# Approximate Orthogonal Sparse Embedding for Dimensionality Reduction

Zhihui Lai, Wai Keung Wong, Yong Xu, *Member, IEEE*, Jian Yang, and David Zhang, *Fellow, IEEE*

***Abstract*— Locally linear embedding (LLE) is one of the most well-known manifold learning methods. As the representative linear extension of LLE, orthogonal neighborhood preserving projection (ONPP) has attracted widespread attention in the field of dimensionality reduction. In this paper, a unified sparse learning framework is proposed by introducing the sparsity or $L_1$-norm learning, which further extends the LLE-based methods to sparse cases. Theoretical connections between the ONPP and the proposed sparse linear embedding are discovered. The optimal sparse embeddings derived from the proposed framework can be computed by iterating the modified elastic net and singular value decomposition. We also show that the proposed model can be viewed as a general model for sparse linear and nonlinear (kernel) subspace learning. Based on this general model, sparse kernel embedding is also proposed for nonlinear sparse feature extraction. Extensive experiments on five databases demonstrate that the proposed sparse learning framework performs better than the existing subspace learning algorithm, particularly in the cases of small sample sizes.**

***Index Terms*— Dimensionality reduction, elastic net, image recognition, manifold learning, sparse projections.**

## I. Introduction

DIMENSIONALITY reduction or feature extraction is widely used in data mining, computer vision, and pattern recognition. The classical dimensionality reduction methods, such as principle component analysis (PCA) [1]–[3] and linear discriminant analysis (LDA) [4]–[6], and their modified methods [7]–[9] are simple, effective, and widely used in different fields, including face recognition, palmprint recognition, and so on. However, these classical methods (i.e., PCA and LDA) only focus on the global structure of a data set for dimensionality reduction.

Roweis and Saul [10] and Tenenbaum *et al.* [11] indicate that images of different objects lie on a low-dimensional manifold embedded in a high-dimensional space. To discover the intrinsic geometry structure of a data set, manifold learning methods have been widely used in the past decade, and the well-known ones, such as locally linear embedding (LLE) [10], ISOMAP [11], and Laplacian eigenmap [12], were proposed. Inspired by these nonlinear methods, a lot of linear dimensionality reduction methods based on manifold learning were proposed for feature extraction. Among these methods, neighborhood preserving embedding (NPE) [13], orthogonal neighborhood preserving projection (ONPP) [14], and locality preserving projections [15], [16] are the representative ones, which preserve the local geometric structure of the manifold using simple linear approximation to the nonlinear mappings. Due to their simplicity and effectiveness, the LLE-based methods were extended to different forms and widely used in facial expression recognition [17], image prediction and retrieval [18], [19], feature fusion [20], face recognition [13], [14], [21], gait recognition [22], and human motion recognition [23].

Recent research shows that the $L_1$-norm sparse learning can enhance the robustness for classification or feature extraction. For example, the sparse representation classifier was proposed for robust face recognition [24]–[26]. The sparse graph or $L_1$ graph was also used in spectral clustering [27] and label propagation [28]. In addition, an important application of the sparse graph or $L_1$ graph is to characterize the robust spare reconstruction relationship among the data points for feature extraction. The representative methods are sparsity preserving projections (SPP) [29] and its supervised extension [30], which aim to learn the linear subspace for dimensionality reduction.

However, the classical (i.e., PCA and LDA), manifold learning-based (i.e., ONPP and NPE) and $L_1$ graph-based (i.e., SPP and its supervised extension [30]), and sparse color component-based [31] linear dimensionality reduction methods can only learn compact projections (i.e., elements in the projections are usually nonzero). Thus, such projections lack reasonable interpretation. On the other hand, the

Manuscript received December 6, 2013; revised April 6, 2015; accepted April 7, 2015. Date of publication May 6, 2015; date of current version March 15, 2016. This work was supported in part by The Hong Kong Polytechnic University, Hong Kong, under Project G-YM53, in part by the National Natural Science Foundation of China under Grant 61203376, Grant 61375012, Grant 61071179, and Grant 61125305, in part by the Hi-Tech Research and Development Program of China under Grant 2006AA01Z119, in part by the Guangdong Natural Science Foundation under Project S2012040007289, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20130329152024199.

Z. Lai is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China, and also with the Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong (e-mail: lai_zhi_hui@163.com).

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

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

J. Yang is with the School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@njust.edu.cn).

D. Zhang is with the Biometrics Research Centre, Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

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.2015.2422994

previous studies show that the introduction of the $L_1$ norm for sparse regression not only enhances the prediction accuracy but also strengthens the generalization ability and the robustness for prediction [32]–[34]. Therefore, the sparse subspace learning methods have attracted much attention. The common properties of the so-called sparse subspace/projection learning methods are that the projection vectors contain a lot of zero elements and can provide the psychological and physiological interpretation. To this end, the representative sparse subspace learning method sparse PCA (SPCA) [35] and double shrinking algorithm (DSA) [36] were proposed, in which the $L_1$-norm penalty was imposed on the regularized optimization problem. Using the label information, Clemmensen *et al.* [37] proposed the sparse discriminant analysis (SDA), which extends the LDA to sparse cases. Cai *et al.* [38] proposed the unified sparse subspace learning (USSL) framework via sparse regression on the spectral of a special neighborhood graph. Using the elastic net [34] for regression, the most important/contributive variables are selected to form the projective vectors in SPCA, SDA, and USSL.

The sparse extensions of the classical methods (i.e., SPCA and SDA) only focus on the global geometric structure in dimensionality reduction. Although USSL takes the local geometric structure into account, it has the following two disadvantages. First, the projections of USSL are independently computed and cannot guarantee complete or approximate orthogonality and thus the effectiveness for feature extraction may be degraded [14], [17], [21]. Second, since the two-step approximation method used in USSL can cause bigger approximation error, the locality preserving ability of USSL may also be affected. Since the data (e.g., images) lie on a low-dimensional manifold embedded in a high-dimensional space, it is very important to explore the local geometric structure in dimensionality reduction. Therefore, it is necessary to develop a new framework preserving the manifold structure and the orthogonality so as to explore the important factors in a sparse manner in feature extraction. It is desirable to further improve the recognition performance of the sparse subspace learning algorithm and strengthen the generalization ability and the robustness.

In this paper, the ONPP is taken as an example to design a novel sparse subspace learning framework, including linear and nonlinear forms, to meet the practical needs in sparse feature extraction. The main contributions of this paper are as follows.

1) We propose a general sparse subspace learning framework called sparse linear embedding (SLE) that can directly integrate the local geometric structure to obtain the sparse projections. We show that the optimal sparse subspace can be computed by iterating elastic net regression and singular value decomposition (SVD).
2) The theoretical relationships between the proposed SLE and ONPP are revealed. In addition, the intrinsic connections between SLE and some sparse subspace learning methods are also discussed.
3) Using the same framework as the platform, Kernel ONPP (KONPP)[14] can also be extended to sparse cases, which shows that the framework is not only suitable for sparse linear subspace learning but also suitable for sparse nonlinear subspace learning.
4) Extensive experimental results show that SLE and sparse kernel embedding (SKE) perform better than the classical sparse subspace learning methods, the sparse, and nonsparse manifold learning-based algorithms in feature extraction and classification.

The rest of this paper is organized as follows. In Section II, LLE and its linear and kernel extensions are reviewed. In Section III, regression analysis is presented as the preparations for SLE. In addition, SLE is proposed in Section IV. Section V simply presents the SKE algorithm. Experiments are carried out to evaluate the SLE and SKE algorithm in Section VI. Finally, the conclusions are given in Section VII.

## II. Related Works

In this section, LLE, NPE, and ONPP are reviewed. Some theoretical analyses are also given for the sake of presenting the proposed sparse embedding framework.

Let the matrix $X = [x_1, x_2, \ldots, x_N]$ be the data matrix, including all the training samples $\{x_i\}_{i=1}^N \in R^m$ in its columns. In practice, the feature dimension $m$ is often very high. The goal of (sparse) linear dimensionality reduction is to transform the data from the originally high-dimensional space to a low-dimensional one

$$y = A^T x \in R^d \tag{1}$$

for any $x \in R^m$ with $d \ll m$, where $A = (a_1, a_2, \ldots, a_d)$ and $a_i (i = 1, \ldots, d)$ is an $m$-dimensional column vector. For the sparse projection learning methods, $a_i (i = 1, \ldots, d)$ shall be sparse (i.e., only a few elements in $a_i$ are nonzero elements/loadings).

