Unified Sparse Subspace Learning via Self-Contained Regression

Shuangyan Yi, *Student Member, IEEE*, Zhenyu He, *Senior Member, IEEE*, Yiu-Ming Cheung, *Senior Member, IEEE*, and Wen-Sheng Chen

***Abstract*— In order to improve the interpretation of principal components, many sparse principal component analysis (PCA) methods have been proposed by in the form of self-contained regression-type. In this paper, we generalize the steps needed to move from PCA-like methods to its self-contained regression-type, and propose a joint sparse pixel weighted PCA method. More specifically, we generalize a self-contained regression-type framework of graph embedding. Unlike the regression-type of graph embedding relying on the regular low-dimensional data, the self-contained regression-type framework does not rely on the regular low-dimensional data of graph embedding. The learned low-dimensional data in the form of self-contained regression theoretically approximates to the regular low-dimensional data. Under this self-contained regression-type, sparse regularization term can be arbitrarily added, and hence, the learned sparse regression coefficients can interpret the low-dimensional data. By using the joint sparse $\ell_{2,1}$-norm regularizer, a sparse self-contained regression-type of pixel weighted PCA can be produced. Experiments on six data sets demonstrate that the proposed method is both feasible and effective.**

***Index Terms*— Weighted PCA, self-contained regression-type, sparse subspace learning.**

I. Introduction

SUBSPACE learning theories, which aim to extract the effective features, can be applied in many fields, such as recognition [2]–[7], [10], [11], [44], [60], writer identification [50], [59], saliency detection [1], [57], and tracking [43], [45]–[49], [56], [58]. The classical linear subspace learning methods include principal component analysis (PCA) [12], [13], locality preserving projections (LPP) [14], neighborhood preserving embedding (NPE) [15], and linear discriminant analysis (LDA) [16]. The classical nonlinear subspace learning methods include Isomap [17], laplacian eigenmaps (LE) [18], locally linear embedding (LLE) [19] and several others. Most of subspace learning methods can be nicely interpreted by the graph embedding (GE) framework [20], requiring the solution of an eigenvalue eigenproblem. The main drawback of graph embedding is that it cannot deal with the newly coming data samples. Linear graph embedding (LGE) [20] has therefore been widely developed and adopted in real-life applications in preference.

In the LGE framework, the classical methods are PCA and LPP, and they play an important role in reconstruction and classification, respectively. However, two key problems still exist: the first is that both PCA and LPP are sensitive to corruption since the former relies on the least squares estimation technique and the latter relies on a graph computation and the second is that it is difficult to interpret their results since each principal component is a linear combination of all the original features. A series of studies have therefore been carried out to deal with these two problems.

In order to deal with the first problem, principal component analysis based on ℓ_1-norm maximization (PCAL1) [52] has been proposed to enhance the robustness to outliers by using the ℓ_1-norm to maximize the transformed data variance. In addition, weighted PCA methods have been proposed, which can generally be classified into two types using weighting of samples and pixels, respectively. An example of this is the use of rotational invariant ℓ_1-norm principal component analysis for robust subspace factorization (R1-PCA) [21] and optimal mean robust principal component analysis (OMRPCA) [22] to enforce the robustness to sample-specific corruption by weighting each sample to soften sample-specific corruption. In fact, most of the weighted PCA methods [23] focus on the weighting of samples while several methods focus on the weighting of pixels [8], [9]. To the best of our knowledge, Torre and Black [24], [25] were the first to avoid the effect of intra-sample corruption by weighting each pixel of each sample. Robust principal component analysis (RPCA) [26] has been proposed to handle the sample-specific corruption by imposing a low-rank constraint on all the data samples, which is very effective in repairing sample-specific corruption. Recently, two-dimensional whiten-

Manuscript received December 31, 2016; revised May 24, 2017; accepted June 14, 2017. Date of publication June 29, 2017; date of current version October 24, 2018. This work was supported in part by the Shenzhen Research Council under Grant JCY20170413104556946, Grant JCYJ20160406161948211, Grant JCYJ20160226201453085, and Grant JSGG20150331152017052, in part by the National Natural Science Foundation of China under Grant 61672183, Grant 61272252, Grant 61272366, and Grant 61672444, in part by the Science and Technology Planning Project of Guanddong Province under Grant 2016B090918047, in part by the Natural Science Foundation of Guangdong Province under Grant 2015A030313544, in part by the Faculty Research Grant of Hong Kong Baptist University under Project FRG2/16-17/051 and Project FRG2/15-16/049, in part by the KTO Grant of HKBU under Project MPCF-004-2017/18, and in part by SZSTI under Grant JCYJ20160531194006833. This paper was recommended by Associate Editor L. Lin. *(Corresponding author: Zhenyu He.)*

S. Yi is with the School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China, and also with the Institute of Research and Continuing Education, Hong Kong Baptist University, Hong Kong (e-mail: shuangyanshuangfei@163.com).

Z. He is with the School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China (e-mail: zhenyuhe@hit.edu.cn).

Y.-M. Cheung is with the Department of Computer Science and the Institute of Research and Continuing Education, Hong Kong Baptist University (HKBU), Hong Kong, and also with the United International College, Beijing Normal University-HKBU, Zhuhai 519000, China (e-mail: ymc@comp.hkbu.edu.hk).

W.-S. Chen is with the Shenzhen Key Laboratory of Media Security, College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, China (e-mail: chenws@szu.edu.cn).

