Sparse Approximation to the Eigensubspace for Discrimination

Zhihui Lai, Wai Keung Wong, Zhong Jin, Jian Yang, and Yong Xu, *Member, IEEE*

***Abstract*— Two-dimensional (2-D) image-matrix-based projection methods for feature extraction are widely used in many fields of computer vision and pattern recognition. In this paper, we propose a novel framework called sparse 2-D projections (S2DP) for image feature extraction. Different from the existing 2-D feature extraction methods, S2DP iteratively learns the sparse projection matrix by using elastic net regression and singular value decomposition. Theoretical analysis shows that the optimal sparse subspace approximates the eigensubspace obtained by solving the corresponding generalized eigenequation. With the S2DP framework, many 2-D projection methods can be easily extended to sparse cases. Moreover, when each row/column of the image matrix is regarded as an independent high-dimensional vector (1-D vector), it is proven that the vector-based eigensubspace is also approximated by the sparse subspace obtained by the same method used in this paper. Theoretical analysis shows that, when compared with the vector-based sparse projection learning methods, S2DP greatly saves both computation and memory costs. This property makes S2DP more tractable for real-world applications. Experiments on well-known face databases indicate the competitive performance of the proposed S2DP over some 2-D projection methods when facial expressions, lighting conditions, and time vary.**

***Index Terms*— Elastic net, face recognition, feature extraction, manifold learning, sparse subspace.**

I. Introduction

TECHNIQUES for linear dimensionality reduction in supervised or unsupervised learning tasks have attracted much attention in the fields of computer vision and pattern recognition. In the past 20 years, many dimensionality reduction methods have been developed to deal with the high-dimensional data, including scene images, face images, palmprint images, and handwritten character images. In the traditional linear dimensionality reduction (or linear feature extraction) methods, one usually first transforms the image matrices to high-dimensional vectors and then projects these high-dimensional vectors into a low-dimensional subspace for feature extraction and classification.

Among the traditional linear dimensionality reduction techniques, principal component analysis (PCA) [1]–[3] and linear discriminant analysis (LDA) [4], [5] are the most frequently used methods in many application fields. However, the resulting high-dimensional image vector space in these traditional feature extraction methods makes it difficult to accurately evaluate the covariance matrices because of the relatively small number of training samples [6], [7]. Furthermore, computing the eigenvectors of a large-size covariance matrix is very time consuming. Another problem is that LDA always encounters the singularity of the within-class scatter matrix when there are small sample size (SSS) problems, which further makes it intractable or unstable to compute the generalized eigenvectors.

In order to avoid the above problems, Yang *et al.* [6] proposed the well-known 2DPCA for image feature extraction. Different from PCA, 2DPCA is based on 2-D image matrices rather than 1-D image vectors. By defining the image-matrix-based covariance matrix, the optimal projections of 2DPCA can be obtained by eigendecomposition of a very small-size matrix. Xu *et al.* [7] showed the theoretical similarities and differences between 2DPCA and PCA. Motivated by 2DPCA, many image-matrix-based methods are proposed for feature extraction. Zuo *et al.* [8] proposed the bidirectional PCA with assembled matrix distance metric for image recognition. Ye [9] proposed the generalized low-rank matrix approximation (GLRMA) for face image feature extraction. 2-DPCA and GLRMA are unsupervised methods, and the classification abilities may be limited, since the label information is not used in the learning steps. By integrating the label information in constructing the within-class and between-class image covariance matrices for discrimination, 2-D linear discriminant analysis (2DLDA) [10], which is a supervised method, can achieve higher classification accuracy in many applications. Based on the same idea of 2DLDA, bilateral 2DLDA [11] is also proposed by using the iteration method, which frequently solves the generalized eigenequations to obtain optimal bilateral projections for feature extraction. The essential idea developed is that classical LDA operates on the image vectors, which are the tensors with order 1, and 2DLDA operates on

Manuscript received May 13, 2011; revised August 20, 2012; accepted August 22, 2012. Date of publication October 23, 2012; date of current version November 20, 2012. This work was supported in part by the General Research Fund of Research Grants Council of Hong Kong under Project 531708, the Hong Kong Polytechnic University Project G-YH24, and the Natural Science Foundation of China under Grant 61203376, Grant 60973098, Grant 61005005, Grant 61071179, and Grant 61125305, the Hi-Tech Research and Development Program of China under Grant 2006AA01Z119, the China Postdoctoral Science Foundation under Project 2012M511479, and the Guangdong Natural Science Foundation under Project S2012040007289.

Z. Lai is with Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China. He was also with School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: lai_zhi_hui@163.com).

W. K. Wong is with Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hong Kong (e-mail: calvin.wong@polyu.edu.hk).

Z. Jin and J. Yang are with School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zhongjin@mail.njust.edu.cn; csjyang@mail.njust.edu.cn).

Y. Xu is with Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: laterfall286@yahoo.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2012.2217154

image matrices, which are regarded as second-order tensors. A natural way is to extend the linear projection methods to higher order tensors. Thus, the tensor-based discriminant analysis methods [12]–[14] were proposed for feature extraction and recognition.

Although there are a number of variations of the image-matrix-based LDA method, the most fundamental ones are still 2DPCA and 2DLDA. However, PCA, LDA, and their 2-D versions cannot preserve the local geometric structure of the image data set. Recent research shows that the high-dimensional data such as face images lie on a low-dimensional nonlinear manifold. How to preserve the manifold's local geometric structure has been an important research field in the past 10 years. Representative nonlinear manifold learning methods were proposed in [15]–[18], and some robust nonlinear methods can be found in [19] and [20]. A tractable method to make the nonlinear techniques more suitable for real-world applications is to learn the explicit linear mapping which can also preserve the geometric structure of the manifold. Therefore, locality preserving projections (LPP) [21]–[23], the linear approximation of the Laplacian eigenmap [17], was proposed to learn the low-dimensional subspace for feature extraction. LPP preserves the local manifold structure modeled by a nearest-neighbor (NN) graph of the high-dimensional data. With the explicit maps, the training and test data can all be directly projected to the low-dimensional subspace for visualization, feature extraction, and classification purposes. From analysis on the essence of the LPP, it can be found that LPP can be seen as a generalization of LDA with the same SSS problems [23]. Thus, LPP cannot be implemented directly because of the singularity of the matrix in its generalized eigenequation [24]. Motivated by the 2DPCA and 2DLDA, which operate directly on 2-D image matrices, 2-D locality preserving projection (2DLPP) [25]–[27] was proposed for linear dimensionality reduction. Recently, some new improved versions of 2DLPP [24], [28]–[30] were also proposed.

However, the classical and the manifold learning-based linear projection methods have limitations. Since the learned projections are the nonzero linear combiner of all the image vectors or image matrices, they regard each variable of the patterns (image vectors or image matrices) as equally important in dimensionality reduction. The important features or variables in the low-dimensional space reduced by such "dense" projections may be destroyed or submerged, thus the classification accuracy may be affected. For simplicity, the projections obtained from directly solving the eigenequation and thus containing almost nonzero elements are called dense projections in this paper.

In recent years, many feature selection techniques have been developed to explore the important information (or say variables/factors in statistic learning) for dimensionality reduction. Sparse PCA (SPCA) [31] was proposed by using the least angle regression [32] and L_1-norm elastic net [33] regression to obtain sparse principle components. d'Aspremont *et al.* [34] relaxed the hard cardinality constraint and obtained a convex approximation using semidefinite programming. In [35] and [36], Moghaddam *et al.* proposed a spectral bounds framework for sparse subspace learning. Particularly, they proposed both exact and greedy algorithms for sparse PCA and sparse LDA, though their sparse LDA can only be applied to the two-class problem. Recently, sparse discriminant analysis (SDA) [37] was proposed for feature extraction. And sparse linear discriminant analysis (SLDA) [38], which combines Lasso regression [39] and SVD to learn the sparse projections, was also developed to deal with the data piling problem. With the same way as SLDA for obtaining the sparse projections, sparse locality-preserving embedding (SLPE) [40] was also proposed for visualization. The recently proposed manifold elastic net (MEN) [41] integrates sparse regression, manifold learning, and dimensionality reduction simultaneously.

However, existing sparse linear projection methods such as SPCA, SDA, SLDA, SLPE, and MEN all operate on the high-dimensional image vectors instead of 2-D image matrices, in which some useful structural information embedded in the original 2-D images may be lost. Another significant disadvantage of the existing sparse feature extraction methods is that it is rather time consuming to directly operate on the very high-dimensional vectors with a large cardinality. For SDA, SLDA, and SLPE, because there are the small sample size problems or the singularities of the local or within-class scatter matrices, it is very difficult to give the theoretical connections between the sparse subspace learning algorithms and the existing dimensionality reduction techniques. The theorems in [38] and [40] show that if the (local) within-class scatter matrices are invertible then the eigensubspace of the corresponding generalized eigenequation is approximated by the ridge regression subspace rather than the sparse subspace. However, in SSS problems, the (local) within-class scatter matrices are always not invertible. The frequently used method for obtaining the sparse solutions is to directly add the L_1-norm term to produce sparse projections, but the effectiveness of such methods may not be guaranteed since they operate on the very high-dimensional image vectors. In addition, it is still unclear for these sparse learning methods to provide a theoretical analysis that the learned sparse subspace approximates the one spanned by the eigenvectors of the corresponding eigenequation.