### A. Locally Linear Embedding

LLE aims to preserve the local linear reconstruction relationship among the data points. In the first step of LLE, each sample $x_i$ is approximated by a weighted linear combination of its $k$ nearest neighbors on the assumption that the neighboring samples lie on a locally linear patch of the nonlinear manifold. The following cost function should be minimized:

$$\varepsilon(W) = \sum_i \Big\| x_i - \sum_{j \in N_k(x_i)} W_{ij} x_j \Big\|^2 \quad \text{s.t.} \sum_{j \in N_k(x_i)} W_{ij} = 1 \tag{2}$$

where $N_k(x_i)$ denotes the index set of $k$ nearest neighbors of $x_i$ and $W_{ij}$ is the optimal local least square reconstruction coefficients.

Once the optimal reconstructive matrix $W$ is obtained, in the second step, it is kept fixed and the final embedding coordinates can be computed by minimizing the following optimization problem:

$$\varepsilon(Y) = \sum_i \Big\| y_i - \sum_j W_{ij} y_j \Big\|^2 = \operatorname{tr}(Y(I - W)^T (I - W) Y) \quad \text{s.t. } YY^T = I \tag{3}$$

where $Y = [y_1, y_2, \ldots, y_d]$. The eigenvectors corresponding to the smaller eigenvalues of the eigenfunction are the final embedding of the data set

$$(I - W)^T (I - W) y = \lambda y. \tag{4}$$

where $y$ is the eigenvector corresponding to eigenvalue $\lambda$. Since there exists the out-of-samples problem for LLE, the linearization method (i.e., ONPP) is proposed to solve this problem.

### B. Orthogonal Neighborhood Preserving Projections

ONPP aims at preserving the local neighborhood geometry structure of the data. The affinity weight matrix of ONPP is from the coefficients of the local least squares approximation as in LLE. To obtain the orthogonal linear projection that can preserve the local linear reconstructive relationship, ONPP uses the strategy of linear approximation to the nonlinear mapping of LLE to learn the projection. The criterion for choosing the optimal projection $P$ is to minimize the cost function

$$\min_a \sum_i \left\| a^T x_i - \sum_j W_{ij} a^T x_j \right\|^2 \quad \text{s.t. } a^T a = 1. \tag{5}$$

The optimal projections of (5) are the orthogonal eigenvectors corresponding to the minimum eigenvalue of the following standard eigenvalue problem:

$$X (I - W)^T (I - W) X^T a = \lambda a \tag{6}$$

or in the matrix form as

$$X (I - W)^T (I - W) X^T A = A \Lambda \tag{7}$$

where $\Lambda$ is the eigenvalue matrix whose diagonal elements are the eigenvalues of $X(I - W)^T (I - W) X^T$.

It should be noted that the orthogonal eigenvectors of ONPP can also be obtained using SVD. Let the SVD of $X(I - W)^T = U D V^T$, where $D$ contains the nonzero singular value in ascending order. Since

$$X (I - W)^T (I - W) X^T = U D V^T V D U^T = U D^2 U^T. \tag{8}$$

It can be seen that the column vectors in $U$ corresponding to the first $d$ smaller nonzero singular values are also the solutions/projections of ONPP.

## III. Regression for Linear Embedding

In this section, we focus on the SLE method (i.e., the sparse extension of ONPP). The idea is to represent the objective function of ONPP in the regression form and provide the theoretical guarantee that the regression solutions/subspace are exactly the solution space of ONPP, which are the eigenvectors of the eigenequation of (6). Thus, the $L_1$-norm penalty term can be added to the regression minimization problem for computing the sparse vectors.

### A. Representation of ONPP

Suppose the weight matrix $W$ in ONPP is given and thus we can define a special matrix associated with the weight matrix as

$$M = (I - W). \tag{9}$$

To preserve the locally linear reconstruction relationship of the data set reflected by $W$ (or $M$), the weight matrix $W$ (or $M$) must be integrated into the new optimization problem, and the solutions' equivalence between the ONPP and the new optimization problem should be guaranteed in theory. In other words, the objective function of ONPP should be rewritten in other forms but with the same optimal solution space as the original ONPP.

Let us take the following optimization problem into consideration:

$$\min_b \sum_i \| b b^T X M_{i,:}^T - X M_{i,:}^T \|_F^2 \tag{10}$$

$$\text{s.t.} \quad b^T b = 1 \tag{11}$$

where $M_{i,:}$ denotes the $i$th row vector in matrix $M$ and $b \in R^m$ is the projection. The original idea in (10) is that any locally linear reconstruction of the data $X M_{i,:}^T$ should be close to the transformed data $b b^T X M_{i,:}^T$. Thus, minimizing the sum of the errors will lead to preserve the locally linear reconstructive relationship as the one in ONPP or LLE. For the optimization problem (10) and (11), we have the following theorem.

*Theorem 1:* The optimization problem in (10) and (11) can derive the same solutions to the eigenequation (6).

*Proof:* From (10) and (11), we have

$$\begin{aligned}
&\sum_i \| b b^T X M_{i,:}^T - X M_{i,:}^T \|_F^2 \\
&= \sum_i \operatorname{tr}\big[ b b^T X M_{i,:}^T \big( b b^T X M_{i,:}^T \big)^T - 2 b b^T X M_{i,:}^T \big( X M_{i,:}^T \big)^T \\
&\qquad + X M_{i,:}^T \big( X M_{i,:}^T \big)^T \big] \\
&= \operatorname{tr}( X M^T M X^T - b b^T X M^T M X^T ).
\end{aligned}$$

The second = satisfies since $b^T b = 1$ is used. Therefore, since $\operatorname{tr}(X M^T M X^T)$ is a constant, the above minimization problem becomes the following maximization problem:

$$\max_b \ \operatorname{tr}(b b^T X M^T M X^T) = \max_b \ \operatorname{tr}(b^T X M^T M X^T b)$$
$$\text{s.t.} \quad b^T b = 1. \tag{12}$$

Using the Lagrange multiplier method, the eigenequation of (6) can be derived from (12). Therefore, any eigenvector corresponding to eigenvalue $\lambda$ in (6) is also the optimal solution to (12) when the Lagrange multiplier is set as $\lambda$. ■

Theorem 1 indicates that the eigenvectors of (6) or (7) can be derived from other form as (10) and (11). In Section III-B, it could be seen that in essence this provides a tractable way to learn the same subspace/solution space as in (6) or (7).

### *B. Projection Relaxation for ONPP*

Theorem 1 provides the equivalent representation of the eigenequation of (6). However, this representation cannot generate the sparse solutions. Our idea is to relax the projection by taking $b^T$ in the term $bb^T XM_{i,:}$ of (10) as a new variable $p^T$ so that (10) is rewritten in a regression form and thus the sparse regression algorithms can be used to compute the sparse solutions. To this end, we first discuss a new optimization problem

$$\begin{aligned}&\min_{b,p} \sum_i \|bp^T XM_{i,:}^T - XM_{i,:}^T\|_F^2\\ &\text{s.t.} \quad b^T b = 1.\end{aligned} \tag{13}$$

*Theorem 2:* Suppose $X(I-W)^T(I-W)X^T$ is the full-rank matrix. Let $p^*$ be the optimal solution to (13), then $p^* = \delta U(:, \text{end})$, where $\delta = 1$ or $-1$ and $U(:, \text{end})$ denote the last left singular vector as in (8).

*Proof:* The proof is a special case of Theorem 3. Thus, it is omitted for avoiding repetition. ■

Theorem 2 shows that the optimal solution (i.e., the subspace spanned by $p^*$) is the same as the one spanned by $U(:, \text{end})$. The only difference is the direction (i.e., positive or negative), which does not affect the theoretical analysis, since we can simply select to set $\delta = 1$ resulting in $p^* = U(:, \text{end})$.

Theorem 2 guarantees in theory that the optimization problem (13) can derive the same solutions to (6) when $X(I - W)^T(I - W)X^T$ is the full-rank matrix. However, $X(I-W)^T(I-W)X^T$ may be not the full-rank matrix, since small sample size problems frequently exist in practice. Thus, it is necessary to develop new models and theories to deal with this case, which will be shown in the next section.

### *C. Modified Ridge Regression Representation of ONPP*

There are two reasons for discussing the ridge regression representation of ONPP. The first is to deal with the singular problem [i.e., $X(I - W)^T(I - W)X^T$ is not the full-rank matrix]. The second is that we plan to use the elastic net algorithm to obtain the sparse solutions, which needs to impose the $L_1$ and $L_2$ norms on the regression-type optimization problem simultaneously. Therefore, let us first consider the modified ridge regression problem as follows:

$$\begin{aligned}&\min_{b,p} \sum_i \|bp^T XM_{i,:}^T - XM_{i,:}^T\|_F^2 + \beta\|p\|^2\\ &\text{s.t.} \quad b^T b = 1.\end{aligned} \tag{14}$$

For the above optimization problem, we have the following theorem.