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/TCSVT.2017.2721541

ing reconstruction (TWR) [27] has been proposed to enhance the robustness of principal component analysis by reducing the number of redundant features and maintaining the important intrinsic features. Moreover, in order to enhance the robustness of LPP, researchers have often introduced the low-rank representation technique into graph construction [28].

In order to deal with the second problem, a series of sparse linear graph embedding methods (SLGE) have been proposed in the literature [29], [51], [54], [55], whose regression way can be classified as regression-type and self-contained regression-type. The goal of both regression-type and self-contained regression-type of graph embedding is to learn a sparse regression coefficients such that the learned low-dimensional data approximates to the regular low-dimensional data. For example, Cai *et al.* [20] proposed a unified sparse subspace learning (USSL) framework by writing a graph embedding formulation [14] into its regression-type. More specifically, USSL includes two steps: the first is the computation of the regular low-dimensional data of graph embedding, and the second is the sparse regression for the regular low-dimensional data. In a similar way to USSL, joint feature selection and subspace learning (JFSSL) [30] also uses a two-step regression method, which aims to regress the regular low-dimensional data of LPP by adding the joint sparse $\ell_{2,1}$-penalized regularization to the regression coefficients. Both USSL and JFSSL rely on obtaining the regular low-dimensional data in advance. Furthermore, Zheng [31] proposed a sparse locality preserving embedding by incorporating LPP into its self-contained regression-type. This can be said to integrate the two steps of USSL into a single step, which does not rely on the regular low-dimensional data. This strategy is named as self-contained regression-type. It is worth noting that the low-dimensional data in PCA-like methods is called principal components. Accordingly, sparse principal component analysis (SPCA) [32] is produced by writing PCA into its sparse self-contained regression-type. Unlike the regression-type of PCA, the self-contained regression-type of PCA does not require prior knowledge of the regular principal components.

It can be observed that the self-contained regression-type of PCA and LPP have been well constructed. Both of them use ℓ_1-norm in their regression-type and form an elastic net framework, which is solved by the LARS-EN algorithm [33]. Once the regression-type or self-contained regression-type of one method is formed, sparse regularization term can be arbitrarily added. Recently, the regularized term based joint sparse $\ell_{2,1}$-norm has been popular and has been widely adopted. For example, methods [30], [34], [35] based on the joint sparse $\ell_{2,1}$-norm can clearly select those effective features for classification. In this paper, we focus on the unsupervised dimensionality reduction methods based on PCA, since PCA has been demonstrated to be a popular technique. Although numerous PCA versions have been developed, they still lack a powerful interpretation of principal components, especially in those methods with regression-type. Inspired by joint sparse principal component analysis (JSPCA) [36], we propose a joint sparse pixel weighted PCA version. The proposed method is the self-contained regression-type of pixel weighted PCA in practice. Experiments on six datasets demonstrate the both feasibility and effectiveness of the proposed method. The main contributions are listed as follows:

1. We generalize a self-contained regression-type framework of graph embedding.

2. Under the generalized framework, we propose a joint sparse pixel weighted PCA method.

The remainder of this paper is organized as follows. In Section II, graph embedding is reviewed and the self-contained regression-type of graph embedding is generalized. Section III presents the proposed joint sparse pixel weighted PCA method with an effective solution. Section IV describes the experiments results and analysis. Finally, a short conclusion is drawn in Section V.

II. Related Work

In this section, we first clarify the concept of principal component and then reveal the essential relationship among PCA, the regression-type of PCA and the self-contained regression-type of PCA.

Notations: For a matrix A, we denote the (i, j)-th element by a_{ij}, the i-th row by $\boldsymbol{a}^i$. In this paper, we denote $\|A\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} {a_{ij}}^2$, $\|\boldsymbol{A}\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$, $\|\boldsymbol{A}\|_{2,1} = \sum_{i=1}^{m}\left\|\boldsymbol{a}^i\right\|_2$, where $\left\|\boldsymbol{a}^i\right\|_2$ means the ℓ_2, 1-norm of vector $\boldsymbol{a}^i$ and $\left\|\boldsymbol{a}^i\right\|_2 = \sqrt{{a^i}^T\boldsymbol{a}^i}$.

A. Graph Embedding Framework

Assume that there are n data samples, i.e., $X = [x_1, x_2, ...x_n] \in \mathbb{R}^{m\times n}$, where n is the number of samples and m is the dimensionality of a sample. Let a graph G be built by these n data samples, and a data sample is represented by each vertex of G. Let W be a symmetric $n \times n$ matrix, whose element W_{ij} means the weight of the edge joining vertices x_i and x_j. Therefore, the defined G and W can be used to characterize the geometric properties of the dataset. The purpose of graph embedding [20] is to represent each vertex of G as a low-dimensional vector that preserves similarities between the vertex pairs [53]. Let $y = [y_1, y_2, ..., y_n] \in \mathbb{R}^n$ represent a one-dimensional data, the objective function is defined as follows:

$$\min_{y}\sum_{i,j}(y_i - y_j)^2 W_{ij}. \tag{1}$$

Since $\sum_{i,j}(y_i - y_j)^2 W_{ij} = 2y^T L y$ [20], Eq. (1) becomes,

$$\min_{y} y^T L y, \tag{2}$$

where $L = D - W$ is a Laplacian matrix, D is a diagonal matrix whose diagonal element is $D_{ii} = \sum_j W_{ji}$. In order to avoid the degenerated solution $y = 0$, the constraint $y^T D y = 1$ is added. Therefore, Eq. (2) becomes,

$$\begin{aligned} &\min_{y}\ y^T L y, \\ &s.t.\ y^T D y = 1, \end{aligned} \tag{3}$$

whose optimization solution can be solved by solving the minimal eigenvalue of the following generalized eigenvalue problem,

$$Ly = \gamma\, Dy. \tag{4}$$

where γ is an eigenvalue of matrix L relative to matrix D. The nonzero solution y corresponding to γ is called eigenvector belonging to γ. Now, we extend the dimensionality of low-dimensional data to d, and denote $Y = [Y_1, Y_2, ..., Y_d] \in \mathbb{R}^{n\times d}$. Then, we have the following graph embedding formula,

$$LY = DY\Upsilon. \tag{5}$$

where Υ is a diagonal matrix whose diagonal elements are assembled by eigenvalues, and the columns of Y are assembled by eigenvectors.