In order to enhance the discriminative ability of 2-D-based projection methods, we propose a novel method called sparse 2-D projections (S2DP) for feature extraction. S2DP combines the L_1-norm elastic net regression and SVD to iteratively learn the sparse projections instead of solving the generalized eigenequation. The essential difference between the existing image-vector-based sparse learning methods and S2DP is that our method directly operates on the image matrix. The contribution of this paper is threefold. First, a theoretical analysis between the well-known and widely evaluated 2-D-based projection methods and the proposed S2DP is conducted. We show that the subspace of 2-D-based projection methods can be approximated by the sparse subspace of S2DP, which guarantees the discriminative ability of the S2DP. Second, by using the proposed framework as a platform, some nonsparse image-based feature extraction methods can be easily extended to sparse cases. Thus, the proposed framework generalizes the nonsparse

2-D linear dimensionality reduction methods into sparse cases. Third, we show that our theorems can also provide the theoretical analysis for the vector-based methods such as SDA, SLDA, and SLPE when each row/column of the weighted image matrix is viewed as an independent pattern vector. Thus, the theorems in [38] and [40] are the special cases in this paper. Finally, the discriminative ability of the sparse projections was evaluated in some well-known databases, which indicates that the proposed S2DP is competitive over some 2-D-based projection methods.

The rest of this paper is organized as follows. In Section II, we briefly review 2DLPP and 2DLGEDA. S2DP algorithm and related analysis are described in Section III. In Section IV, experiments are carried out to evaluate our S2DP algorithm. The conclusions are given in Section V.

II. 2-D Feature Extraction Methods

A. 2-D LPP

2DLPP works directly on 2-D image matrices and was proposed in [25]–[27] as an extension of LPP. Assume that $X_i \in R^{n_1 \times n_2}$ $(i = 1, 2, \ldots, m)$ are the 2-D image matrices of the training images and $n_1 > n_2$. Suppose Φ is an $n_2 \times d$ matrix, where each column of Φ is an n_2-dimensional unit vector and $d \leq n_2$. The purpose of 2DLPP is to seek an optimal projective matrix $\Phi = [\varphi_1, \varphi_2, \ldots, \varphi_d]$ and map a 2-D image from $n_1 \times n_2$-dimensional image space into an $n_1 \times d$-dimensional Euclidean space by the following linear projection:

$$Y_i = X_i\Phi = X_i[\varphi_1, \ldots, \varphi_d], \qquad (i = 1, 2, \ldots, m). \tag{1}$$

The objective function of 2DLPP is to preserve the 2-D images' local similarities in the projective space by solving the following optimization problem:

$$\begin{aligned} \arg\min_{\Phi} \sum_{i=1}^{m}\sum_{j=1}^{m} \|Y_i - Y_j\|_F^2 W_{ij} \\ = \arg\min_{\Phi} \sum_{i=1}^{m}\sum_{j=1}^{m} \|X_i\Phi - X_j\Phi\|_F^2 W_{ij} \end{aligned} \tag{2}$$

where $\|\cdot\|_F$is Frobenius norm and W_{ij} is the similarity defined as

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|X_i - X_j\|_F^2}{t}\right), & X_i \in N_k(X_j) \text{ or } X_j \in N_k(X_i) \\ 0, & \text{otherwise} \end{cases}$$

where $N_k(X_i)$ denotes the k NNs of X_i, and t is the Gaussian kernel parameter.

After some matrix analysis steps, the minimization problem of (2) with a constraint [25]–[27] can be equivalently represented as

$$\arg\max_{\varphi} \varphi^T X^T (W \otimes I_{n_1}) X\varphi \tag{3}$$

$$\text{s.t.} \quad \varphi^T X^T (D \otimes I_{n_1}) X\varphi = 1 \tag{4}$$

where $X = [X_1^T, X_2^T, \ldots, X_m^T]^T$ is the 2-D image training sample matrix of size $mn_1 \times n_2$, and D is a diagonal matrix whose entries are column or row sums of W; I_{n_1} is an identity matrix of order n_1; operator $\otimes$ is the Kronecker product of the matrices.

The optimal d projections that maximize the objective function are computed by the maximum eigenvalue solutions to the generalized eigenvalue problem [25]–[27]

$$X^T (W \otimes I_{n_1}) X\varphi = \lambda X^T (D \otimes I_{n_1}) X\varphi \tag{5}$$

where φ is the eigenvector corresponding to eigenvalue λ.

From the derivation of 2DLPP, it can be found that the main idea of 2DLPP is to preserve the local similarities of the date set, that is, keep the connected points stay as close together as possible and thus the intrinsic local geometric structure of the image manifold is preserved. However, since the label information is not used, the performance of 2DLPP is limited. Thus the following 2DLGEDA algorithm was proposed to address this problem.

B. 2-D Local Graph Embedding Discriminant Analysis

The goal of 2DLGEDA [29] is to preserve the 2-D image within-class compactness and maximize the between-class separability. The main idea of 2DLGEDA is to use the label information to construct the local within-class graph and local between-class graph to reflect the compactness and separablity of the image manifold. 2-D image within-class compactness is characterized from the intrinsic graph by the term

$$\begin{aligned} S_w &= \sum_{i=1}^{m}\sum_{j=1}^{m} \|Y_i - Y_j\|_F^2 W_{ij}^w = \sum_{i=1}^{m}\sum_{j=1}^{m} \|X_i\Phi - X_j\Phi\|_F^2 W_{ij}^w \\ &= 2tr(\Phi^T X^T ((D^w - W^w) \otimes I_{n_1}) X\Phi) \\ &= 2tr(\Phi^T X^T (L^w \otimes I_{n_1}) X\Phi) \\ W_{ij}^w &= \begin{cases} 1, & X_i \in N_{k_w}^+(X_j) \text{ or } X_j \in N_{k_w}^+(X_i) \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

where $tr(\cdot)$ denotes the trace of a matrix and D^w is a diagonal matrix whose entries are column or row sums of W^w; $N_k^+(X_i)$ indicates the samples in the k_w NNs of X_i in the same class, and $L^w = D^w - W^w$.

Similarly, 2-D image between-class separability is characterized from the between-class graph by the term

$$\begin{aligned} S_b &= \sum_{i=1}^{m}\sum_{j=1}^{m} \|Y_i - Y_j\|_F^2 W_{ij}^b = \sum_{i=1}^{m}\sum_{j=1}^{m} \|X_i\Phi - X_j\Phi\|_F^2 W_{ij}^b \\ &= 2tr(\Phi^T X^T ((D^b - W^b) \otimes I_{n_1}) X\Phi) \\ &= 2tr(\Phi^T X^T (L^b \otimes I_{n_1}) X\Phi) \\ W_{ij}^b &= \begin{cases} 1, & \text{if } (i, j) \in P_{k_b}(c_i) \text{ or } (i, j) \in P_{k_b}(c_j) \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

where c_i is the class label of X_i, and $P_{k_b}(c_i)$ is a set of data pairs that are in the k_b NN pairs among the set $\{(i, j)|i \in \pi_{c_i}, j \notin \pi_{c_i}\}$, where π_{c_i} denotes the index set of the samples with label c_i. D^b is a diagonal matrix whose entries are column or row sums of W^b, and $L^b = D^b - W^b$.

The criterion of 2DLGEDA is formally similar to the Fisher criterion since they are both Rayleigh quotients, and the optimal projections can be obtained from solving the generalized eigenequation

$$X^T (L^b \otimes I_{n_1}) X\varphi = \lambda X^T (L^w \otimes I_{n_1}) X\varphi \tag{8}$$

where λ is the generalized eigenvalue corresponding to the eigenvector φ. The optimal transformation matrix of 2DLGEDA is composed of the eigenvectors associated with the d largest eigenvalues.

C. 2-D Feature Extraction Framework

2DLGEDA, 2DLPP, and the algorithms proposed in [28] and [30] are graph-based representative methods of the following general 2-D dimensionality reduction framework:

$$X^T(L_b \otimes I)X\varphi = \lambda X^T(L_w \otimes I)X\varphi \tag{9}$$

where L_b and L_w represent the local neighborhood graph (or their graph Laplacian) defined in different ways. Note that the notations in (9) are different from (8) since we use subscripts instead of superscripts. When $L_b = L^b$ and $L_w = L^w$, (9) becomes (8), which is the generalized eigenequation of 2DLGEDA. When $L_b = L$ and $L_w = D$, (9) becomes the generalized eigenequation of 2DLPP. Many existing 2-D-based methods such as those in [25]–[30] can also be described in this framework.