*Theorem 3:* Let $p^*$ be the optimal solution of (14), then $p^* = \delta(D_1^2/D_1^2 + \beta)U(:, \text{end})$, where $\delta = 1$ or $-1$ and $D_1$ denote the largest singular value of $X(I - W)^T$. If $\delta = 1$ and let $\widehat{p} = p^*/\|p^*\|$, then $\widehat{p} = U(:, \text{end})$.

*Proof:* The proof is in the Appendix. ■

Thus, the theoretical analysis shows that (14) also derives the same solution as the original ONPP in (6). However, the above analysis only reveals the relationship between the first projection of ONPP and the regression optimization problem. Usually, a single projection is not enough for feature extraction or dimensionality reduction. Thus, a set of projections is needed. We should consider a more complex case to compute a set of projections which are expected to be the same as the projections of ONPP so as to optimally preserve the local geometry relationship.

The next theorem extends Theorem 3 to derive the whole sequence of projections and reveals the relationship between the set of projections of ONPP and the new optimization problem. Let us take the optimization problem in the matrix form as follows:

$$\begin{aligned}&\min_{B,P} \sum_i \|BP^T XM_{i,:}^T - XM_{i,:}^T\|_F^2 + \beta\|P\|_F^2\\ &\text{s.t.} \quad B^T B = I_d\end{aligned} \tag{15}$$

where both $B = (b_1, b_2, \ldots, b_d)$ and $P = (p_1, p_2, \ldots, p_d)$ are of the size $m \times d$ matrices. This optimization problem can provide the sequential solutions of ONPP [i.e., the solutions of (7) instead of a single solution/projection]. This can be guaranteed by the following theorem.

*Theorem 4:* Let $P^*$ be the optimal solution to (15), then $P^* = [p_1^*, p_2^*, \ldots, p_d^*] = U(D^2/D^2 + \beta I)\Delta$, where diagonal matrix $\Delta$ only contains 1 or $-1$ in its diagonal elements. If the diagonal elements in $\Delta$ are all set to 1, and let $\widehat{p}_i = p_i^*/\|p_i^*\|$, then $\widehat{P} = [\widehat{p}_1, \ldots, \widehat{p}_1] = U(:, 1:d)$.

*Proof:* The proof is similar to the one in Theorem 3 with $b$ and $p$ replaced by matrix $B$ and $P$, respectively. Thus, it is omitted. ■

Theorem 4 shows whether the matrix $X(I-W)^T(I-W)X^T$ is singular or not, the modified ridge regression representation in optimization (15) can always derive the sequential solutions of ONPP in (7) or (8).

Theorem 4 also shows that if the normalized steps are operated on $P^*$, the model can derive exact solutions of ONPP. However, the optimal solutions to (15) are usually nonsparse, and therefore the $L_1$-norms penalty is also imposed on the model, which is presented in the next section.

## IV. Sparse Linear Embedding

In this section, the model of the proposed SLE will be introduced and the optimization method is also presented.

### *A. Model and Its Solutions of Sparse Linear Embedding*

Since the modified regression optimization problem (14) or (15) cannot give the sparse projections, a tractable method is to add the $L_1$ norm to the optimization problem in (14) or (15). Thus, we have the following SLE model combining $L_1$ and $L_2$ norms for regression:

$$\begin{aligned}&\min_{b,p} \sum_i \|bp^T XM_{i,:}^T - XM_{i,:}^T\|_F^2 + \beta\|p\|^2 + \gamma|p|\\ &\text{s.t.} \quad b^T b = 1\end{aligned} \tag{16}$$

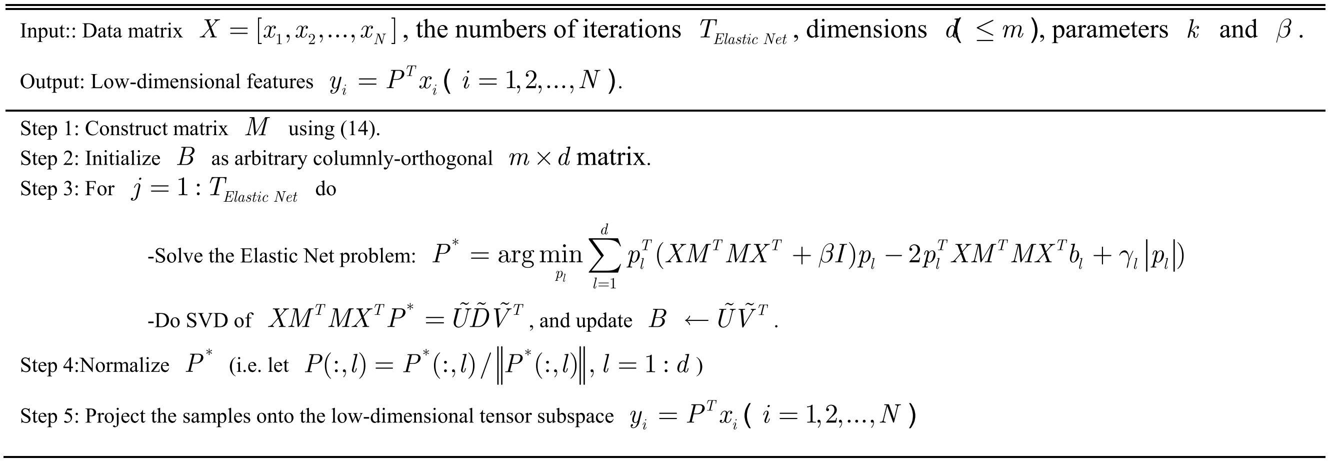
or in the matrix form

$$\begin{aligned}&\min_{B,P} \sum_i \|BP^T XM_{i,:}^T - XM_{i,:}^T\|_F^2 + \beta\|P\|_F^2 + \sum_{l=1}^{d} \gamma_l |p_l|\\ &\text{s.t.} \quad B^T B = I_d\end{aligned} \tag{17}$$

where $\beta > 0$, $\gamma_l > 0$, and $|\cdot|$ denote the $L_1$ norm.

TABLE I

SPARSE LINEAR EMBEDDING ALGORITHM



```
Input:: Data matrix $X=[x_1,x_2,...,x_N]$, the numbers of iterations $T_{Elastic\ Net}$, dimensions $d(\le m)$, parameters $k$ and $\beta$.
Output: Low-dimensional features $y_i = P^T x_i\ (i=1,2,...,N)$.
Step 1: Construct matrix $M$ using (14).
Step 2: Initialize $B$ as arbitrary columnly-orthogonal $m \times d$ matrix.
Step 3: For $j = 1 : T_{Elastic\ Net}$ do
        -Solve the Elastic Net problem: $P^* = \arg\min_{p_l} \sum_{l=1}^{d} p_l^T (XM^TMX^T + \beta I)p_l - 2p_l^T XM^TMX^T b_l + \gamma_l |p_l|)$
        -Do SVD of $XM^TMX^TP^* = \tilde{U}\tilde{D}\tilde{V}^T$, and update $B \leftarrow \tilde{U}\tilde{V}^T$.
Step 4:Normalize $P^*$ (i.e. let $P(:,l) = P^*(:,l)/\|P^*(:,l)\|, l = 1:d$)
Step 5: Project the samples onto the low-dimensional tensor subspace $y_i = P^T x_i\ (i = 1,2,...,N)$
```

It can be seen from (16) and (17) that it is difficult to directly solve the optimization problem with $L_1$- and $L_2$-norm penalty. Therefore, we propose to use the alternative iteration method to solve this problem. Since a single solution is insufficient for feature extraction, the optimization problem in the matrix form (17) is considered in Section IV-B.

### *B. Solutions of SLE*

Similar to the proof in Theorem 3, on the one hand, (17) can be rewritten as

$$\begin{aligned}\sum_i \|BP^TXM_{i,:}^T - XM_{i,:}^T\|_F^2 + \beta\|P\|_F^2 + \sum_{l=1}^{d}\gamma_l|p_l| \\ = \mathrm{tr}(XM^TMX^T) + \sum_{l=1}^{d} p_l^T(XM^TMX^T + \beta I)p_l \\ - 2p_l^TXM^TMX^Tb_l + \gamma_l|p_l|. \qquad (18)\end{aligned}$$

For given $B$ (i.e., $b_l$ is fixed), since $\mathrm{tr}(XM^TMX^T)$ is a constant and thus it can be ignored. Then, minimizing (18) is equivalent to solve the $d$-independent elastic net problems so as to get the optimal $p_l$ $(l = 1, 2, \ldots, d)$, which constitutes matrix $P$.