Denote $Y = X^T Q$, Eq. (5) becomes,

$$LX^T Q = DX^T Q\Upsilon. \tag{6}$$

Multiplying X with both sides of Eq. (6), the following linear graph embedding formula is obtained,

$$XLX^T Q = XDX^T Q\Upsilon. \tag{7}$$

Equivalently, we have,

$$XDX^T Q = XLX^T Q\Upsilon^{-1}. \tag{8}$$

Denote $M_D = XDX^T$, $M_L = XLX^T$, and $\Upsilon^{-1} = \Sigma$, then we need to solve the maximum eigenvalues of the following generalized eigenvalue problem,

$$M_D Q = M_L Q\Sigma. \tag{9}$$

B. The Self-Contained Regression-Type of Graph Embedding Framework

If M_L is invertible and has a cholesky decomposition $M_L = G_L G_L^T$ where $G_L \in \mathbb{R}^{m\times m}$ is a lower triangle matrix, LGE will have the self-contained regression-type.

More specifically, denote the following notations:

$$\begin{aligned} M_D &= F_D^T F_D, \quad M_L = F_L^T F_L, \\ F_D &= \sqrt{D}X^T, \quad F_L = \sqrt{L}X^T. \end{aligned} \tag{10}$$

Note that, both M_D and M_L are symmetric and positive semidefinite. Substitute $M_L = G_L G_L^T$ into Eq. (9), we have,

$$G_L^{-1} M_D Q = G_L^T Q\Sigma, \tag{11}$$

and

$$\begin{aligned} G_L^{-1} F_D^T F_D G_L^{-T}(G_L^T Q) &= (G_L^{-1} F_D^T)(G_L^{-1} F_D^T)^T (G_L^T Q) \\ &= (G_L^T Q)\Sigma. \end{aligned} \tag{12}$$

Let $C = G_L^T Q \in \mathbb{R}^{m\times d}$, similar to the self-contained regression-type of PCA, we generalize the following self-contained regression-type framework of graph embedding,

$$\begin{aligned} &\min_{B,C} \ \| G_L^{-1} F_D^T - BC^T G_L^{-1} F_D^T \|_F^2 + \lambda \| C \|_F^2, \\ &s.t.\ B^T B = I, \end{aligned} \tag{13}$$

where $B \in \mathbb{R}^{m\times d}$. Eq. (13) is equivalent to,

$$\begin{aligned} &\min_{B,C} \ \sum_{i=1}^{n} \| G_L^{-1} F_{D,i} - BC^T G_L^{-1} F_{D,i} \|^2 \\ &\qquad + \lambda \sum_{j=1}^{d} (G_L^{-T} C_j)^T M_L (G_L^{-T} C_j), \\ &s.t.\ B^T B = I, \end{aligned} \tag{14}$$

where $F_{D,i}$ is the transpose of the i-th row of F_D.

Let $Q = G_L^{-T} C$, we have,

$$\begin{aligned} &\min_{B,Q} \ \sum_{i=1}^{n} \| G_L^{-1} F_{D,i} - BQ^T F_{D,i} \|^2 + \lambda \sum_{j=1}^{d} Q_j^T M_L Q_j. \\ &s.t.\ B^T B = I, \end{aligned} \tag{15}$$

Eq. (15) is just the self-contained regression-type of LPP in [31]. Besides, when $M_L = I$ and $M_D = XX^T$, Eq. (9) becomes $XX^T Q = Q\Sigma$. Correspondingly, Eq. (13) becomes the self-contained regression-type of PCA.

III. Unified Sparse Subspace Learning via Self-Contained Regression

A. Joint Sparse Pixel Weighted PCA via Self-Contained Regression

Inspired by the generalized self-contained regression-type framework and the previous study [36], we propose the following optimization formulation:

$$\begin{aligned} &\arg\min_{Q,P} \left\| X - PQ^T X \right\|_{2,1} + \lambda \| Q \|_{2,1} + \beta \left\| \sqrt{D_1}^{-1} Q \right\|_F^2, \\ &s.t.,\ P^T D_1 P = I, \end{aligned} \tag{16}$$

where $\sqrt{D_1}$ is a diagonal matrix, whose diagonal elements are the adaptive weight of each pixel, and D_1 is designed according to the reconstruction case of each pixel as follows:

$$D_1 = \begin{bmatrix} \frac{1}{2\left\| [X - PQ^T X]^1 \right\|_2} & & \\ & \frac{1}{2\left\| [X - PQ^T X]^2 \right\|_2} & \\ & & \ddots \end{bmatrix}, \tag{17}$$