There is an obvious disadvantage of the 2-D-based feature extraction framework. Taking all the image pixels equally in dimensionality reduction cannot explore the intrinsic information since the 2-D images are of great information redundancy. When using the 2-D-based methods for dimensionality reduction, the whole 2-D image is projected to the nonsparse optimal projection φ, that is, $Y_l = X_l\varphi$ $(l = 1, 2, \ldots, m)$, and the locality of the 2-D image manifold is preserved in the low-dimensional subspace. Since the elements in φ are nonzero, each pixel of the image matrix contributes to the low-dimensional feature Y_l. However, if φ is a sparse vector, that is, $\varphi = [0, \ldots, \bar{\varphi}_i, 0, \ldots, 0, \bar{\varphi}_j, 0, \ldots, 0]^T$ (we take two nonzero elements as an example), then $Y_l = X_l\varphi = X_l[0, \ldots, \bar{\varphi}_i, 0, \ldots, 0, \bar{\varphi}_j, 0, \ldots, 0]^T$ shows that only the ith and jth column features contribute to Y_l. If the sparseness and locality preserving criteria (or local discriminative criteria) are combined, some column features of the 2-D images in the local neighborhood can be preserved in the low-dimensional subspace. Thus, the following proposed S2DP algorithm can extract the latent intrinsic features embedded in the image matrices so as to overcome this disadvantage in the other 2-D feature extraction methods.

The following section describes how the proposed method addresses the limitations of the 2-D-based feature extraction framework.

III. S2DP

A. Motivations of S2DP

How to obtain a good classification result is an important issue in image recognition and computer vision. However, images contain much redundant information, and the discriminative information is not decided by all the pixels. As discussed in Section II-C, this paper aims to extract the latent intrinsic discriminative features from the image matrices to increase the classification performance and enhance the robustness of the algorithm. Recent research [31]–[33], [35]–[38] indicates that introducing the L_1-norm can enhance discriminative feature selection and also avoid overfitting and thus improve the prediction accuracy. Introducing the L_1-norm regression to learn the sparse discriminant projections seems to be a good choice to accomplish the purpose desired. Thus, the natural idea is to append the sparseness constraint to the existing 2-D-based method and guarantee the effectiveness of the sparse projections from the theoretical aspect.

B. Model of S2DP

The problem at hand is to find the optimal sparse projections that approximate the optimal (discriminant) vectors of the generalized eigenequation of (9). Our idea is to impose a sparseness constrained condition in (9) as a general framework. The model of S2DP is given as follows:

$$\begin{cases} X^T(L_b \otimes I)X\varphi = \lambda X^T(L_w \otimes I)X\varphi \\ \text{s.t.} \quad \text{Card}(\varphi) \le K \end{cases} \tag{10}$$

where φ is the column vector corresponding to eigenvalue λ and $\text{Card}(\varphi)$ denotes the number of nonzero elements of φ.

The only difference between (9) and (10) is that a sparseness constraint is imposed in (10). Directly solving the generalized eigenfunction of (9) cannot obtain the sparse projections. Therefore, in order to obtain the sparse solutions by using L_1-norm regression, we should rewrite the above equation $X^T(L_b \otimes I)X\varphi = \lambda X^T(L_w \otimes I)X\varphi$. The following subsections develop some properties about the Kronecker product of the graph Laplacian matrices, which help reduce the computation cost and give representations of the two terms $X^T(L_b \otimes I)X$ and $X^T(L_w \otimes I)X$ in (10).

C. Idea and Background for Obtaining the Sparse Projections

Directly solving the generalized eigenequation (10) with a sparseness constraint is a difficult problem, but we can represent and transfer this problem by using the L_1-norm regression to obtain the sparse subspace, which also approximates the subspace spanned by the generalized eigenvectors associated with the larger (or smaller) eigenvalues. Thus, the discriminative abilities or locality preserving abilities can be maintained by the sparse projections.

SDA [37] aims to minimize the regression error of the class index matrix by iteratively using the L_1-norm regression method and provides us with the most direct idea for obtaining the sparse projections. However, a significant drawback of SDA is that it is difficult to conduct a theoretical analysis on the connections of the sparse projections of SDA and the dense projections of LDA.

In [38] and [40], a theoretical analysis was conducted on the relationships between the iterative ridge regression subspace and the corresponding eigensubspace. Then L_1-penalized term was directly added to the iterative ridge regression procedures, in which no theory is presented to analyze the latent relationship between the ridge regression and the complex norm regression. In addition, the proof of the theorems in [38] and [40] needs a strict assumption that vector-based within-class scatter matrix must be positive definite. This assumption usually cannot be satisfied because of SSS

problems. Under the background of 2-D image-based methods, however, this assumption or condition can be satisfied naturally. This is the other reason why the 2-D-based theoretical analysis was examined in this paper. Our theoretical analysis presented in Section III-E shows that the theorems in [38] and [40] can be regarded as a special case of the Theorem 2 and is represented as the propositions.

As seen from Sections III-B and III-E, the most essential difference between SDA and the proposed framework is that the proposed sparse projection framework directly works on the image matrices which construct the corresponding generalized eigenequation. It will be shown that the sparse projections approximate the dense projections of the corresponding 2-D projection methods. On the one hand, the connections of theoretical analysis between dense projections and sparse projections are bridged. On the other hand, the effectiveness of the sparse projections of S2DP can be guaranteed since the sparse projection subspace approximates the dense projection subspace and the effectiveness of 2-D-based dense projection methods have been widely evaluated. Since the most important factors are selected by sparse regression for discrimination, the sparse feature selection can enhance the performance of the algorithm.

D. Singular Value Decomposition on Graph Laplacian With Kronecker Product

The Kronecker product has the following properties.

Lemma 1: [42] $(A \otimes B)(C \otimes D) = AC \otimes BD$.

Corollary 1: $(A \otimes B)(C \otimes D)(E \otimes F) = ACE \otimes BDF$.

Corollary 2: $(A \otimes I)(C \otimes I)(E \otimes I) = ACE \otimes I$.

With the above preparation, we derived theorem 1.

Theorem 1: Denote the singular value decomposition of the symmetric graph Laplacian L_* as $L_* = UDU^T$, then $L_* \otimes I = (UDU^T) \otimes I = (UD^{1/2} \otimes I)((UD^{1/2})^T \otimes I)$.

From Theorem 1, instead of directly performing the SVD on the large-size graph Laplacian matrix $L_* \otimes I$, we can perform the SVD on the small-size graph Laplacian matrix L_*, and then perform Kronecker products on them, which will greatly save the computational cost.

After we perform SVD on the graph Laplacian matrices, we can directly give the decompositions of $X^T(L_b \otimes I)X$ and $X^T(L_w \otimes I)X$ in (9) and denote them as

$$\begin{aligned} M_w &= X^T(L_w \otimes I)X = F_w F_w^T \\ &= X^T(U_w D_w^{1/2} \otimes I)((U_w D_w^{1/2})^T \otimes I)X \qquad (11) \\ M_b &= X^T(L_b \otimes I)X = F_b F_b^T \\ &= X^T(U_b D_b^{1/2} \otimes I)((U_b D_b^{1/2})^T \otimes I)X \qquad (12) \end{aligned}$$

where $F_w = X^T(U_w D_w^{1/2} \otimes I)$ and $F_b = X^T(U_b D_b^{1/2} \otimes I)$.

E. Optimal Sparse Solutions

Based on the idea mentioned in Sections III-A and III-C, the key problem for obtaining the sparse solutions is to transfer the generalized eigenequation with sparseness constraint into the L_1-norm regression form and guarantee that the optimal eigensubspace spanned by the generalized eigenvectors associated with the larger (or smaller) eigenvalues of (9) are optimally approximated by the learned sparse subspace. To prove the theorem, we need the following conclusions.

Lemma 2: For any column orthogonal $m \times d$ $(m > d)$ matrix V satisfying $V^TV = I$, V can be represented as $V = AU^T$, where $m \times d$ matrix A has the orthogonal columns and U is an arbitrary orthogonal $d \times d$ matrix.

Proof: $V^TV = I = UIU^T = UA^TAU^T = (AU^T)^T AU^T$.

Lemma 3: Let M be an $m \times m$ symmetric positive semidefinite matrix; if $V = \arg \max_{V^TV=I} Tr(V^TMV)$, then for any arbitrary orthogonal matrix U, there exists $m \times d$ matrix A^* such that $V = A^*U^T$ and A^* satisfies $A^* = \arg \max_{A^TA=I} Tr(A^TMA)$.

Proof: From Lemma 2, we know that there are U and A^* satisfying $V = A^*U^T$; since U is the orthogonal matrix, let $V = AU^T$, it is easy to see that the optimization problem of $\arg \max_{V^TV=I} Tr(V^TMV)$ can be converted to

$$\begin{aligned} &\max Tr[(AU^T)^TM(AU^T)] \quad \text{s.t. } (AU^T)^T(AU^T) = \mathrm{I} \\ &\Leftrightarrow \max Tr(A^TMA) \quad \text{s.t. } AA = \mathrm{I}. \end{aligned}$$

Therefore, if $V = A^*U^T$ be the optimal solution of $\arg \max_{V^TV=I} Tr(V^TMV)$, then A^* also satisfies the second optimization problem, that is, $A^* = \arg \max_{A^TA=\mathrm{I}} Tr(A^TMA)$. ■

Lemma 4: Assume that the Cholesky decomposition of the positive definite matrix M_w is $M_w = G_wG_w^T$, and denote the singular value decomposition of $G_w^{-1}M_bG_w^{-T} = \bar{V}\bar{\Lambda}\bar{V}^T$, where $\bar{\Lambda}$ is an $N \times N$ diagonal matrix of eigenvalues with decreasing order, then $G_w^{-T}\bar{V}$ is the eigenvector matrix of the generalized eigenfuction (9) corresponding to the decreasing ordered eigenvalues.