On the other hand, (18) can also be represented as

$$\begin{aligned}\sum_i \|BP^TXM_{i,:}^T - XM_{i,:}^T\|_F^2 + \beta\|P\|_F^2 + \sum_{l=1}^{d}\gamma_l|p_l| \\ = \mathrm{tr}[XM^TMX^T + P^T(XM^TMX^T + \beta I)P] + \sum_{l=1}^{d}\gamma_l|p_l| \\ - \mathrm{tr}(2P^TXM^TMX^TB). \qquad (19)\end{aligned}$$

For given matrix $P$, the term $\mathrm{tr}[XM^TMX^T + P^T(XM^TMX^T + \beta I)P] + \sum_{l=1}^{d}\gamma_l|p_l|$ becomes a constant, and thus can be ignored. Then, minimizing (19) for the given matrix $P$ becomes the following maximization problem:

$$\begin{aligned}\max_B\ & \mathrm{tr}(B^TXM^TMX^TP) \\ \text{s.t.}\ & B^TB = I_d. \qquad (20)\end{aligned}$$

The following theorem gives the optimal solutions to the optimization problem (20).

*Theorem 5:* Let the SVD of $XM^TMX^TP = \tilde{U}\tilde{D}\tilde{V}^T$, and then $\tilde{B} = \tilde{U}\tilde{V}^T$ is the optimal solution to (20).

*Proof:* The proof is similar to the one in [35, Th. 4]. ■

If $\gamma_l \to 0_+$, the optimal $\widehat{B} \to U(:1:d)$, which indicates that the optimal $B$ in the iteration procedures always lies on the same subspace as ONPP. Furthermore, according to Theorem 4, this property makes the (sparse) matrix $\widehat{P}$ also close to this orthogonal subspace (when $\gamma_l \to 0_+$), which indicates that the sparse projection matrix $\widehat{P}$ is approximately orthogonal. Thus, the effectiveness of using $\widehat{P}$ for feature extraction and discrimination is guaranteed (when ONPP is extended to sparse cases). These properties also indicate that the local geometry structure of the data set can be optimally preserved in a sparse manner.

### *C. Convergence of the Proposed Algorithm*

In this section, we discuss the convergence of the iterative algorithm.

*Theorem 6:* The iterative procedures of SLE presented in Table I will converge to a local optimum.

*Proof:* The original objective function of SLE in each iteration step can be rewritten as follows:

$$\begin{aligned}J(B^{(t-1)}, P^{(t-1)}) = \sum_i \|B^{(t-1)}P^{(t-1)T}XM_{i,:}^T - XM_{i,:}^T\|_F^2 \\ + \beta\|P^{(t-1)}\|_F^2 + \sum_{l=1}^{d}\gamma_l|p_l^{(t-1)}|\end{aligned}$$

where $B^{(t-1)}$ and $P^{(t-1)}$ are the optimal solutions of the objective function in the $t-1$th iteration, and $p_l^{(t-1)}$ is the column vector in $P^{(t-1)}$.

For the given $B^{(t-1)}$, the elastic net algorithm can give the optimal solution $P^{(t)}$, which decreases the objective function. Thus, we have $J(B^{(t-1)}, P^{(t)}) \le J(B^{(t-1)}, P^{(t-1)})$.

On the other hand, for the given $P^{(t)}$, the SVD of $XM^TMX^TP^{(t)}$ provides the optimal solution of $B^{(t)}$, which further decreases the objective function. Thus, we have $J(B^{(t)}, P^{(t)}) \leq J(B^{(t-1)}, P^{(t)})$.

From the above analysis, we have $J(B^{(t)}, P^{(t)}) \leq J(B^{(t-1)}, P^{(t-1)})$. Therefore, the objective function will converge to a local optimum in the iteration. ■

Since the object function is strict convex with respect to the variable $P$ (mainly because of the quadratic term of $P$), according to [39, Proposition 2.1.2], the optimal solution is unique. Hence, the convergence of the function values implies the convergence of the variable $P$ [40, Corollary 27.2.2]. On the other hand, the object function is convex with respect to the variable $B$ (the object function is a linear function of $B$). According to [40, Corollary 27.2.1], the sequence of $B^{(t)}$ is bounded and its every cluster point is a minimum point of the object function. Therefore, the convergence of the variable $P$ is guaranteed and the weak convergence of the variable $B$ can also be obtained.

### D. Discussion and Comparison

SPCA is obtained by extending the PCA to sparse cases. The following proposition shows the close relationship between SPCA and SLE.

*Proposition 1:* SPCA is a special case of the SLE model.

*Proof:* We only need to show that the optimization problem of SPCA is exactly the modified one of SLE when the local neighborhood matrix is defined in a special way as follows:

$$\tilde{M} = \sqrt{\frac{1}{N}\left(I - \frac{1}{N}ee^T\right)}$$

where $e$ is an $N$-dimensional vector whose elements are all 1s. Then, the optimization problem of SLE in (18) with the specially defined neighborhood matrix $\tilde{M}$ converses to be

$$\begin{aligned}
\min_{B,P} \; & \sum_i \|BP^T X\tilde{M}_{i,:} - X\tilde{M}_{i,:}\|_F^2 + \beta\|P\|_F^2 + \sum_{l=1}^{d} \gamma_l |p_l| \\
&= \min_{b_l,p_l} \; \mathrm{tr}(X\tilde{M}^T\tilde{M}X^T) + \sum_{l=1}^{d} p_l^T(X\tilde{M}^T\tilde{M}X^T + \beta I)p_l \\
&\quad - 2p_l^T X\tilde{M}^T\tilde{M}X^T b_l + \gamma_l|p_l| \\
&= \min_{b_l,p_l} \; \mathrm{tr}(S_T) + \sum_{l=1}^{d} p_l^T(S_T + \beta I)p_l - 2p_l^T S_T b_l + \gamma_l|p_l| \\
\text{s.t.} \quad & B^T B = I_d
\end{aligned}$$

where $S_T = X(1/N)(I - (1/N)ee^T)X^T$ is exactly the total scatter matrix in PCA. It can be seen from [35] that the above optimization problem is exactly the SPCA criterion. Therefore, SPCA is a special case of the SLE model. ■

DSA aims to minimize the following problem:

$$\min \; p_l^T S p_l + \gamma_l|p_l| \quad \text{s.t. } p_l^T p_l = 1.$$

It is easy to obtain the relationship of SLE and DSA.

*Proposition 2:* For the given matrix $S$, SLE degrades to be DSA when $B = P$ and $\beta = 0$.

*Proof:* Let $S = XM^TMX^T$. From (19), we have

$$\min_{B,P} \; \mathrm{tr}[P^T(S + \beta I)P - \mathrm{tr}(2P^TSB)] + \sum_{l=1}^{d} \gamma_l|p_l|$$
$$(\text{s.t. } B^TB = I_d).$$

If $B = P$ and $\beta = 0$, the above optimization problem derives

$$\min_{P} \; \mathrm{tr}[P^T(S + 0I)P - \mathrm{tr}(2P^TSP)] + \sum_{l=1}^{d} \gamma_l|p_l|$$
$$(\text{s.t. } P^TP = I_d)$$
$$\Rightarrow \min \sum_{l=1}^{d} p_l^T(-S)p_l + \gamma_l|p_l| \quad \left(\text{s.t. } p_l^T p_l = 1\right).$$

It is easy to find that SLE degrades to be $d$-independent DSA with the given scatter matrix $S$ for each projection $p_l$. The only difference is the sign of the matrix, which has no essential effect to the optimization problem. ■

Therefore, SLE not only intrinsically includes the previous SPCA algorithm and the DSA algorithm for the given scatter matrix but also integrates the local geometric structure to extend the existing subspace learning method to sparse cases. By defining different matrices $M$, the SLE model can derive different sparse subspace learning algorithms.

## V. Sparse Kernel Embedding

In this section, sparse kernel extension using the SLE model is presented and some theoretical analyses are also provided. Since the idea of SKE is similar to SLE except for operating on the kernel space, the modified kernel ridge regression is directly given and then the SKE algorithm is presented briefly.

### A. Kernel Technique

Recently, the sparse learning method was widely used in kernel learning [6], [41], [42]. However, unlike these methods, the novel sparse kernel model presented in this paper can be directly derived by the SLE model. Since this paper mainly focuses on the sparse linear model, the kernelized SLE, which is called as SKE, will be simply presented as an example to show that the proposed model can also be used for sparse nonlinear subspace learning as a general framework. It is known that KONPP [14] uses the kernel technique which maps the samples in the original input space into a potentially much higher dimensional space by a nonlinear mapping $\varphi : x \to \varphi(x)$. SKE also aims to preserve the local geometric structure in the kernel subspace for dimensionality reduction. In the kernel space, the kernel matrix can be defined as

$$K_{i,j} \triangleq k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle. \quad (21)$$

That is, $K \triangleq \varphi(X)^T\varphi(X)$. Then, the reconstruction matrix $W$ of the SKE can be obtained by the following optimization problem:

$$\min_W \sum_i \left\| \varphi(x_i) - \sum_{j \in N_k(x_i)} W_{ij}\varphi(x_j) \right\|^2 \quad \text{s.t. } \sum_{j \in N_k(x_i)} W_{ij} = 1.$$

The optimal reconstructive coefficient matrix in the kernel feature space can be used in the SKE algorithm.