where projection matrix $Q \in \mathbb{R}^{m\times d}$ is first used to project the data matrix X onto a low-dimensional subspace and another projection matrix $P \in \mathbb{R}^{m\times d}$ is then used to recover the data matrix X. Here, we relax the orthogonal constraint of projection matrix Q, introduce another projection matrix P and add the joint sparse $\ell_{2,1}$-norm onto the loss term $X - PQ^T X$ and the projection matrix Q, where the $\ell_{2,1}$-norm is defined to encourage the rows of a matrix to be zero. For example, the penalty term $\|Q\|_{2,1}$ penalizes all d elements in each row and finally obtains m penalizing values. The obtained penalizing values are usually used to indicate the significance of the features. That is, when the penalizing value approximates to 0, the corresponding feature is regarded as the redundant feature. $\left\| \sqrt{D_1}^{-1} Q \right\|_F^2$ aims to obtain the stable optimal solution. $\lambda \geq 0$ and $\beta \geq 0$, as the regularization parameters, are used to balance the relation between three terms.

Directly solving or interpreting Eq (16) is difficult. Using some mathematical techniques for Eq. (16), we have,

$$\begin{aligned}&\arg\min_{Q,P} \left\|\sqrt{D_1}(X - PQ^TX)\right\|_F^2 + \lambda\left\|\sqrt{D_2}Q\right\|_F^2 \\ &\qquad + \beta\left\|\sqrt{D_1}^{-1}Q\right\|_F^2, \\ &s.t.,\ \ P^TD_1P = I,\end{aligned} \tag{18}$$

where D_1 is computed according to Eq. (17), and

$$D_2 = \begin{bmatrix} \frac{1}{2\|q^1\|_2} & & \\ & \frac{1}{2\|q^2\|_2} & \\ & & \ddots \end{bmatrix}, \tag{19}$$

are two $m \times m$ diagonal matrices. Note that $[X - PQ^TX]^i$ (i=1,2,…,m) means the i-th row of matrix $X - PQ^TX$, and q^i (i=1,2,…,m) means the i-th row of matrix Q. When $\left\|[X - PQ^TX]^i\right\|_2 = 0$, we let $D_1^{ii} = \frac{1}{2\left\|[X-PQ^TX]^i\right\|_2+\zeta}$ (ζ is a very small constant). Similarly, when $\left\|q^i\right\|_2 = 0$, we let $D_2^{ii} = \frac{1}{2\|q^i\|_2+\zeta}$. In this way, the smaller D_2^{ii} is, the more important the i-th feature is. Moreover, we can see that if $\left\|[X - PQ^TX]^i\right\|_2$ and $\left\|q^i\right\|_2$ are small, D_1 and D_2 are large and thus the minimization of $2tr((X - PQ^TX)^TD_1(X - PQ^TX)) + 2\lambda tr(Q^TD_2Q)$ (i.e., $\left\|X - PQ^TX\right\|_{2,1} + \lambda\|Q\|_{2,1}$) tends to force $\left\|[X - PQ^TX]^i\right\|_2$ and $\left\|q^i\right\|_2$ to be a very small value. After several iterations, some $\left\|[X - PQ^TX]^i\right\|_2$ and $\left\|q^i\right\|_2, (i = 1, 2, \ldots, m)$ may be close to zero, and thus we obtain a row-sparse Q and a row-sparse $X - PQ^TX$.

Next, let $\sqrt{D_1}P = \bar{P} \in \mathbb{R}^{m\times d}$, and $\sqrt{D_1}^{-1}Q = \bar{Q} \in \mathbb{R}^{m\times d}$. Then, the formulation in Eq. (18) can be rewritten as the following self-contained regression-type of weighted PCA,

$$\begin{aligned}&\arg\min_{\bar{Q},\bar{P}} \left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2 + \lambda\left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2 \\ &\qquad + \beta\left\|\bar{Q}\right\|_F^2, \\ &s.t.,\ \ \bar{P}^T\bar{P} = I.\end{aligned} \tag{20}$$

More specifically, Eq. (20) is the self-contained regression-type of weighted PCA (see Eq. (22)) in terms of weighted data $\sqrt{D_1}X$. From Eq. (26), we can see that the principal component of our method is $X^T\sqrt{D_1}\bar{P}$, $\bar{Q}$ is row-sparse and $\bar{Q}$ can be used to indicate those redundant features not participating in the principal component $X^T\sqrt{D_1}\bar{P}$. Since Eq. (20) is fully equivalent to Eq. (16), the proposed method is therefore named as joint sparse pixel weighted PCA. Once the optimal solution $\bar{P}$ and $\bar{Q}$ of Eq. (20) are obtained, the optimization solution P and Q of Eq. (16) are obtained.

Connection to the Self-Contained Regression-Type of Graph Embedding Framework: When $M_L = I$ and $M_D = \sqrt{D_1}X(\sqrt{D_1}X)^T$, according to the general self-contained regression-type framework (see Eq. (13)), we have,

$$\begin{aligned}&\arg\min_{\bar{Q},\bar{P}} \left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2 + \beta\left\|\bar{Q}\right\|_F^2, \\ &s.t.,\ \ \bar{P}^T\bar{P} = I,\end{aligned} \tag{21}$$

which is just the self-contained regression-type of pixel weighted PCA.