Proof:

$$\begin{aligned} G_w^{-1}M_bG_w^{-T} &= \overline{V\Lambda V}^T \Rightarrow G_w^{-T}G_w^{-1}M_bG_w^{-T} = G_w^{-T}\overline{V\Lambda V}^T \\ &\Rightarrow (G_wG_w^T)^{-1}M_bG_w^{-T}\overline{V} = G_w^{-T}\overline{V\Lambda} \\ &\Rightarrow (M_w)^{-1}M_bG_w^{-T}\overline{V} = G_w^{-T}\overline{V\Lambda}. \end{aligned}$$ ■

From Lemma 4, we directly have Corollary 3.

Corollary 3: The first d eigenvectors corresponding to the largest eigenvalues of the generalized eigenfuction (9) are exactly the first d columns of $G_w^{-T}\overline{V}$.

With the above preparation, we derived Theorem 2 as the main theory in this paper. We only give the theorem and its proof in the case of the eigenvectors corresponding to the larger eigenvalues with (9). For the case of the eigenvectors corresponding to the smaller eigenvalues, it is easy to obtain similar results in the same way as in this paper.

Theorem 2: Since M_w is positive definite, its Cholesky decomposition can be presented as $M_w = G_wG_w^T$, where G_w is an $n_2 \times n_2$ lower triangular matrix with full rank. Let $\Phi = [\varphi_1, \varphi_2, \ldots, \varphi_d]$ be the eigenvectors of (9) associated with the first d largest eigenvalues. Without misleading, we still let $P = [p_1, p_2, \ldots, p_d]$ and $Q = [q_1, q_2, \ldots, q_d]$ be the

optimal solutions to the following regression problem:

$$(P, Q) = \arg\min_{P,Q} \sum_{i=1}^{mn_1} \left\| G_w^{-1} F_b(:,i) - P Q^T F_w(:,i) \right\|^2 + \alpha \sum_{j=1}^{d} q_j^T M_w q_j + \beta \sum_{j=1}^{d} \|q_j\|^2 + \gamma \sum_{j=1}^{d} |q_j| \quad (13)$$

$$\text{s.t. } P^T P = I \quad (14)$$

where $\alpha \geq 0$, $\beta \geq 0$, $\gamma \geq 0$, $F_b(:,i)$ and $F_w(:,i)$ are the ith column of F_b and F_w, and $\|\cdot\|$, $|\cdot|$ denote the L_2- and L_1-norm, respectively. Then the columns of Q span the linear space approximating Φ when $\beta \to 0_+$ and $\gamma \to 0_+$.

Proof: The proof of the theorem was divided into two steps, and the representation of the sparse subspace Q was obtained by using the generalized eigenvectors of (9), that is, Φ.

Step 1: Suppose P is given.

If P is fixed, the update of Q is a regression-type problem. There exists an orthogonal matrix $P_\perp$ such that $[P, P_\perp]$ is $n_2 \times n_2$ column orthogonal matrix. The first part of (13) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{mn_1} \left\| G_w^{-1} F_b(:,i) - P Q^T F_b(:,i) \right\|^2 = \left\| G_w^{-1} F_b - P Q^T F_b \right\|^2 \\
&= \left\| F_b^T G_w^{-T} - F_b^T Q P^T \right\|^2 \\
&= \left\| F_b^T G_w^{-T} [P, P_\perp] - F_b^T Q P^T [P, P_\perp] \right\|^2 \\
&= \left\| [F_b^T G_w^{-T} P, F_b^T G_w^{-T} P_\perp] - [F_b^T Q P^T P, F_b^T Q P^T P_\perp] \right\|^2 \\
&= \left\| [F_b^T G_w^{-T} P - F_b^T Q P^T P, F_b^T G_w^{-T} P_\perp - 0] \right\|^2 \\
&= \left\| F_b^T G_w^{-T} P - F_b^T Q \right\|^2 + \left\| F_b^T G_w^{-T} P_\perp \right\|^2 \\
&= \sum_{j=1}^{d} \left\| F_b^T G_w^{-T} p_j - F_b^T q_j \right\|^2 + \left\| F_b^T G_w^{-T} P_\perp \right\|^2
\end{aligned}$$

since for fixed P, $\left\| F_b^T G_w^{-T} P_\perp \right\|^2$ is a constant and can be ignored. We have the following regression problem to compute the sparse matrix Q:

$$\arg\min_{Q} \left(\sum_{j=1}^{d} \left\| F_b^T G_w^{-T} p_j - F_b^T q_j \right\|^2 + \alpha \sum_{j=1}^{d} q_j^T M_w q_j + \beta \sum_{j=1}^{d} \|q_j\|^2 + \gamma \sum_{j=1}^{d} |q_j| \right) \quad (15)$$

or

$$\arg\min_{Q} \left(\left\| F_b^T G_w^{-T} P - F_b^T Q \right\|^2 + \alpha \left\| G_w^T Q \right\|^2 + \beta \|Q\|^2 + \gamma \sum_{j=1}^{d} |q_j| \right). \quad (16)$$

Let $Y = \begin{bmatrix} F_b G_w^{-T} P \\ 0_{n_2 \times d} \end{bmatrix}$ and $\widehat{X} = \begin{bmatrix} F_b \\ \sqrt{\alpha} G_w^T \end{bmatrix}$, then (15) and (16) become the elastic net regression problem

$$\arg\min_{Q} \left(\left\| Y - \widehat{X} Q \right\|^2 + \beta \|Q\|^2 + \gamma \sum_{j=1}^{d} |q_j| \right). \quad (17)$$

When taking the limit of the above optimal problem in (16) or (17), that is, $\beta \to 0_+$, $\gamma \to 0_+$, we have

$$\arg\min_{Q} \left(\sum_{j=1}^{d} \left\| F_b G_w^{-T} p_j - F_w q_j \right\|^2 + \alpha \sum_{j=1}^{d} q_j^T M_b q_j \right). \quad (18)$$

By requiring the derivative of the above optimal problem with respect to q_i to be 0, we get

$$q_i = (M_b + \alpha M_w)^{-1} M_b G_b^{-T} p_i \quad (19)$$

or in the matrix form

$$Q = (M_b + \alpha M_w)^{-1} M_b G_b^{-T} P. \quad (20)$$

Step 2: Suppose Q is given. In this step, the proof is similar to the proof in [34] and [35] but with some variations. For the completeness of the proof and for drawing the following corollaries and conclusions, it is necessary for us to restate it more clearly.

If Q is fixed, then the other terms in (13) are constants, and the update of P becomes a Procrustes problem [43]

$$\begin{aligned}
&\sum_{i=1}^{mn_1} \left\| G_w^{-1} F_b(:,i) - P Q^T F_b(:,i) \right\|^2 \\
&= \left\| G_w^{-1} F_b - P Q^T F_b \right\|^2 \\
&= \left\| F_b^T G_w^{-T} - F_b^T Q P^T \right\|^2 \\
&= tr(F_b^T G_w^{-T} G_w^{-1} F_b + F_b^T Q Q^T F_b) - 2tr(Q^T F_b F_b^T G_w^{-T} P) \\
&\quad \text{s.t. } P^T P = I.
\end{aligned} \quad (21)$$

Thus, the update of P minimizing (21) is equivalent to maximizing the following problem:

$$tr(Q^T F_b F_b^T G_w^{-T} P) = tr(Q^T M_b G_w^{-T} P). \quad (22)$$

Substituting (20) into (22) gives

$$\max \; tr\{P^T G_w^{-1} M_b (M_b + \alpha M_w)^{-1} M_b G_w^{-T} P\} \quad (23)$$

subject to $P^T P = I$.

The middle term in (23) can be rewritten as

$$\begin{aligned}
&G_w^{-1} M_b (M_b + \alpha M_w)^{-1} M_b G_w^{-T} \\
&\quad = G_w^{-1} M_b G_w^{-T} (G_w^{-1} M_b G_w^{-T} + \alpha I)^{-1} G_w^{-1} M_b G_w^{-T}.
\end{aligned} \quad (24)$$

Denote the SVD of $G_w^{-1} M_b G_w^{-T} = V \Lambda V^T$, where Λ is a diagonal matrix with decreasing eigenvalues. From (24), we can conclude that the columns of $V$are the eigenvectors of matrix $G_w^{-1} M_b (M_b + \alpha M_w)^{-1} M_b G_w^{-T}$.

