambridge, 2003, pp. 153–160.

[22] He, X.—Cai, D.—Yan, S.—Zhang, H.-J.: Neighborhood Preserving Embedding. Tenth IEEE International Conference on Computer Vision (ICCV 2005), 2005, Vol. 2, pp. 1208–1213.

[23] Guha, T.—Ward, R. K.: Learning Sparse Representations for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, 2012, No. 8, pp. 1576–1588.

[24] Qiao, L.—Chen, S.—Tan, X.: Sparsity Preserving Discriminant Analysis for Single Training Image Face Recognition. Pattern Recognition Letters, Vol. 31, 2010, No. 5, pp. 422–429.

[25] Deng, L.—Li, X.: Machine Learning Paradigms for Speech Recognition: An Overview. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, 2013, No. 5, pp. 1060–1089.

[26] Zhou, Y.—Liu, K.—Carrillo, R. E.—Barner, K. E.—Kiamilev, F.: Kernel-Based Sparse Representation for Gesture Recognition. Pattern Recognition, Vol. 46, 2013, No. 12, pp. 3208–3222.

[27] Tong, T.—Wolz, R.—Coupé, P.—Hajnal, J. V.—Rueckert, D.—The Alzheimer's Disease Neuroimaging Initiative: Segmentation of MR Images via Discriminative Dictionary Learning and Sparse Coding: Application to Hippocampus Labeling. NeuroImage, Vol. 76, 2013, pp. 11–23, doi: 10.1016/j.neuroimage.2013.02.069.

[28] Cetingul, H. E.—Wright, M. J.—Thompson, P. M.—Vidal, R.: Segmentation of High Angular Resolution Diffusion MRI Using Sparse Riemannian Manifold Clustering. IEEE Transactions on Medical Imaging, Vol. 33, 2014, No. 2, pp. 301–317, doi: 10.1109/TMI.2013.2284360.

[29] Wang, L.—Shi, F.—Gao, Y.—Li, G.—Gilmore, J. H.—Lin, W.—Shen, D.: Integration of Sparse Multi-Modality Representation and Anatomical Constraint for Isointense Infant Brain MR Image Segmentation. NeuroImage, Vol. 89, 2014, pp. 152–164, doi: 10.1016/j.neuroimage.2013.11.040.

[30] Bai, T.—Li, Y.-F.—Zhou, X.: Learning Local Appearances with Sparse Representation for Robust and Fast Visual Tracking. IEEE Transactions on Cybernetics, Vol. 45, 2015, No. 4, pp. 663–675.

[31] Zhang, S.—Yao, H.—Zhou, H.—Sun, X.—Liu, S.: Robust Visual Tracking Based on Online Learning Sparse Representation. Neurocomputing, Vol. 100, 2013, pp. 31–40, doi: 10.1016/j.neucom.2011.11.031.

[32] Yang, S.—Wang, M.—Chen, Y.—Sun, Y.: Single-Image Super-Resolution Reconstruction via Learned Geometric Dictionaries and Clustered Sparse Coding. IEEE Transactions on Image Processing, Vol. 21, 2012, No. 9, pp. 4016–4028.

[33] Gao, X.—Zhang, K.—Tao, D.—Li, X.: Image Super-Resolution with Sparse Neighbor Embedding. IEEE Transactions on Image Processing, Vol. 21, 2012, No. 7, pp. 3194–3205.

[34] Dong, W.—Fu, F.—Shi, G.—Cao, X.—Wu, J.—Li, G.—Li, X.: Hyperspectral Image Super-Resolution via Non-Negative Structured Sparse Representation. IEEE Transactions on Image Processing, Vol. 25, 2016, No. 5, pp. 2337–2352.

[35] Qiao, L.—Chen, S.—Tan, X.: Sparsity Preserving Projections with Applications to Face Recognition. Pattern Recognition, Vol. 43, 2010, No. 1, pp. 331–341.