Eq. (20) is the proposed joint sparse pixel weighted PCA version. Assume that $\sqrt{D_1}$ is fixed, when $\bar{P} = \bar{Q}$, $\lambda \to 0$ and $\beta \to 0$, Eq. (20) becomes the weighted PCA as follows,

$$\begin{aligned}&\arg\min_{\bar{Q}} \left\|\sqrt{D_1}X - \bar{Q}\bar{Q}^T\sqrt{D_1}X\right\|_F^2, \\ &s.t.,\ \ \bar{Q}^T\bar{Q} = I,\end{aligned} \tag{22}$$

whose generalized eigenequation is listed as follows,

$$\sqrt{D_1}X(\sqrt{D_1}X)^T\bar{Q} = \alpha\bar{Q}. \tag{23}$$

Given the SVD of $\sqrt{D_1}X(\sqrt{D_1}X)^T = E\Lambda E^T$, we have $\sqrt{D_1}X(\sqrt{D_1}X)^TE = E\Lambda$. Therefore, the first d columns of E span the subspace Φ.

Similar to the self-contained regression-type of PCA, we argue that Eq. (21) is the self-contained regression-type of Eq. (22). Obviously, when $\lambda \to 0$, the obtained optimal subspace spanned by $\bar{Q} = E(\beta I + \Lambda)^{-1}\Lambda U^T$ is same with Φ. Once the self-contained regression-type is formed, arbitrary sparse terms can be added. When the sparse term $\lambda\left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2$ is added, Eq. (20) is produced.

B. The Optimization Solution

The solution of Eq. (20) is divided into the below two steps.

Step 1: Given $\bar{P}$, there exists a column-orthogonal matrix $\bar{P}_\perp$ such that $[\bar{P}, \bar{P}_\perp]$ is $m \times m$ orthogonal matrix. Then, optimization problem in Eq. (20) becomes,

$$\begin{aligned}\arg\min_{\bar{Q}} &\left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2 + \lambda\left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2 \\ &+\beta\left\|\bar{Q}\right\|_F^2.\end{aligned} \tag{24}$$

The first part of Eq. (24) can be rewritten as,

$$\begin{aligned}&\left\|\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X\right\|_F^2 \\ &= \left\|X^T\sqrt{D_1} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\right\|_F^2 \\ &= \left\|X^T\sqrt{D_1}[\bar{P}, \bar{P}_\perp] - X^T\sqrt{D_1}\bar{Q}\bar{P}^T[\bar{P}, \bar{P}_\perp]\right\|_F^2 \\ &= \left\|X^T\sqrt{D_1}\bar{P} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\bar{P}\right\|_F^2 \\ &\quad + \left\|X^T\sqrt{D_1}\bar{P}_\perp - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\bar{P}_\perp\right\|_F^2 \\ &= \left\|X^T\sqrt{D_1}\bar{P} - X^T\sqrt{D_1}\bar{Q}\right\|_F^2 + \left\|X^T\sqrt{D_1}\bar{P}_\perp\right\|_F^2.\end{aligned} \tag{25}$$

Since $\bar{P}$ is fixed, and $\left\|X^T\sqrt{D_1}\bar{P}_\perp\right\|_F^2$ is a constant, optimization problem in Eq(24) becomes the following optimization problem,

$$\begin{aligned}\arg\min_{\bar{Q}} &\left\|X^T\sqrt{D_1}\bar{P} - X^T\sqrt{D_1}\bar{Q}\right\|_F^2 + \lambda\left\|\sqrt{D_2}\sqrt{D_1}\bar{Q}\right\|_F^2 \\ &+\beta\left\|\bar{Q}\right\|_F^2.\end{aligned} \tag{26}$$

By the derivative of Eq. (26) with respect to $\bar{Q}$ to be 0, we get,

$$\bar{Q} = (\lambda\sqrt{D_1}D_2\sqrt{D_1} + \beta I + \sqrt{D_1}XX^T\sqrt{D_1})^{-1}\sqrt{D_1}XX^T\sqrt{D_1}\bar{P}. \tag{27}$$

Hence,

$$Q = (\lambda D_2 + \beta D_1^{-1} + XX^T)^{-1} XX^T \sqrt{D_1}\bar{P}. \quad (28)$$

When $\lambda \to 0$, Eq. (27) becomes as follows,

$$\bar{Q} = (\beta I + \sqrt{D_1}XX^T\sqrt{D_1})^{-1}\sqrt{D_1}XX^T\sqrt{D_1}\bar{P}. \quad (29)$$

Step 2: Given $\bar{Q}$ to compute $\bar{P}$, optimization problem in Eq. (20) becomes,

$$\arg\min_{\bar{P}} \left\| \sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X \right\|_F^2, \quad s.t. \bar{P}^T\bar{P} = I. \quad (30)$$

The first part of Eq(30) can be rewritten as,

$$\begin{aligned}
&\left\| \sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X \right\|_F^2 \\
&= tr((\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X)^T(\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X)) \\
&= tr((X^T\sqrt{D_1} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T)(\sqrt{D_1}X - \bar{P}\bar{Q}^T\sqrt{D_1}X)) \\
&= tr(X^T D_1 X - X^T\sqrt{D_1}\bar{P}\bar{Q}^T\sqrt{D_1}X \\
&\quad - X^T\sqrt{D_1}\bar{Q}\bar{P}^T\sqrt{D_1}X + X^T\sqrt{D_1}\bar{Q}\bar{P}^T\bar{P}\bar{Q}^T\sqrt{D_1}X) \\
&= tr(X^T D_1 X + X^T\sqrt{D_1}\bar{Q}\bar{Q}^T\sqrt{D_1}X) \\
&\quad - 2tr(\bar{Q}^T\sqrt{D_1}XX^T\sqrt{D_1}\bar{P}). \quad (31)
\end{aligned}$$