### B. Kernel Formulation

Suppose the neighborhood matrix $M = (I - W)$ is given. Without misleading and losing generality, still let $B = (b_1, b_2, \ldots, b_d)$ and $P = (p_1, p_2, \ldots, p_d)$ be the matrices with the size of $N \times d$ that should be optimized in the kernel space

$$\min_{B,P} \sum_i \|BP^T\varphi(X)M_{i,:} - \varphi(X)^T\varphi(X)M_{i,:}\|_F^2$$
$$\text{s.t.} \quad B^TB = I_d. \tag{22}$$

Similar to other kernel algorithms, suppose $P$ are in the range of $\varphi(X)$, i.e., $P = \varphi(X)\tilde{P}$. Substituting $P = \varphi(X)\tilde{P}$ to (22) and adding the regularization term $\beta\|\tilde{P}\|_F^2$, we have the following optimization problem:

$$\begin{aligned}\min_{\tilde{B},\tilde{P}} &\sum_i \|B(\varphi(X)\tilde{P})^T\varphi(X)M_{i,:} - KM_{i,:}\|_F^2 + \beta\|\tilde{P}\|_F^2\\ &= \min_{\tilde{B},\tilde{P}} \operatorname{tr}[B\tilde{P}^TKM^TMK - 2\tilde{P}^TKM^TMK\tilde{B}\\ &\qquad + KM^TMK] + \beta\|\tilde{P}\|_F^2\\ &= \min_{\tilde{B},\tilde{P}} \operatorname{tr}[\tilde{P}^T(KM^TMK + \beta I)\tilde{P}\\ &\qquad - 2\tilde{P}^TKM^TMK\tilde{B} + KM^TMK].\end{aligned}$$

Thus, by discarding the constant term $KM^TMK$, the constraint regression problem becomes

$$\min_{\tilde{B},\tilde{P}} \operatorname{tr}[\tilde{P}^T(KM^TMK + \beta I)\tilde{P} - 2\tilde{P}^TKM^TMKB]$$
$$\text{s.t.} \quad B^TB = I_d. \tag{23}$$

For the above optimization problem, the following theorem is obtained.

*Theorem 7:* Supposed $Q = [q_1, q_2, \ldots, q_d]$ is the projection matrix of KONPP. Denote the SVD of $KM^T = \tilde{U}\tilde{D}\tilde{V}^T$, where $D$ contains the singular value in descending order. Let $\tilde{P}^*$ be the optimal solution to (23), then $\tilde{P}^* = [\tilde{p}_1^*, \tilde{p}_2^*, \ldots, \tilde{p}_d^*] = \tilde{U}(\tilde{D}^2/\tilde{D}^2 + \beta I)\Delta$, where diagonal matrix $\Delta$ only contains 1 or $-1$ in its diagonal elements. If the diagonal elements in $\Delta$ are all set to 1 and let $\underset{\sim}{p}_i^* = \tilde{p}_i^*/\|\tilde{p}_i^*\|$ for any $i$, then $\underset{\sim}{P}^* = [\underset{\sim}{p}_1^*, \ldots, \underset{\sim}{p}_d^*] = Q$.

*Proof:* The proof is similar to Theorem 3. ■

Theorem 7 reveals the relationship between KONPP and the regression problem (23), which also provides a tractable method for solving the sparse nonlinear learning problem by adding the $L_1$-norm penalty. The details are presented in Section V-C.

### C. SKE Algorithm

To obtain the sparse kernel subspace, we obtain the $L_1$ and $L_2$-norm penalty problem of SKE as follows:

$$\begin{aligned}\min_{B,\tilde{P}} &\sum_i \|B(\varphi(X)\tilde{P})^T\varphi(X)M_{i,:} - \varphi(X)\varphi(X)M_{i,:}\|_F^2\\ &\qquad + \beta\|\tilde{P}\|_F^2 + \beta\|\tilde{P}\|_F^2 + \sum_{l=1}^d \gamma_l|\tilde{p}_l|\end{aligned}$$
$$\text{s.t.} \quad B^TB = I_d \tag{24}$$

where $\beta > 0$ and $\gamma_l > 0$ are used for penalizing the loadings of different projection vectors. Similar to the formulation of the SLE algorithm, we obtain the final model of SKE as follows:

$$\begin{aligned}\min_{B,\tilde{P}} &\operatorname{tr}[\tilde{P}^T(KM^TMK + \beta I)\tilde{P} - 2\tilde{P}^TKM^TMKB]\\ &\qquad + \beta\|\tilde{P}\|_F^2 + \sum_{l=1}^d \gamma_l|\tilde{p}_l|\end{aligned}$$
$$\text{s.t.} \quad B^TB = I_d. \tag{25}$$

Following the same way as SLE, we can obtain the similar algorithm procedures. Since the details are very similar to SLE, it is omitted for avoiding repetition. Comparing (17) with (24), we can obtain the general framework for linear and nonlinear sparse embedding. The essential difference is to use the data in original space or kernel space for sparse regression.

## VI. Experiment

In this section, a set of experiments is presented to evaluate the proposed SLE algorithm for recognition tasks against the classical sparse learning methods (i.e., SPCA and SDA), the most related manifold learning methods, (i.e., NPE and ONPP), the manifold learning-based sparse subspace learning method USSL, the $L_1$-norm-based SPP, and the $L_1$ graph method proposed in [27]. The Yale face database was used to explore the robustness of SLE on the variations in expressions and illumination, and the performance of the properties on different parameters. The AR face database was employed to test the performance of SLE when there was a variation in time, facial expressions, and lighting conditions. The Carnegie Mellon University Pose, Illumination, and Expression (CMU PIE) database was used to evaluate the performance of these methods when face poses and lighting conditions vary dramatically. The COIL-20 and Caltech 101 databases were used to test the performance of the proposed algorithms in objective recognition. The nearest neighborhood classifier with the Euclidean distance was used in all experiments.

### A. Databases

The Yale face database (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 images from 15 individuals (each 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(a) shows sample images of one person in the Yale database.

The AR face database [43] contains over 4000 color face images from 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of 120 individuals (65 men and 55 women) were categorized into two sections (separated by two weeks) and each section contained 13 color images; 20 images from 120 individuals were selected and used in our experiments. The face portion of each image was manually cropped and then normalized to 50 × 40 pixels for computational efficiency. The sample images of one person are shown in Fig. 1(b).