[36] Yin, F.—Jiao, L.—Shang, F.—Xiong, L.—Wang, X.: Sparse Regularization Discriminant Analysis for Face Recognition. Neurocomputing, Vol. 128, 2014, pp. 341–362, doi: 10.1016/j.neucom.2013.08.032.

[37] Lai, Z.—Xu, Y.—Yang, J.—Tang, J.—Zhang, D.: Sparse Tensor Discriminant Analysis. IEEE Transactions on Image Processing, Vol. 22, 2013, No. 10, pp. 3904–3915.

[38] Guan, N.—Tao, D.—Luo, Z.—Shawe-Taylor, J.: MahNMF: Manhattan Non-Negative Matrix Factorization. arXiv preprint arXiv:1207.3438, 2012.

[39] Lu, G.-F.—Jin, Z.—Zou, J.: Face Recognition Using Discriminant Sparsity Neighborhood Preserving Embedding. Knowledge-Based Systems, Vol. 31, 2012, pp. 119–127, doi: 10.1016/j.knosys.2012.02.014.

[40] Zang, F.—Zhang, J.: Discriminative Learning by Sparse Representation for Classification. Neurocomputing, Vol. 74, 2011, No. 12, pp. 2176–2183.

[41] Tikhonov, A. N.—Arsenin, V. Y.: Solutions of Ill-Posed Problems. Winston, 1977.

[42] Van Overschee, P.—De Moor, B.: Subspace Identification for Linear Systems: Theory-Implementation-Applications. Springer Science & Business Media, 2012.

[43] Cootes, T.—Taylor, C.: Anatomical Statistical Models and Their Role in Feature Extraction. The British Journal of Radiology, 2014.

[44] Lou, S.—Zhang, G.—Pan, H.—Wang, Q.: Supervised Laplacian Discriminant Analysis for Small Sample Size Problem with Its Application to Face Recognition. Journal of Computer Research and Development, 2012, No. 8.

[45] Samaria, F. S.—Harter, A. C.: Parameterisation of a Stochastic Model for Human Face Identification. Proceedings of the Second IEEE Workshop on Applications of Computer Vision, 1994, pp. 138–142, doi: 10.1109/ACV.1994.341300.

[46] Lee, K.-C.—Ho, J.—Kriegman, D. J.: Acquiring Linear Subspaces for Face Recognition under Variable Lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, 2005, No. 5, pp. 684–698.

[47] Baraniuk, R. G.—Cevher, V.—Duarte, M. F.—Hegde, C.: Model-Based Compressive Sensing. IEEE Transactions on Information Theory, Vol. 56, 2010, No. 4, pp. 1982–2001.

[48] Golub, G. H.—Van Loan, C. F.: Matrix Computations. 3^{rd} Ed. JHU Press, 2012.

Yingchun Ren is currently a Ph.D. candidate in the Research Center of CAD at Tongji University, Shanghai, China. He received his B.Sc. degree in mathematics from Zhengzhou University, Zhengzhou, in 2005 and the M.Sc. degree in applied mathematics from the University of Shanghai for Science and Technology, Shanghai, in 2008. His research interests include pattern recognition, machine learning and computer vision.

Yufei Chen is presently a Senior Lecturer in the CAD Research Center of Tongji University. She was a postdoctoral researcher in control science and engineering of Tongji Univerisity from 2010 to 2012. She received her Ph.D. degree from Tongji Univerisity in 2010. She was also a guest researcher in Fraunhofer Institute for Computer Graphics Research, Germany, from 2008 to 2009. Her research topics include image processing and data analysis.

Xiaodong Yue is presently a Senior Lecturer in the School of Computer Engineering and Science, Shanghai University, China. Before joining Shanghai University, he received his Ph.D. degree from Tongji University, China, and worked as a postdoctoral researcher in the Department of Computer Science and Technology at Tongji University. He was a research assistant in the Department of Computer Science at Hong Kong Baptist University in 2009 and also a research fellow in the Faculty of Engineering and Information Technology at University of Technology in Sydney (UTS), Australia, from 2011 to 2012. His research topics include pattern recognition, soft computing and multimedia and he has published more than thirty international journal and conference papers.