Since $\bar{Q}$ is given, Eq. (30) becomes,

$$\arg\max_{\bar{P}} tr(\bar{Q}^T\sqrt{D_1}XX^T\sqrt{D_1}\bar{P}), \quad s.t. \bar{P}^T\bar{P} = I. \quad (32)$$

Substituting Eq. (29) into Eq. (32), we have,

$$\begin{aligned}
&\arg\max_{\bar{P}} tr(\bar{P}^T\sqrt{D_1}XX^T\sqrt{D_1}(\beta I + \sqrt{D_1}XX^T\sqrt{D_1})^{-T} \\
&\qquad \sqrt{D_1}XX^T\sqrt{D_1}\bar{P}), \\
&s.t. \ \bar{P}^T\bar{P} = I. \quad (33)
\end{aligned}$$

Given the SVD of $\sqrt{D_1}XX^T\sqrt{D_1} = E\Lambda E^T$. We can conclude that the columns of E are the eigenvectors of matrix $\sqrt{D_1}XX^T\sqrt{D_1}(\beta I + \sqrt{D_1}XX^T\sqrt{D_1})^{-T}\sqrt{D_1}XX^T\sqrt{D_1}$.

On the other hand, optimization problem in Eq. (30) is equal to,

$$\arg\min_{\bar{P}} \left\| X^T\sqrt{D_1} - X^T\sqrt{D_1}\bar{Q}\bar{P}^T \right\|_F^2, \quad s.t. \bar{P}^T\bar{P} = I. \quad (34)$$

The update of $\bar{P}$ of minimizing Eq. (27) with the constraint of $\bar{P}^T\bar{P} = I \in \mathbb{R}^{d\times d}$ means that $\bar{P}$ is column-orthogonal. In order to compute $\bar{P}$, we introduce the following [32, Lemma 1].

Lemma 1: Let $Z^{n\times m}$ and $V^{n\times d}$ be two matrices. Consider the constrained minimization problem,

$$\arg\min_{P} \left\| Z - VM^T \right\|^2, \quad s.t. \ M^TM = I. \quad (35)$$

Suppose the SVD of Z^TV is EDU^T, then the optimization solution is $M = EU^T \in \mathbb{R}^{m\times d}$.

According to Lemma 1, we have $Z^TV = \sqrt{D_1}XX^T\sqrt{D_1}\bar{Q}$. Let the SVD of $\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = EDU^T$, then the optimal $\bar{P}$ can also be directly obtained from SVD of $\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = EDU^T$, i.e.,

$$\bar{P} = EU^T. \quad (36)$$

Thus,

$$P = \sqrt{D_1}^{-1}EU^T. \quad (37)$$

Substituting $\bar{P} = EU^T$ into Eq. (29), we have,

$$\begin{aligned}
\bar{Q} &= (\beta I + \sqrt{D_1}XX^T\sqrt{D_1})^{-1}\sqrt{D_1}XX^T\sqrt{D_1}EU^T \\
&= (\beta I + E\Lambda E^T)^{-1}E\Lambda E^TEU^T \\
&= (\beta I + E\Lambda E^T)^{-1}E\Lambda U^T \\
&= E(\beta I + \Lambda)^{-1}\Lambda U^T. \quad (38)
\end{aligned}$$

Algorithm 1 The Proposed Algorithm

Input: Training sample set X, parameter λ, dimensionality d.

1: Initialize D_1, D_2 as $I \in \mathbb{R}^{m\times m}$ and random $\bar{P} \in \mathbb{R}^{m\times d}$.
2: **while** not converge **do**
2.1: Compute $\bar{Q}$ according to Eq. (27)
2.2: Compute Q according to Eq. (28)
2.3: Compute $\bar{P}$ according to Eq. (36)
2.4: Compute P according to Eq. (37)
2.5: Compute D_1 according to Eq. (17)
2.6: Compute D_2 according to Eq. (19)
end while

Output: Projection matrix Q.

C. Computational Complexity Analysis

The main computational complexity of joint sparse pixel weighted PCA has two steps in each iteration: the first step is to compute $Q = (\lambda D_2 + XX^T)^{-1}XX^T\sqrt{D_1}\bar{P}$ with a computational complexity $O(m^3)$; the second step is to compute SVD of $\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = EDU^T$, whose computational complexity is also $O(m^3)$ at most. Therefore, the computational complexity of one iteration does not exceed $O(m^3)$. If this algorithm needs t iterations, the total computational complexity is $O(tm^3)$.

IV. Experiments

To evaluate the proposed method, we compare it with PCA [37], PCAL1 [52], OMRPCA [22], and SPCA [32]; these are selected as PCA plays an important role in reconstruction, PCAL1 maximizes the L1 dispersion (i.e., using L1-norm in the feature space) to improve the robustness of PCA to outliers. R1PCA and OMRPCA are two weighted PCA versions by using weighting of the data samples, and SPCA is the first method to propose the self-contained regression-type of PCA. It is shown here that the proposed method not only achieves a good reconstruction result but also offers a better interpretation of the principal components.

The codes of all the comparison methods are downloaded from the author's Web sites. All the methods have a common parameter d, i.e., the reduced dimensionality. In all the methods, d is of the same value. Besides, the other parameters are searched from their papers or coded by us. All codes are implemented using MATLAB on a computer with a 3.30-GHz duo core CPU and 8-GB memory.