On the other hand, the update of P minimizing (21) with the constraint of $P^T P = I$ means that P is orthonormal in

the columns. Thus, the optimal P can also be directly obtained from SVD

$$G_w^{-1}F_bF_b^TQ = G_w^{-1}M_bQ = VDU^T. \tag{25}$$

According to Lemmas 2 and 3, the optimal $P$of (23) with the constraint of $P^TP = I$ satisfies

$$P = VU^T. \tag{26}$$

Substituting $P = VU^T$ into (20) gives

$$\begin{aligned} Q &= (M_b + \alpha M_w)^{-1}M_bG_w^{-T}P \\ &= (M_b + \alpha M_w)^{-1}M_bG_w^{-T}VU^T \\ &= G_w^{-T}(G_w^{-1}M_bG_w^{-T} + \alpha I)^{-1}G_w^{-1}M_bG_w^{-T}VU^T \\ &= G_w^{-T}(V\Lambda V^T + \alpha I)^{-1}V\Lambda V^TVU^T \\ &= G_w^{-T}V(\Lambda + \alpha I)^{-1}\Lambda U^T. \end{aligned} \tag{27}$$

From Lemma 4 or Corollary 3, it should be noted that the first d leading eigenvectors of the generalized eigenvalue problem of (9) are exactly the first d columns of $\Phi = G_w^{-T}V$. Therefore, for each iteration, we have

$$Q = \Phi(\Lambda + \alpha I)^{-1}\Lambda U^T. \tag{28}$$

Thus, when $\beta \to 0_+$ and $\gamma \to 0_+$, the columns of Q approximate the linear subspace Φ in each iteration. ■

According to Corollary 3 and the proof of Theorem 2, we have the following conclusions.

Corollary 4: When $d < n_2$, the first d eigensubspace $\Phi = [\varphi_1, \varphi_2, \ldots, \varphi_d]$ is approximated by the first d sparse columns of $Q(:, 1:d) = [q_1, q_2, \ldots, q_d]$.

From Theorem 2, it is easy to have the following corollary.

Corollary 5: When $\beta = 0$ and $\gamma = 0$, the linear subspace spanned by the column vectors of Φ is the same linear subspace as spanned by the columns of Q.

In fact, the vector-based discriminant analysis method is a special form of the image-matrix-based method if we view each column/row of the matrices F_w and F_b as an independent pattern vector. With the same way as the theoretical connections between image-vector-based LDA and 2DLDA presented in [44], it is easy to have the following propositions from the proof of Theorem 2.

Proposition 1: If each column (row) of the matrices F_w and F_b is regarded as an independent high-dimensional pattern vector, and vector-based scatter matrix M_w is positive definite, with the same notations as in Theorem 2, when $\beta \to 0_+$ and $\gamma \to 0_+$, the vector-based sparse subspace Q will span the linear space approximating Φ.

Proposition 2: With the same assumptions and notations as in Proposition 1, when $\beta = 0$ and $\gamma = 0$, the linear subspace spanned by the columns of Φ is the same linear subspace as spanned by the columns of Q.

When $\beta = 0$ and $\gamma = 0$, it is easy to find that Proposition 2 exactly is the same as the theorems in [38] and [40] when each row (or column) of the matrices F_w and F_b is viewed as an independent pattern vector. Thus, we give a more general theory that not only suits 2-D methods but can also be specialized to vector-based methods and thus generalizes the vector-based theorems in [38] and [40] into the sparse case and 2-D case. Therefore, Propositions 1 and 2 provide the theoretical guarantee for the effectiveness of SDA, SLDA, and SLPE when vector-based scatter matrix M_w is positive definite.

F. Algorithm Steps

From the proof of Theorem 2, it is easy to get the algorithm details of S2DP for feature extraction, which are described in the following steps.

Step 1: Construct the graph Laplacian matrices L_w and L_b.

Step 2: Perform SVD decomposition on L_w and L_b, obtain $F_w = X^T(U_wD_w^{1/2} \otimes I)$ and $F_b = X^T(U_bD_b^{1/2} \otimes I)$.

Step 3: Compute M_w, M_b, and the Cholesky decomposition of $M_w = G_wG_w^T$.

Step 4: Generate a random matrix P and orthogonalize its columns.

Step 5: Iterate until converges or achieve the number of iterations set by users.

- Step a: For the given matrix P, solve the elastic net problem of (17).
- Step b: Compute the SVD decomposition $G_w^{-1}M_bQ = VDU^T$ and update $P = VU^T$.

Step 6: Project the samples on the low-dimensional sparse subspace $Q(1:d) = [q_1, q_2, \ldots, q_d]$ to obtain the low-dimensional feature matrices $Y_i = X_iQ$ $(i = 1, 2, \ldots, m)$.

Once the feature matrix Y_i (where $Y_i = [y_{i1}, y_{i2}, \ldots, y_{id}]$ and $y_{ij}(j = 1, 2, \ldots, d)$ is the column vector in the low-dimensional feature matrix) is obtained, it can be transformed to be a vector using the formulation $Y_i' = [y_{i1}^T, y_{i2}^T, \ldots, y_{id}^T]^T$, and then a desired classifier [such as the NN classifier or support vector machine (SVM)] can be used for classification on the vector Y_i' $(i = 1, 2, \ldots, m)$.

G. Fast S2DP

In the above sections the essence of the S2DP algorithm is revealed and a general method for obtaining the optimal sparse projections was proposed. However, direct regression on the large-size matrices $F_w = X^T(U_wD_w^{1/2} \otimes I)$ and $F_b = X^T(U_bD_b^{1/2} \otimes I)$ in the iterations is very time-consuming since they are with the size $n_2 \times n_1m$. In this section, we describe the fast version of S2DP.

Denote the SVD of the following two scatter matrices as:

$$M_w = \tilde{F}_w\tilde{F}_w^T = \left(\tilde{U}_w\tilde{D}_w^{\frac{1}{2}}\right)\left(\tilde{U}_w\tilde{D}_w^{\frac{1}{2}}\right)^T \tag{29}$$

$$M_b = \tilde{F}_b\tilde{F}_b^T = \left(\tilde{U}_b\tilde{D}_b^{\frac{1}{2}}\right)\left(\tilde{U}_b\tilde{D}_b^{\frac{1}{2}}\right)^T \tag{30}$$

where $\tilde{F}_w = \tilde{U}_w\tilde{D}_w^{1/2}$ and $\tilde{F}_b = \tilde{U}_b\tilde{D}_b^{1/2}$. It is clear that the size of the matrices $\tilde{F}_w$ and $\tilde{F}_b$ is $n_2 \times n_2$, which is far less than the size of F_w and F_b. It is easy to see that when the matrices F_w and F_b are replaced by $\tilde{F}_w$ and $\tilde{F}_b$, Theorem 2 is also true. Thus we obtain the fast version of S2DP.

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

H. Comparisons of the Computational Complexity and Space Complexity

Assume that there are m training samples, and the size of the image matrix is $n \times n$ (i.e., $n_1 = n_2 = n$). The main computations of the sparse learning methods lie in the iteration procedures. For the vector-based SPCA, SDA, SLDA, and SLPE, the complexity is $O(tn^6)$ at most, where t denotes the iteration number. However, for the image-based S2DP, the computational complexity in the iteration procedures is $O(tn^3)$. It is clear that the computational complexity of S2DP is far less than that of the existing vector-based sparse learning methods. Since the elastic net and SVD are used in the iteration procedures of S2DP, S2DP is more time consuming than 2-D-based methods, such as 2DPCA and 2DLPP, whose computational complexities in solving the generalized eigenfuntion need only $O(n^3)$.

For the vector-based sparse projection learning methods, that is, SPCA, SDA, SLDA, and SLPE, the space complexity needs $O(n^4)$. However, the S2DP algorithm framework can work on the 2-D scatter matrix and needs only $O(n^2)$. In this way, S2DP greatly saves the memory cost, which is the same as the other 2-D-based dense projection methods such as 2DPCA, 2DLDA, and 2DLPP.

IV. Experiments and Analyses

To evaluate the proposed S2DP algorithm, we compared it with 2DPCA, 2DLDA, 2DLPP, and 2DLGEDA on Yale, AR, FERET, and CMU PIE face databases. The Yale database was used to examine the performance when both facial expressions and illumination were varied. The AR face database was employed to test the performance of S2DP when there was a variation in time, facial expressions, and lighting conditions. The FERET face database was used to evaluate the performance of S2DP in a larger number of individuals with variations in facial expression, illumination, and pose. The CMU PIE face database was used to evaluate the performance of these methods when facial expressions and lighting conditions varied in large ranges.

A. Experiments on Yale Face Database

The Yale face database (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 images of 15 individuals (each person providing 11 different images) with various facial expressions and lighting conditions. In our experiments, each image was manually cropped and resized to 50 × 40 pixels. Fig. 1 shows sample images of one person in the Yale database.