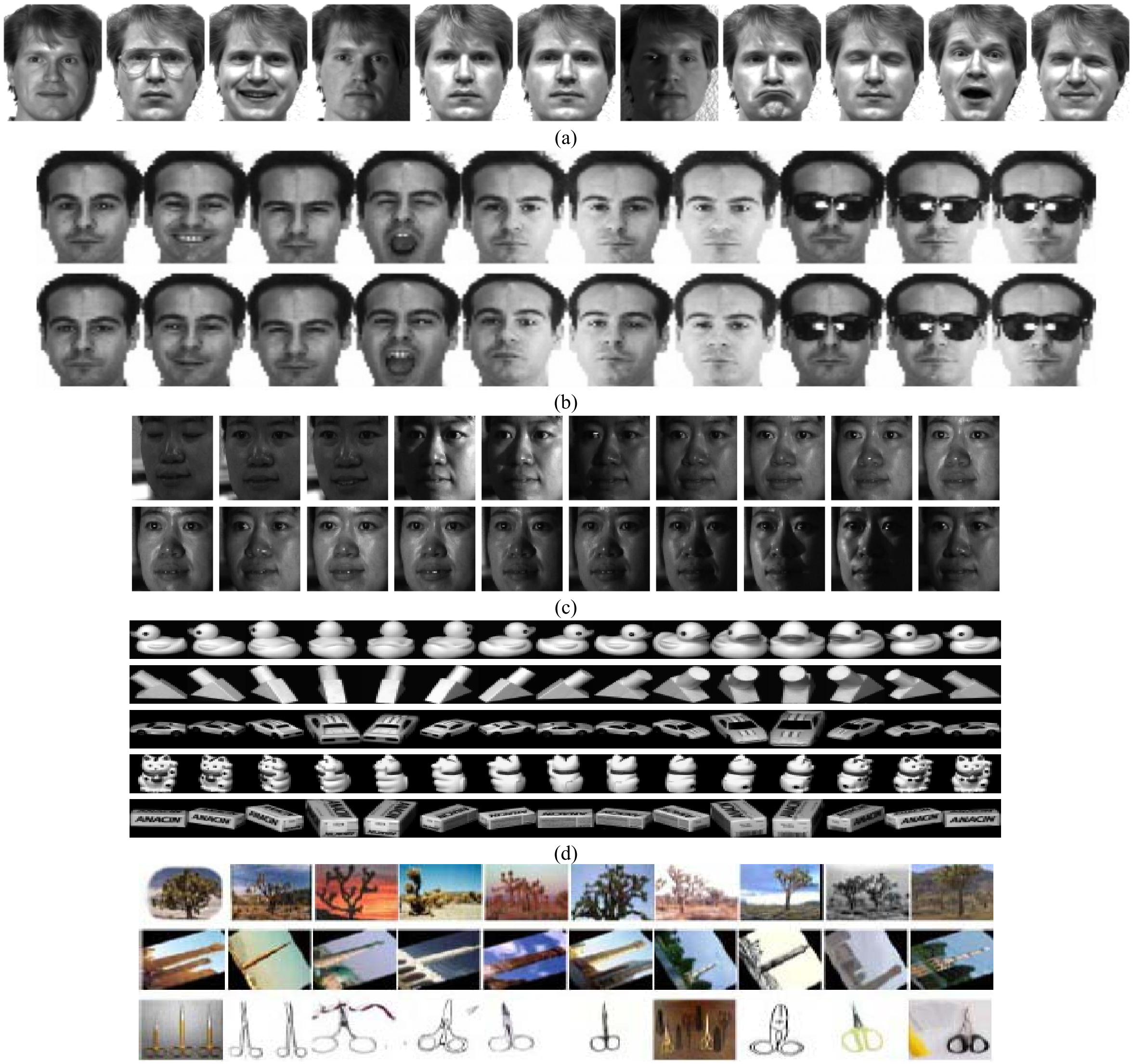


Fig. 1. Image samples used in the experiments. (a) Yale face database. (b) AR face database. (c) CMU PIE face database. (d) COIL-20 object image database. (e) Caltech 101 image database.

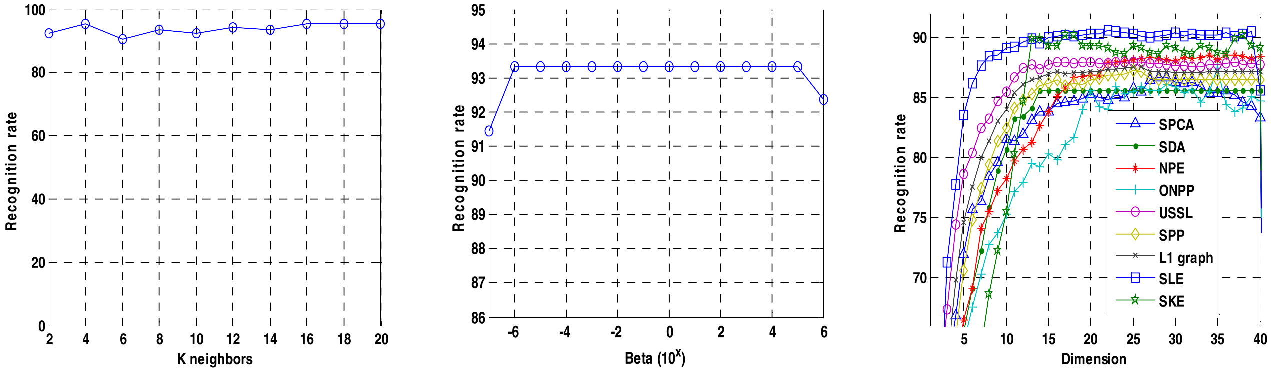
The CMU PIE face database [44] contains 68 individuals with 41 368 face images as a whole. The face images are captured in various poses, illuminations, and expressions. In our experiments, we selected a subset (C29), which contains 1632 images from 68 individuals (each providing 24 images). The C29 subset involves variations in illumination, facial expression, and pose. All of these face images were aligned-based one-eye coordinates and cropped to 32 × 32 pixels. Fig. 1(c) shows the sample images from this database.

The COIL-20 database (http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php) consists of $20 \times 72 = 1440$ images from 20 objects where the images of the object are taken at pose intervals of 5° (i.e., 72 poses per object). The original images were normalized to $128 \times 128$ pixels. All images were converted into a gray-scale image of $32 \times 32$ pixels for computational efficiency in the experiments. Some sample images of five objects are shown in Fig. 1(d).

Caltech 101 data set contains 9144 images from 102 classes (i.e., 101 object classes and a background class), including animals, vehicles, flowers, and so on. The samples from each category have significant shape variability. The number of images in each category varies from 31 to 800. For computational efficiency, a subset of 50 categories and each category contains 31 images was used in the experiment. Fig. 1(e) shows some images from the data set.

### B. Experimental Settings

In the experiments, $T (T = 4, 6)$ images of each individual were randomly selected and used as the training set, and one half of the remaining images as the validation set and the remainders as the test set. The experiments were independently conducted 10 times and the average recognition rates on the test set were calculated and reported. For each run, the validation set was used for parameter selection. When PCA



(a) (b) (c)

Fig. 2. Some properties of SLE. (a) Variations of recognition rate versus the parameter $k$. (b) Variation of recognition rate versus parameter $\beta$. (c) Recognition rates versus the dimension of each method on the Yale face database.

TABLE II
Comparison of Recognition Rate (Percentage), Standard Deviation, and Optimal Dimension on Yale Face Database

| T | SPCA | SDA | NPE | ONPP | USSL | SPP | L1 graph | SLE | SKE |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 86.63±3.27<br>28 | 85.57±2.91<br>14 | 86.67±5.17<br>39 | 87.42±5.16<br>35 | 87.23±2.77<br>25 | 87.90±4.27<br>16 | 87.19±4.53<br>37 | **90.57±2.70**<br>**22** | 90.19±4.14<br>17 |
| 6 | 89.55±2.55<br>35 | 90.27±7.41<br>14 | 89.93±3.80<br>31 | 90.31±4.03<br>31 | 91.07±2.73<br>35 | 92.00±5.15<br>24 | 92.21±3.53<br>36 | **93.66±3.43**<br>**25** | 93. 40±3.14<br>16 |

TABLE III
Comparison of Recognition Rate (Percentage), Standard Deviation, and Optimal Dimension on AR Face Database

| T | SPCA | SDA | NPE | ONPP | USSL | SPP | L1 graph | SLE | SKE |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 83.39±4.29<br>195 | 89.13±7.47<br>119 | 82.67±3.87<br>190 | 81.51±7.62<br>170 | 86.03±2.86<br>195 | 79.66±7.08<br>185 | 82.78±7.08<br>195 | **91.56±3.44**<br>**190** | 88.43±4.08<br>200 |
| 6 | 88.46±3.52<br>195 | 91.22±2.82<br>119 | 90.38±3.61<br>185 | 89.15±2.21<br>170 | 91.79±2.39<br>195 | 88.56±6.67<br>190 | 90.47±3.61<br>190 | 93.45±2.17<br>185 | **94.30±1.98**<br>**190** |

is used for preprocessing, about 98% of image energy is preserved as in the references. The optimal subspace dimensions are ranged in [1, 40] on the Yale and COIL-20 databases with step 1, and in [1, 150], [1, 200], and [1, 200] on the CMU PIE, AR, and Caltech 101 databases, respectively, with step 5, since within the higher dimension subspace, algorithms cannot achieve a higher recognition rate. In SPCA, SDA, USSL, SLE, and SKE, when the elastic net is used, $\beta$ is selected from $10^{-5}, 10^{-4}, \ldots, 10^{5}$, and the parameters $\gamma_l$ can be automatically determined, since the elastic net algorithm can provide the optimal solution path of $\gamma_l$ for given $\beta$ [34]. The nearest neighbor parameter $k$ in USSL, NPE, ONPP, SLE, and SKE is selected from $1, 2, 4, \ldots, N-1$. The Gaussian kernel function is used for SKE to construct the nonlinear data, in which the kernel width parameter $\delta$ is selected from the 0.001, 0.01, and 0.1 times the mean distance of the training data points. For each run, the optimal parameters determined by the validation set are used in the algorithms to learn the projections.

The variations of the parameters versus the recognition rates in the Yale face database based on a single run are shown in Fig. 2, which shows that when $k = 4$ or $k \geq 15$ SLE achieves its best performance, the top recognition rates have slight variations to the value of $k$. However, as shown in Fig. 2(b), SLE is very robust to the value of $\beta$ in a large range. In other words, the top recognition rate can be achieved when $\beta$ varies from $10^{-6}$ to $10^{5}$. Similar properties exist on other databases.