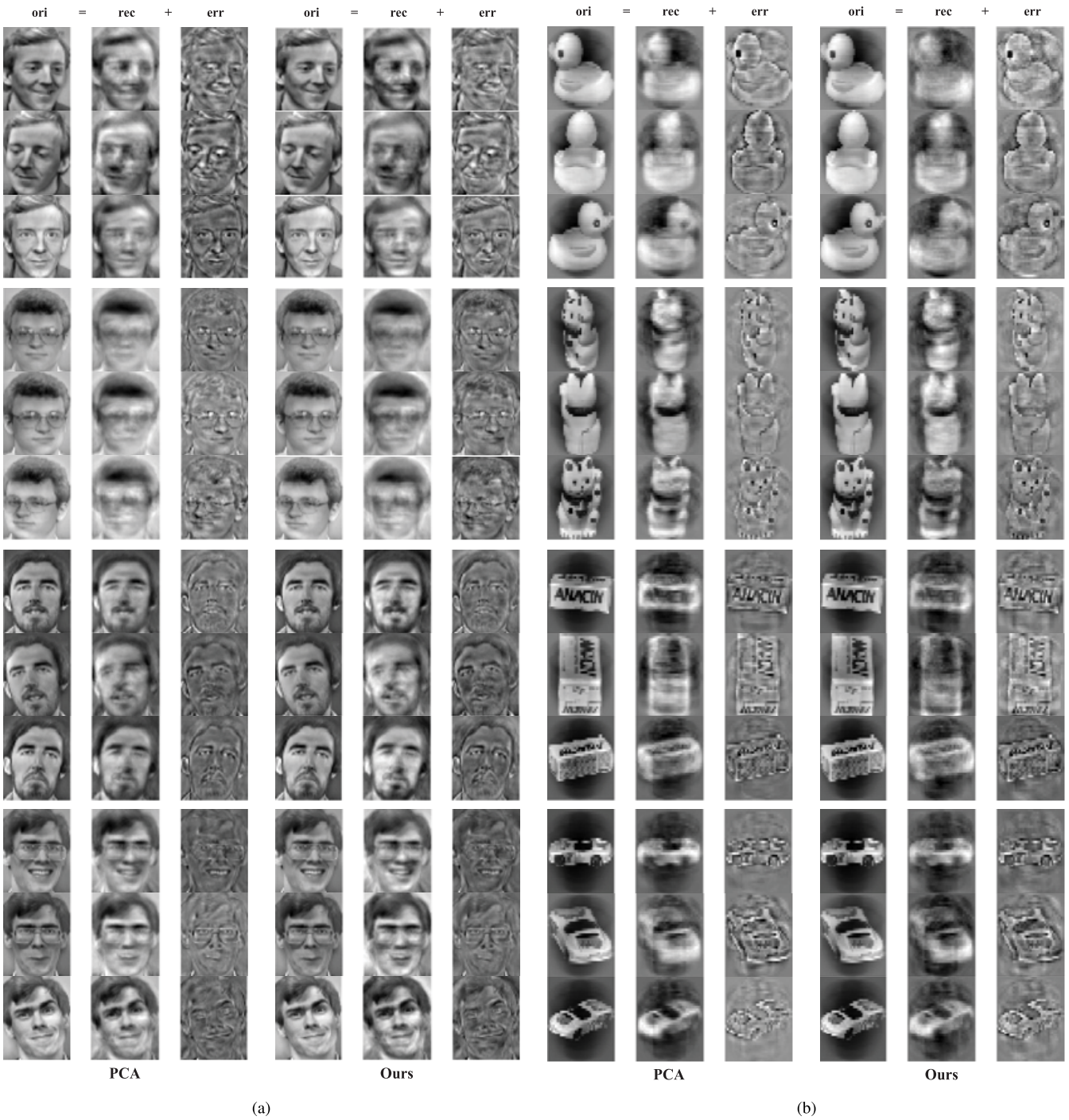

Fig. 1. Visualization of the comparison of reconstruction results in terms of PCA and the proposed method on the ORL and COIL20 datasets. This figure includes four groups and each group includes three columns, which corresponds to the original image, reconstruction image and the error image, respectively. The first two groups are implemented on the ORL dataset, and while the last two groups are implemented on the COIL20 dataset.

A. Datasets

Two Facial Datasets: The ORL face dataset [38] contains 40 individuals, and each of which has 10 face images. Here, every image is resized to 56×46 pixels. The FERET face dataset [39] contains 1400 images from 200 individuals, each of which has seven images. Every image is resized to 40×40 pixels.

Two Object Datasets: The COIL20 image dataset [40] contains 20 subjects. Each subject contains 72 images, and each image is taken at pose intervals of 5^o. Here, each image is converted to a gray-scale image of size 36×37 pixels. The COIL100 image dataset [41] contains 100 subjects, where each subject contains 72 images and each image is taken at pose intervals of 5^o. Here, each image is converted into a gray-scale image of size 32×32 pixels.

A Digital Dataset: The USPS dataset [42] contains 9298 digit images in total from 0 to 9, each of which is of size 16×16 pixels, with 256 gray levels per pixel.