TABLE I
Recognition Rates (Percentage) and the Corresponding Dimensions of the Five Methods on the Yale Face Database

Methods	2DPCA	2DLDA	2DLPP	2DLGEDA	S2DP
Recognition rate (%)	88.00	88.00	92.00	90.67	97.33
Dimension	40×14	40×3	40×9	40×11	40×8

1) Properties of the S2DP:

a) Robustness of the S2DP: In the first experiment, we tested the robustness of the proposed S2DP. The first four images (i.e., center-light, with glasses, happy, left-light) were used for training, and other two images per individual were randomly selected and used for validation. The remaining five images were used for test. In the experiments, the left-light, right-light, and surprised images can be viewed as outliers, since the distances among the three images and the remaining images are far larger than that of the others. Therefore, there were outliers in the training, validation, and test set. For feature extraction, we used 2DPCA, 2DLDA, 2DLPP, 2DLGEDA, and the proposed S2DP. Regarding the parameters in 2DLPP, 2DLGEDA, and S2DP, we set the Gaussian kernel parameter $t = \infty$ for simplicity (i.e., 0–1 pattern is used in all the methods) and thus reduce the selection of the parameters. In the experiments, the neighborhood k were selected from the set $\{1, 2, 4, \ldots, m-1\}$, cardinality K and dimension d were selected from the set $\{1, 2, 3, \ldots, n_2\}$, and the weighted parameters α and β were selected from the set $\{0.01, 0.1, 1, 10, 10, 100\}$ by using the validation set. γ is not necessary to select because it can be automatically determined by the elastic net algorithm [31]. The optimal parameters of S2DP (i.e., $k = 5$, $\alpha = 1$, $\beta = 100$, $K = 6$, and $d = 8$) corresponding to the best performance of the validation set were used in the algorithm to learn the projections for feature extraction and recognition on the test set.

The recognition rates with the NN classifier and the corresponding dimensions of the five methods are shown in Table I. Fig. 2(a) shows the variations of the dimensions versus the recognition rates. Note that the dimension marked on the horizontal abscissa in the figures means the number of the projections. The recognition rates indicate that the performance of S2DP is more robust than the other methods under the facial expressions and illumination variations. The reason is that the sparse projection method selects the most important discriminative factor to form the sparse projections for dimensionality reduction and thus reduces the effect of facial expressions and illumination variations to a certain extent.

From the viewpoint of feature extraction and dimensionality reduction, the robustness comes from the sparsity of the projections. When a face image X_i with strong illumination is projected to the sparse axis φ (i.e., $Y_i = X_i\varphi = X_i[0, \ldots, \overline{\varphi}_i, 0, \ldots, 0, \overline{\varphi}_j, 0, \ldots, 0]^T$), most of the elements in image matrix X_i have no contributions to the low-dimensional feature Y_i. The sparse projection performs as the "filter" and can greatly reduce the negative effect as a result of the facial expressions and illumination variations. Thus, the

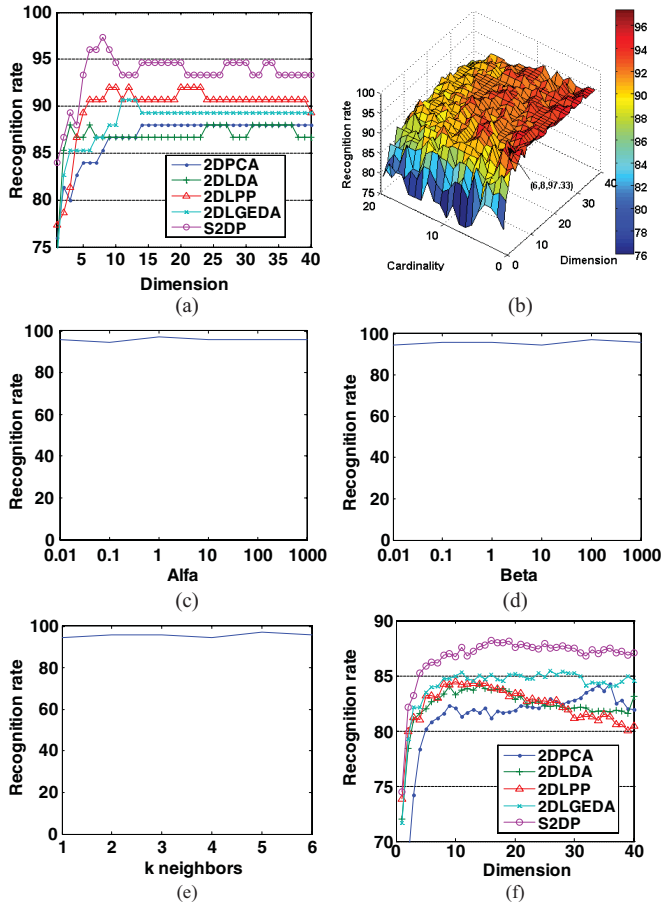

Fig. 2. (a) Recognition rate versus corresponding dimension of the five methods. (b) Variation of cardinality (i.e., K), dimension, and recognition rate of S2DP. (c) Recognition rate versus α. (d) Recognition rate versus β. (e) Recognition rate versus the neighborhood k. Note that the 1NN was used as the classifier. (f) Average recognition rates (%) versus dimensions on the Yale face database. (Classifier: 1NN).

sparse projections can obtain good performance when using low-dimensional features for classification.

b) Effectiveness and efficiency of S2DP against the vector-based sparse subspace learning algorithms: The recognition rates of S2DP, SPCA, and SLDA are 97.33%, 87.27%, and 85.45%, and the computational time [CPU: Intel(R) Core, 2.67 GHz; RAM: 4 GB] cost in training is 5.748 (1.607 for fast version of S2DP), 93.683, and 43.588 s, respectively. Thus, the vector-based sparse subspace learning algorithms achieve only lower recognition rates and need more time in training than S2DP. Although S2DP is far more efficient than the vector-based sparse subspace learning methods, it still costs more time in training procedures than the other 2-D-based methods such as 2DPCA, 2DLPP, and 2DLGEDA which cost less than 1 s in average for training. In other words, this experiment supports the theoretical analyses on the computational complexity presented in Section III-H. Since S2DP is in essence a 2-D-based method and since the vector-based sparse subspace learning methods are time consuming and perform poorly to the 2-D methods, we focus only on comparing the S2DP with the other 2-D methods in the following experiments.

TABLE II

AVERAGE RECOGNITION RATES (PERCENTAGE), STANDARD DEVIATION, AND THE CORRESPONDING DIMENSIONS OF THE FIVE METHODS ON THE YALE FACE DATABASE

Classifiers	2DPCA	2DLDA	2DLPP	2DLGEDA	S2DP
1NN	84.23	84.16	84.47	85.43	88.21
	±3.58	±2.87	±5.43	±4.63	±3.38
	40×36	40×14	40×10	40×26	40×16
3NN	83.67	83.86	82.62	84.38	86.14
	±3.31	±3.25	±4.89	±7.85	±2.78
	40×15	40×10	40×12	40×40	40×11
MD	84.20	84.70	84.72	84.13	86.17
	±2.47	±3.65	±4.63	±4.88	±3.21
	40×11	40×35	40×18	40×16	40×6
SVM	86.16	86.38	86.86	87.97	89.53
	±2.45	±3.89	±4.41	±4.87	±1.72
	40×21	40×16	40×8	40×24	40×7

c) Influence of the parameters: In order to further investigate the properties of the S2DP algorithm, the recognition rates of S2DP versus the variations of cardinality (i.e., K) and dimension are shown in Fig. 2(b), which indicates that S2DP can achieve the top recognition rate with only six nonzero elements (i.e., $\text{Card}(\varphi) = 6$). Fig. 2(c)–(e) also show that S2DP is robust to the parameters α, β, k, respectively. That is, the top recognition rates of S2DP have no significant variations when different parameters are used. Similar properties also exist in other databases.

2) Performance of the S2DP: In this section, we evaluate the performance of the proposed method using the first 10 images of each person. Two images per individual were randomly selected as the training set and validation set, respectively. The remaining six images of each individual were used for test. For each run, the parameters were selected by using the same method as in Section IV-A ($k = 6$, $\alpha = 0.01$, $\beta = 0.1$, and $K = 7$ were the optimal parameters of S2DP in average). The experiments were independently repeated 10 times for avoiding the bias of the random experiments, and four classifiers, that is, 1NN classifier, 3NN classifier, minimum distance (MD) classifier, and SVM with linear kernel, were used for classification. The average recognition rate of each method and the corresponding dimension are given in Table II. The results of the average recognition rates (%) by using 1NN as the classifier versus the different dimensions are shown in Fig. 2(f). Experimental results show that S2DP performs better than the other 2-D methods with different classifiers. On the one hand, since S2DP introduces the elastic net or L_1-norm regression to perform feature selection, the most important features/factors are selected to form the sparse projections according to the criterion. On the other hand, since the 2-D images are of great information redundancy, the low-dimensional features obtained from the dense projections still contain much redundant information, which degrades their performance. Thus S2DP is far more effective than other methods.

B. Experiments on AR Face Database

The AR face database [45] contains over 4000 color face images of 126 people (70 men and 56 women), including

Fig. 3. Sample images of one person on the AR face database.