### C. Experimental Results and Its Analysis

The average recognition rates of each method in the Yale, AR, CMU PIE face databases, the COIL-20, and the Caltech 101 image database are shown in Tables II–VI, respectively. The average recognition rates versus the dimension of each method on these databases are shown in Figs. 2 and 3. In Fig. 3(a) and (b), five times of the numbers marked on the horizontal axes are the real values of the optimal dimensions shown in Tables III and IV (the dimension step is set to be 5). Based on these experimental results, we have the following observations and corresponding analyses.

1) SLE and SKE consider both the locality of the image manifold structure, the approximate orthogonality and the sparsity in subspace learning. Therefore, it performs better than other methods, which only have one or

TABLE IV
COMPARISON OF RECOGNITION RATE (PERCENTAGE), STANDARD DEVIATION, AND OPTIMAL DIMENSION ON CMU PIE FACE DATABASE

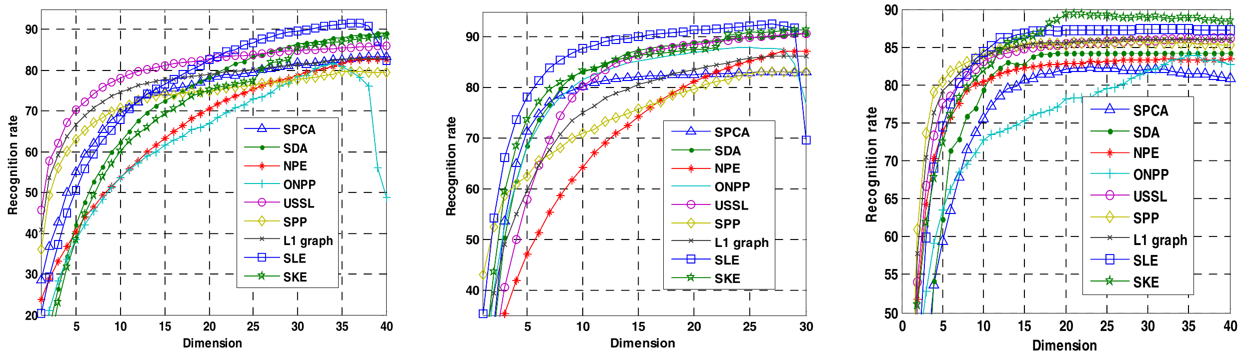
| T | SPCA | SDA | NPE | ONPP | USSL | SPP | L1 graph | SLE | SKE |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 68.01±15.37<br>145 | 75.50±7.24<br>67 | 74.59±7.11<br>135 | 74.98±8.11<br>125 | 75.51±6.04<br>140 | 61.19±13.09<br>130 | 63.45±11.42<br>135 | **77.89±5.62**<br>**130** | 76.32±4.54<br>150 |
| 6 | 83.12±6.62<br>150 | 90.90±4.05<br>67 | 87.27±5.66<br>145 | 88.05±7.73<br>125 | 90.61±5.39<br>145 | 83.17±8.13<br>140 | 86.30±6.22<br>145 | **92.51±3.48**<br>**135** | 91.38±3.64<br>145 |

TABLE V
COMPARISON OF RECOGNITION RATE (PERCENTAGE), STANDARD DEVIATION, AND OPTIMAL DIMENSION ON COIL-20 FACE DATABASE

| T | SPCA | SDA | NPE | ONPP | USSL | SPP | L1 graph | SLE | SKE |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 76.87±3.32<br>26 | 78.49±4.68<br>19 | 77.48±9.57<br>38 | 80.06±3.33<br>37 | 80.97±3.70<br>38 | 80.79±2.89<br>16 | 80.82±2.98<br>35 | 82.43±2.83<br>39 | **84.05±2.73**<br>**21** |
| 6 | 82.32±2.01<br>23 | 84.14±4.06<br>19 | 83.37±6.88<br>33 | 83.91±2.56<br>36 | 86.18±2.26<br>39 | 85.72±2.87<br>27 | 86.05±2.71<br>33 | 86.12±2.61<br>38 | **89.28±2.12**<br>**22** |

TABLE VI
COMPARISON OF RECOGNITION RATE (PERCENTAGE), STANDARD DEVIATION, AND OPTIMAL DIMENSION ON CALTECH 101 IMAGE DATABASE

| T | SPCA | SDA | NPE | ONPP | USSL | SPP | L1 graph | SLE | SKE |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 37.64±1.52<br>120 | 49.05±0.74<br>49 | 45.16±1.19<br>75 | 43.98±0.89<br>90 | 40.18±1.10<br>145 | 43.20±0.85<br>85 | 45.87±1.25<br>90 | 52.05±1.54<br>80 | **53.44±1.93**<br>**55** |
| 25 | 46.00±2.01<br>110 | 56.63±1.41<br>49 | 51.26±2.20<br>90 | 50.93±1.90<br>85 | 46.84±2.03<br>145 | 49.63±2.27<br>85 | 52.28±1.77<br>135 | **59.00±1.92**<br>**190** | 58.10±1.91<br>50 |



(a) (b) (c)

Fig. 3. Recognition rates versus the dimension of each method on (a) AR, (b) CMU PIE, and (c) COIL-20 face database.

two properties. There is no complete vector for the proposed SKE and SLE, which can be found in Tables II, III, and VI.

2) Since USSL and SLE take the local geometric structure into account, they usually perform better than SPCA and SDA, which only take the global information in the learning steps. SLE performs better than USSL, since SLE has approximate orthogonal projections and the projections of USSL are learnt independently and the orthogonality cannot be preserved. For the challenging data set, such as Caltech 101 database, SLE (and SDA) significantly performs better than NPE and ONPP through sparse feature selection.

3) The common property of NPE, ONPP, SPP, $L_1$ graph, and SLE is that they all aim to preserve the reconstructive coefficient relationship. It is found that SPP, $L_1$ graph, and SLE introducing the sparsity perform better than NPE and ONPP, which only use the $L_2$ norm as the measurement. However, it is also found in Tables III and IV ($T = 4$) that when there is only a very small number of training samples per individual, the recognition rate of SPP is lower than that of the other method and is significantly lower than those of SLE and SKE. This indicates that in the cases of very small sample size, introducing the sparsity to projections can obtain better performance than to enforce the sparsity on the reconstruction coefficients. This also demonstrates that the methods with the sparsity constraint on the projections can be more robust/stable with the lack of training samples.

4) SDA, which only focuses on the global information, is a supervised sparse subspace learning method, which extends LDA to sparse case. It can only obtain $C-1$-dimensional subspace. Moreover, it is difficult for SDA to accurately estimate the optimal score in a very high-dimensional space (i.e., in the original space of the data set) for regression. Therefore, SDA cannot perform better than SLE and SKE even though label information is considered. Another potential reason may be that the supervised method (i.e., SDA) is more likely to overfit with a few training samples, whereas the unsupervised methods are able to find a common face manifold before classification and thus avoid overfitting.

## VII. Conclusion

In this paper, a general framework preserving the local geometric structure and orthogonality among the projections is proposed for sparse subspace learning. Theoretical analyses show that the optimal sparse subspace of SLE approximates to the subspace of ONPP, which guarantees the effectiveness in dimensionality reduction. It is also shown that this framework can be easily used in kernel form, which derives the SKE for nonlinear dimensionality reduction. The optimization problem can be solved using the iterative procedures, which combine the elastic net algorithm and SVD together. Experiments on four well-known image databases demonstrate that SLE and SKE perform better than the previous sparse subspace learning algorithms and the most related nonsparse subspace algorithms in feature extraction. In the future, we will explore a more general framework for sparse subspace learning when there exist multiconstraints and preserve different geometry structures. In addition, the supervised SLE and SKE are also the important research issues, which are beyond the scope of this paper and will be explored in another paper in detail.