TABLE III
RECOGNITION RATES (PERCENTAGE) AND THE CORRESPONDING DIMENSIONS OF THE FIVE METHODS ON THE AR FACE DATABASE

Classifiers	2DPCA	2DLDA	2DLPP	2DLGEDA	S2DP
1NN	58.13	56.98	58.13	58.33	62.08
	20×19	20×17	20×16	20×14	20×16
3NN	58.13	57.71	58.02	58.13	60.62
	20×19	20×16	20×17	20×14	20×17
MD	54.90	55.73	51.98	56.77	58.13
	20×20	20×18	20×19	20×20	20×19
SVM	60.31	58.96	61.25	61.15	65.10
	20×18	20×20	20×19	20×20	20×18

frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of 120 individuals (65 men and 55 women) were taken in two sessions (separated by 2 weeks), and each section contained 13 color images. Twenty images of these 120 individuals were selected and used in the two experiments. The face portion of each image was manually cropped and then normalized to 25 × 20 pixels for computational efficiency. The sample images of one person are shown in Fig. 3.

In order to test the performance of the 2-D methods when there were variations in time, facial expressions, lighting conditions, and occlusions, 10 images in the first section (the images in the first line in Fig. 3) were used as the training set and two images in the second section (images in the second line in Fig. 3) were randomly selected and used as the validation set. The remaining images were used for the test. The method for selecting the optimal parameters in each algorithm was the same as in Section IV-A The optimal parameters of S2DP were set as $k = 6$, $\alpha = 0.01$, $\beta = 0.01$, and $K = 8$. The top recognition rates and the corresponding dimensions with four classifiers are shown in Table III. The recognition rates (%) of 1NN versus the dimensions are shown in Fig. 4(a). As seen from Table III and Fig. 4(a), the top recognition rate of S2DP is the highest. The experiment also supports our experimental analysis mentioned above and suggests that S2DP is more robust than 2DPCA, 2DLDA, 2DLPP, and 2DLGEDA on facial expressions, lighting conditions, and time variations.

C. Experiments on FERET Face Database

The FERET face database is a result of the FERET program, which was sponsored by the US Department of Defense through the Defense Advanced Research Projects Agency (DARPA) Program [46]. It has become a standard database for testing and evaluating state-of-the-art face recognition algorithms. The proposed method was tested on a subset of the FERET database. This subset includes 1400 images of 200 individuals (each individual has seven images) and involves variations in facial expression, illumination, and pose. In the experiment, the facial portion of each original image was automatically cropped based on the location of the eyes, and the cropped images was resized to 40 × 40 pixels. The sample images of one person are shown in Fig. 5.

TABLE IV
RECOGNITION RATES (PERCENTAGE) AND CORRESPONDING DIMENSION OF THE FIVE METHODS ON THE FERET FACE DATABASE

Classifiers	2DPCA	2DLDA	2DLPP	2DLGEDA	S2DP
1NN	47.00	43.00	50.00	51.50	57.00
	40×37	40×16	40×11	40×12	40×19
3NN	40.50	42.00	49. 50	50.00	55.00
	40×30	40×39	40×39	40×39	40×7
MD	33.50	34.00	51.00	42.50	54.50
	40×36	40×39	40×39	40×30	40×6
SVM	45.00	43.50	57.00	53.00	59.50
	40×37	40×9	40×11	40×6	40×6

In the experiments, in order to test the performance of the proposed method in the variations of facial expression and lighting condition, five images were selected from the image gallery of each individual to form the training sample set, and one image per individual was used for validation. The remaining images were used for the test. The method in selecting the optimal parameters was the same as in Section IV-A. The optimal parameters of S2DP were set as $k = 6$, $\alpha = 0.01$, $\beta = 0.01$, and $K = 10$. The maximal recognition rates of different methods by using the four classifiers and the corresponding dimensions are given in Table IV. The recognition rate curves using 1NN classifier versus the variation of dimensions are shown in Fig. 4(b). Table IV and Fig. 4(b) show that S2DP obtains the highest recognition rate. With the SVM as the classifier, S2DP can obtain higher classification accuracy than the other classifiers.

D. Experiments on CMU PIE Face Database

The CMU PIE database [47] contains 68 people, and each person has 13 pose variations that ranged from right-to-left profile images and 43 different lighting conditions, which have 21 flashes with ambient light on or off. Twenty-three frontal-view images of each person were used in our experiments. Original images were aligned, cropped, and then resized to 46 × 46 pixels. Fig. 6 shows some sample images in the CMU PIE face database. This database was used for evaluating the performance of different methods when the face poses and lighting conditions varied in large ranges.

In the experiments, 10 and 2 images were randomly selected from each individual for training and validation, respectively, while the remaining images of each individual were used for the actual test. For each run, the parameters in each algorithm were selected using the same method as in Section IV-A ($k = 10$, $\alpha = 0.01$, $\beta = 0.01$, and $K = 30$ were the optimal parameters of S2DP on average). The experiments were repeated 20 times for avoiding the bias of the random experiments. The average recognition rates and the corresponding dimensions with different classifiers are reported

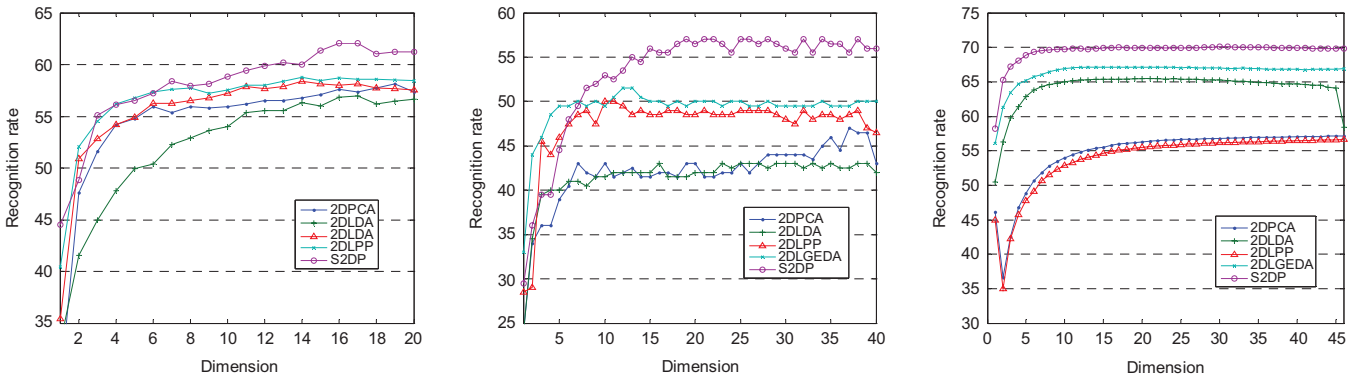

Fig. 4. Recognition rates (%) versus the dimensions on the (a) AR, (b) FERET, and (c) CMU PIE face databases, respectively. Classifier: 1NN.

Fig. 5. Sample images of one person on FERET face database.

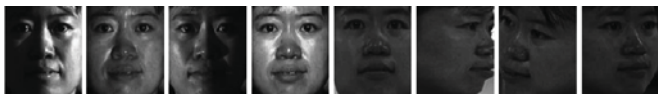

Fig. 6. Some sample images of one person in the CMU PIE face database.

TABLE V
AVERAGE RECOGNITION RATES (PERCENTAGE), STANDARD DEVIATION, AND THE CORRESPONDING DIMENSIONS OF THE FIVE METHODS ON THE CMU PIE FACE DATABASE

Classifiers	2DPCA	2DLDA	2DLPP	2DLGEDA	S2DP
1NN	57.11 ±6.61 46×46	65.46 ±4.04 46×20	56.55 ±5.87 46×46	67.13 ±5.46 46×20	70.12 ±4.53 46×30
3NN	56.03 ±6.88 46×42	64.72 ±4.21 46×21	56.38 ±5.83 46×41	67.00 ±5.69 46×40	68.24 ±4.57 46×35
MD	25.58 ±7.63 46×42	62.72 ±5.74 46×40	45.77 ±6.99 46×43	63.18 ±6.01 46×41	67.42 ±4.66 46×41
SVM	56.27 ±13.22 46×14	65.38 ±8.48 46×35	60.41 ±13.54 46×28	66.32 ±7.81 46×28	69.55 ±4.68 46×28

in Table V. The recognition rates using 1NN classifier versus the dimensions of each method are plotted in Fig. 4(c).

As seen from Fig. 6, the face poses and lighting conditions varied in large ranges. In this case, the unsupervised methods (i.e., 2DPCA and 2DLPP) obtained bad results. By using the label information, 2DLDA, 2DLGEDA, and S2DP obtained better performance. By introducing the L_1-norm elastic net regression, S2DP can select the most important discriminative factors to form the sparse projections, thus the recognition rates of S2DP were significantly higher than the recognition rates of other methods. The only difference between the S2DP and the other 2-D-based methods is that S2DP introduces L_1-norm for discriminative projection learning. This indicates that using the sparse projections or introducing the L_1-norm for discriminative projection learning can greatly improve the robustness for the variations of facial expressions and lighting conditions. The results are consistent with the experiments mentioned above. Due to the large variations in lighting conditions and poses, different classifiers perform very differently from other classifiers. However, the 1NN and SVM still obtain higher accuracies, and S2DP can also achieve the best performance. These series of experiments presented above show that S2DP is more robust when different classifiers were used for classification.

V. CONCLUSION

In this paper, an image-matrix-based sparse projections framework called S2DP was proposed for face feature extraction and recognition. S2DP combines the L_1-norm elastic net regression and SVD to iteratively learn the sparse projections instead of solving the generalized eigenequation. Thus, we generalized the vector-based sparse projections learning to the image-matrix-based cases. Our theoretical analysis showed that the sparse projections learned by our method approximate the eigensubspace of the corresponding image-based generalized eigenequation. According to the connections between the vector-based and image-matrix-based methods, our theorems also provide theoretical guarantees for the effectiveness of vector-based sparse learning methods. The results of the theoretical analysis also showed that S2DP is more efficient and costs less memory space than the vector-based sparse projection methods. Experiments on various face databases indicate that the proposed framework outperforms other non-sparse 2-D projection methods with different classifiers.

REFERENCES

[1] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
[2] I. Joliffe, *Principal Component Analysis*. London, U.K.: Springer-Verlag, 1986.
[3] K. Fukunnaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1991.
[4] A. M. Martinez and A. C. Kak, "Principal components analysis versus linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriengman, "Eigenfaces versus fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[6] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional principal components analysis: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[7] Y. Xu, D. Zhang, J. Yang, and J. Yang, "An approach for directly extracting features from matrix data and its application in face recognition," *Neurocomputing*, vol. 71, nos. 10–12, pp. 1857–1865, 2008.

[8] W. Zuo, D. Zhang, and K. Wang, "Bidirectional principal components analysis with assembled matrix distance metric for image recognition," *IEEE Trans. Syst., Man, Cybern., B*, vol. 36, no. 4, pp. 863–872, Aug. 2006.

[9] J. Ye, "Generalized low rank approximations of matrices," *Mach. Learn.*, vol. 61, no. 1, pp. 167–191, 2005.

[10] M. Li and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognit. Lett.*, vol. 26, no. 5, pp. 527–532, 2005.

[11] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. 18th Annu. Conf. Neural Inf. Process. Syst.*, 2004, pp. 1569–1576.

[12] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang, "Discriminant analysis with tensor representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 526–532.

[13] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.

[14] S.-J. Wang, J. Yang, M.-F. Sun, X.-J. Peng, M.-M. Sun, and C.-G. Zhou, "Sparse tensor discriminant color space for face verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 876–888, Jun. 2012.

[15] J. B. Tenenbaum, V. DeSilva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, Dec. 2000.

[16] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, Dec. 2000.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, 2001, pp. 589–591.

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

[19] H. Huang and H. He, "Super-resolution method for face recognition using nonlinear mappings on coherent features," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 121–130, Jan. 2011.

[20] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.

[21] Y. Xu, F. Song, F. Ge, and Y. Zhao, "A novel local preserving projection scheme for use with face recognition," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6718–6721, 2011.

[22] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press, 2003.

[23] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacian faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[24] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2005.

[25] B. Niu, Q. Yang, S. C. K. Shiu, and S. K. Pal, "Two-dimensional laplacianfaces method for face recognition," *Pattern Recognit.*, vol. 41, no. 10, pp. 3237–3243, 2008.

[26] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections with its application to palmprint recognition," *Pattern Recognit.*, vol. 40, no. 1, pp. 339–342, 2007.

[27] S. Chen, H. Zhao, M. Kong, and B. Luo, "A two-dimensional extension of locality preserving projections," *Neurocomputing*, vol. 70, nos. 4–6, pp. 912–921, 2007.

[28] R. Zhi and Q. Ruan, "Facial expression recognition base on two-dimensional discriminant locality preserving projections," *Eurocomputing*, vol. 70, no. 7, pp. 1543–1546, 2007.

[29] M. Wan, Z. Lai, J. Shao, and Z. Jin, "Two-dimensional local graph embedding discriminant analysis with its application to face and palm biometrics," *Eurocomputing*, vol. 73, nos. 1–3, pp. 197–203, 2009.

[30] Y. Xu, G. Feng, and Y. Zhao, "One improvement to two-dimensional locality preserving projection method for use with face recognition," *Neurocomputing*, vol. 73, nos. 1–3, pp. 245–249, 2009.

[31] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

[32] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annal. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

[33] H. Zhou and T. Hastie, "Regression shrinkage and selection via the elastic net with applications to microarrays," Dept. Stat., Stanford Univ., Stanford, CA, Tech. Rep., 2003.

[34] A. d'Aspremont, L. E. Chaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," in *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press, 2004.

[35] B. Moghaddam, Y. Weiss, and S. Avidan, "Spectral bounds for sparse PCA: Exact and greedy algorithms," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2005.

[36] B. Moghaddam, Y. Weiss, and S. Avidan, "Generalized spectral bounds for sparse LDA," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 641–648.

[37] L. Clemmensen, T. Hastie, and B. Ersboll, "Sparse discriminant analysis," U.S. Dept. Stat., Stanford Univ., Stanford, CA, Tech. Rep. 1-25, Jun. 2008.

[38] Z. Qiao, L. Zhou, and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data," *Int. Assoc. Eng., J. Appl. Math.*, vol. 39, no. 1, pp. 48–60, 2010.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[40] Z. Zheng, "Sparse local preserving embedding," in *Proc. 2nd Int. Congr. Image Signal Process.*, 2009, pp. 1–5.

[41] T. Zhou, D. Tao, and X. Wu, "Manifold elastic net: A unified framework for sparse dimension reduction," *Data Mining Knowl. Discovery*, vol. 22, no. 3, pp. 340–371, 2011.

[42] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[43] L. Elden, "Algorithms for the regularization of ill-conditioned least squares problems," *BIT*, vol. 17, no. 2, pp. 134–145, 1997.

[44] L. Wang, X. Wang, and J. Feng, "On image matrix based feature extraction algorithms," *IEEE Trans. Syst. Man Cybern. B*, vol. 36, no. 1, pp. 194–197, Feb. 2006.

[45] A. M. Martinez and R. Benavente, "The AR face database," Centre de Visio per Computador, Univ. Autonoma de Barcelona, Bellaterra, Barcelona, Tech. Rep. 24, Jun. 1998.

[46] P. J. Phillips. (2004). *The Facial Recognition Technology (FERET) Database* [Online]. Available: http://www.itl.nist.gov/iad/humanid/feret/feret_master.html

[47] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

Zhihui Lai received the B.S. degree in mathematics from South China Normal University, Guangzhou, China, the M.S. degree from Jinan University, Guangzhou, and the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology, Nanjing, China, in 2002, 2007, and 2011, respectively.

He has been a Research Associate with The Hong Kong Polytechnic University, Hong Kong, since 2010. Currently, he is also a Postdoctoral Fellow with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive senses, human vision modeling, and applications in intelligent robot research.

Wai Keung Wong received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong.

He is currently an Associate Professor with The Hong Kong Polytechnic University. He has authored or co-authored more than 50 scientific articles in refereed journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, the *International Journal of Production Economics*, the *European Journal of Operational Research*, the *International Journal of Production Research, Computers in Industry*, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, among others. His recent research interests include artificial intelligence, pattern recognition, and optimization of manufacturing scheduling, planning and control.

Zhong Jin received the B.S. degree in mathematics, the M.S. degree in applied mathematics, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), China, in 1982, 1984, and 1999, respectively.

He is currently a Professor in the Department of Computer Science, NUST. He was a Research Assistant with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, from 2000 to 2001. He was a Visitor with the Laboratoire HEUDIASYC, Universite de Technologie de Compiegne, France, from 2001 to 2002. He was with the Centre de Visio per Computador, Universitat Autonoma de Barcelona, Spain, as a Ramon y Cajal Program Research Fellow from September 2005 to October 2005. His current interests include pattern recognition, computer vision, face recognition, facial expression analysis, and content-based image retrieval.

Jian Yang received the B.S. degree in mathematics from the Xuzhou Normal University, Xuzhou, China, the M.S. degree in applied mathematics from the Changsha Railway University, Changsha, China, and the Ph.D. degree on pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 1995, 1998, and 2002, respectively.

He was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain, in 2003. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark. He is currently a Professor with the School of Computer Science and Technology, NUST. He has authored or co-authored more than 50 scientific papers on pattern recognition and computer vision. His current research interests include pattern recognition, computer vision, and machine learning.

Dr. Yang was a recipient of the RyC Program Research Fellowship sponsored by the Spanish Ministry of Science and Technology in 2003.

Yong Xu (M'06) received the B.S. and M.S. degrees from the Air Force Institute of Meteorology, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST) , Nanjing, China, in 2005.

He was a Post-Doctoral Research Fellow with the Shenzhen Graduate School, Harbin Institute of Technology, from 2005 to 2007, where he is currently a Professor. He is also a Research Assistant Researcher with The Hong Kong Polytechnic University, Hong Kong, from 2007 to 2008. He has authored or co-authored more than 40 scientific papers. His current interests include pattern recognition, biometrics, and machine learning.