## Appendix
## Proof of Theorem 3

$$\begin{aligned}
&\sum_i \|bp^T XM_{i,:}^T - XM_{i,:}^T\|_F^2 + \beta\|p\|^2 \\
&= \sum_i \operatorname{tr}\big[bp^T XM_{i,:}^T\big(bp^T XM_{i,:}^T\big)^T - 2bp^T XM_{i,:}^T\big(XM_{i,:}^T\big)^T \\
&\qquad + XM_{i,:}^T\big(XM_{i,:}^T\big)^T\big] + \beta\operatorname{tr}(pp^T) \\
&= \operatorname{tr}(XM^TMX^T - 2bp^TXM^TMX^T \\
&\qquad + pp^T(XM^TMX^T + \beta I)). \qquad\qquad (A1)
\end{aligned}$$

For fixed $b$, taking the partial deviation of (A1) with respect to $p$ to be zero gives

$$-2XM^TMX^Tb + 2(XM^TMX^T + \beta I)p = 0$$

thus

$$p = (XM^TMX^T + \beta I)^{-1}XM^TMX^Tb. \qquad (A2)$$

Substituting (A2) in (A1) derives the following optimization problem:

$$\begin{aligned}
\min\ & \operatorname{tr}(XM^TMX^T - bb^TXM^TMX^T \\
&\quad \times (XM^TMX^T + \beta I)^{-1}XM^TMX^T) \\
\text{s.t.}\ & b^Tb = 1.
\end{aligned}$$

Since $XM^TMX^T$ is a constant, the minimization problem converts to the maximization problem

$$\begin{aligned}
\max\ & \operatorname{tr}(b^TXM^TMX^T(XM^TMX^T + \beta I)^{-1}XM^TMX^Tb) \\
\text{s.t.}\ & b^Tb = 1. \qquad\qquad (A3)
\end{aligned}$$

Let the SVD of $X(I-W)^T = UDV^T$, where $D$ contains the singular value in ascending order. We have

$$\begin{aligned}
&XM^TMX^T(XM^TMX^T + \beta I)^{-1}XM^TMX^T \\
&\qquad = U\frac{D^4}{D^2 + \beta I}U^T. \qquad (A4)
\end{aligned}$$

Hence, $b^* = \delta U(:,\text{end})$ is the optimal solution to (A3), where $\delta = 1$ or $-1$ and $U(:,\text{end})$ denote the last left singular vector (corresponding to the largest singular value).

Let $D_1$ be the singular value corresponding to $U(:,\text{end})$. Substituting $b^* = \delta U(:,\text{end})$ in (A2), we obtain the optimal $p^*$

$$\begin{aligned}
p^* &= (XM^TMX^T + \beta I)^{-1}XM^TMX^T\delta U(:,\text{end}) \\
&= \delta\frac{D_1^2}{D_1^2 + \beta}U(:,\text{end}) \qquad (A5)
\end{aligned}$$

this gives the results. ■

## References

[1] L. Sirovich and M. Kirby, “Low-dimensional procedure for the characterization of human faces,” *J. Opt. Soc. Amer. A, Opt. Image Sci. Vis.*, vol. 4, no. 3, pp. 519–524, 1987.

[2] M. Kirby and L. Sirovich, “Application of the Karhunen–Loeve procedure for the characterization of human faces,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[3] M. Turk and A. Pentland, “Eigenfaces for recognition,” *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[4] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[5] C. Fraley and A. E. Raftery, “Model-based clustering, discriminant analysis, and density estimation,” *J. Amer. Statist. Assoc.*, vol. 97, no. 458, pp. 611–631, 2002.

[6] Q. Liu, H. Lu, and S. Ma, “Improving kernel Fisher discriminant analysis for face recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 42–49, Jan. 2004.

[7] J. Li and D. Tao, “Simple exponential family PCA,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 3, pp. 485–497, Mar. 2013.

[8] D. Tao, X. Li, S. J. Maybank, and X. Wu, “Geometric mean for subspace selection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[9] J. Li and D. Tao, “On preserving original variables in Bayesian PCA with application to image analysis,” *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4830–4843, Dec. 2012.

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

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

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

[13] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2005, pp. 1208–1213.

[14] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2143–2156, Dec. 2007.

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

[16] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, vol. 16. Cambridge, MA, USA: MIT Press, 2003, pp. 153–160.

[17] S. Liu and Q. Ruan, "Orthogonal tensor neighborhood preserving embedding for facial expression recognition," *Pattern Recognit.*, vol. 44, no. 7, pp. 1497–1513, Jul. 2011.

[18] M. Türkan and C. Guillemot, "Image prediction based on neighbor-embedding methods," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1885–1898, Apr. 2012.

[19] W. Bian and D. Tao, "Biased discriminant Euclidean embedding for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 545–554, Feb. 2010.

[20] B.-Y. Sun, X.-M. Zhang, J. Li, and X.-M. Mao, "Feature fusion using locally linear embedding for classification," *IEEE Trans. Neural Netw.*, vol. 21, no. 1, pp. 163–168, Jan. 2010.

[21] T. Zhang, K. Huang, X. Li, J. Yang, and D. Tao, "Discriminative orthogonal neighborhood-preserving projections for classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 253–263, Feb. 2010.

[22] X. Li, S. Lin, S. Yan, and D. Xu, "Discriminant locally linear embedding with high-order tensor data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 2, pp. 342–352, Apr. 2008.

[23] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[24] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Two-stage nonnegative sparse representation for large-scale face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 35–46, Jan. 2013.

[25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[26] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of $L_1$-optimizer in pattern classification," *Pattern Recognit.*, vol. 45, no. 3, pp. 1104–1118, Mar. 2012.

[27] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with $\ell^1$-graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

[28] F. Zang and J.-S. Zhang, "Label propagation through sparse neighborhood and its applications," *Neurocomputing*, vol. 97, pp. 267–277, Nov. 2012.

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

[30] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji, "Discriminant sparse neighborhood preserving embedding for face recognition," *Pattern Recognit.*, vol. 45, no. 8, pp. 2884–2893, Aug. 2012.

[31] 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.

[32] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

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

[34] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

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

[36] T. Zhou and D. Tao, "Double shrinking sparse dimension reduction," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 244–257, Jan. 2013.

[37] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[38] D. Cai, X. He, and J. Han, "Spectral regression: A unified approach for sparse subspace learning," in *Proc. 7th IEEE Int. Conf. Data Mining*, Oct. 2007, pp. 73–82.

[39] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA, USA: Athena Scientific, 2003.

[40] R. T. Rockafellar, *Convex Analysis*, 2nd ed. Princeton, NJ, USA: Princeton Univ. Press, 1972.

[41] H. Yang, Z. Xu, J. Ye, I. King, and M. R. Lyu, "Efficient sparse generalized multiple kernel learning," *IEEE Trans. Neural Netw.*, vol. 22, no. 3, pp. 433–446, Mar. 2011.

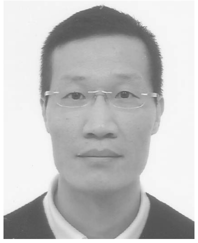
[42] A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu, "$\ell_p$–$\ell_q$ penalty for sparse linear and sparse multiple kernel multitask learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 22, no. 8, pp. 1307–1320, Apr. 2011.

[43] A. Martinez and R. Benavente, "The AR face database," CVC, Tech. Rep. 24, 1998.

[44] 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, in 2002, the M.S. degree from Jinan University, Guangzhou, in 2007, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, China, in 2011.

He was a Research Associate and a Post-Doctoral Fellow with The Hong Kong Polytechnic University, Hong Kong, from 2010 to 2014. His current research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization, and applications in the fields of 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 over 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*, and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS. His current research interests include artificial intelligence, pattern recognition, and optimization of manufacturing scheduling, planning and control.

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

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

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

He was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain, in 2003. He was a Post-Doctoral Fellow with the Biometrics Centre, The Hong Kong Polytechnic University, Hong Kong, from 2004 to 2006, and the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA, from 2006 to 2007. He is currently a Professor with the School of Computer Science and Technology, NUST. He has authored over 80 scientific papers in pattern recognition and computer vision. His journal papers have been cited over 2000 times in the ISI Web of Science, and 4000 times in a Google Scholar. His current research interests include pattern recognition, computer vision, and machine learning.

Prof. Yang received the RyC Program Research Fellowship from the Spanish Ministry of Science and Technology.

**David Zhang** (F'08) received the Degree in computer science from Peking University, Beijing, China, and the M.Sc. degree in computer science and the Ph.D. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively, and the second Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 1994.

He was a Post-Doctoral Fellow with Tsinghua University, Beijing, and an Associate Professor with Academia Sinica, Beijing, from 1986 to 1988. He is currently the Head of the Department of Computing and the Chair Professor with The Hong Kong Polytechnic University, Hong Kong. He also serves as a Visiting Chair Professor with Tsinghua University, and an Adjunct Professor with Peking University, Shanghai Jiao Tong University, Shanghai, HIT, and the University of Waterloo. He has authored over ten books and 200 journal papers.

Prof. Zhang is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, and a fellow of the International Association for Pattern Recognition. He is the Founder and Editor-in-Chief of the *International Journal of Image and Graphics*, a Book Editor of *International Series on Biometrics* (Springer), an Organizer of the International Conference on Biometrics Authentication, and an Associate Editor of over ten international journals, including the IEEE TRANSACTIONS and *Pattern Recognition*.