TABLE I

RECONSTRUCTION ERROR COMPARISONS OF SIX PCA METHODS ON SIX DATASETS USING DIFFERENT DIMENSIONS, WHERE D MEANS THE DIMENSION

Datasets	ORL($\times 10^1$)						FERET($\times 10^2$)						COIL20($\times 10^2$)					
Algorithms	PCA	PCAL1	R1PCA	OMRPCA	SPCA	Ours	PCA	PCAL1	R1PCA	OMRPCA	SPCA	Ours	PCA	PCAL1	R1PCA	OMRPCA	SPCA	Ours
10 D	7.4993	7.6535	7.5056	7.4929	7.9137	7.6520	2.5290	2.5425	2.5831	2.5250	2.5495	2.7420	4.9337	4.9637	4.8957	4.8054	5.0882	5.0433
20 D	6.3987	6.6204	6.4868	6.3879	6.8203	6.6098	2.1493	2.1723	2.1677	2.1437	2.1495	2.3985	4.0646	4.1206	4.0217	3.9831	4.3421	4.2727
30 D	5.7198	5.9873	5.7771	5.7079	5.9962	5.9850	1.9261	1.9598	1.9325	1.9216	1.9264	2.2210	3.6054	3.6917	3.5825	3.5539	3.9619	3.8822
40 D	5.2082	5.5039	5.2526	5.1955	5.5412	5.5078	1.7691	1.8128	1.7755	1.7648	1.7697	2.0836	3.3054	3.4130	3.2866	3.2621	3.5396	3.6092
50 D	4.8054	5.1305	4.8472	4.7925	5.1122	5.1272	1.6500	1.7033	1.6565	1.6452	1.6510	1.9944	3.0761	3.1983	3.0597	3.0401	3.1277	3.3992
60 D	4.4762	4.8332	4.5119	4.4621	4.7668	4.8293	1.5505	1.6094	1.5566	1.5456	1.5520	1.9199	2.8996	3.0255	2.8824	2.8639	3.0884	3.2512
Datasets	COIL100($\times 10^3$)						USPS($\times 10^3$)						LUNG($\times 10^1$)					
Algorithms	PCA	PCAL1	R1PCA	OMRPCA	SPCA	Ours	PCA	PCAL1	R1PCA	OMRPCA	SPCA	Ours	PCA	PCAL1	R1PCA	OMRPCA	SPCA	Ours
10 D	2.1308	2.1625	2.1580	2.1244	2.6702	2.3413	2.9706	2.9876	3.0631	2.9362	3.6936	3.6487	1.7966	1.8327	1.8366	1.7948	1.8078	1.9045
20 D	1.8254	1.8521	1.8283	1.8161	2.3791	2.2758	1.9887	2.0058	2.0240	1.9671	2.9018	2.8833	1.5878	1.6571	1.6101	1.5851	1.5878	1.7200
30 D	1.6392	1.6663	1.6390	1.6312	2.3550	2.1998	1.4427	1.4692	1.4556	1.4266	2.4241	2.3873	1.4554	1.5399	1.4796	1.4516	1.4554	1.5943
40 D	1.5178	1.5497	1.5154	1.5091	2.1148	2.0534	1.0720	1.1059	1.0780	1.0561	2.1133	2.0692	1.3460	1.4521	1.3698	1.3303	1.3460	1.5088
50 D	1.4261	1.4565	1.4221	1.4183	2.1015	1.9504	0.8194	0.8514	0.8207	0.8086	1.5912	1.8549	1.2490	1.3711	1.2783	1.2216	1.2490	1.4280
60 D	1.3536	1.3867	1.3489	1.3457	2.0206	1.8787	0.6387	0.6698	0.6367	0.6280	1.0705	1.8373	1.1607	1.2865	1.1918	1.1157	1.1607	1.3719

A Gene Dataset: The LUNG dataset [34] contains 203 gene samples in total with five classes, which contain 139, 21, 20, 6, 17 samples, respectively. Every gene sample has 3312 genes.

Note that all the above datasets are centered and normalized in our experiments.

B. Experimental Results

1) Reconstruction: For these six datasets, the reconstruction errors of PCA, PCAL1, R1PCA, OMRPCA, SPCA and the proposed method are carried out, and the experimental results are listed in Table 1.

From Table 1, it can be seen that PCA, PCAL1, R1PCA and OMRPCA often perform favorably against SPCA and the proposed method. This is due to the addition of the sparse regularization term, which means that SPCA and the proposed method inevitably suffer from some loss of information, while PCA, PCAL1, R1PCA and OMRPCA do not suffer from this drawback. Furthermore, the proposed method often performs favorably against SPCA. This may because the reconstruction error of the proposed method arises from $\ell_{2,1}$-norm, while the reconstruction error of the SPCA method arises from ℓ_2-norm.

Recall Table 1, the reconstruction error is computed according to the original images before corruption and the reconstruction images of corrupted images. However, the original images itself often include some inevitable corruption such as glasses or illumination and several other noise. At this moment, it is invalid using the criteria of the reconstruction error for evaluating reconstruction methods. This is also the reason why the proposed method does not achieve better results in Table 1. To this end, we visualize some reconstruction images for both ORL and COIL20 datasets in Fig. 1. It can be seen from Fig. 1 that although the proposed method has the worst result from Table 1, it obtains the reconstruction results similar to those of PCA. Furthermore, we compare different methods on toy dataset (see Fig. 2). The toy dataset contains 180 data points (colored in black) near a straight line and 20 data points (colored in orange) far away this straight line (i.e., the outliers). Every data point has two features, and the first feature (i.e., x feature) is significantly important while the second feature (i.e., y feature) is redundant. It can be seen from Fig. 2 that the proposed method, OMRPCA, and SPCA perform better than PCA and PCAL1. As we known, PCA (in an L2 sense) is sensitive to outliers. PCAL1 (in an L1 sense) performs better than PCA but is still sensitive to outliers. OMRPCA and SPCA have similar results with the proposed method on the toy dataset. This is because OMRPCA is robust to outliers and SPCA can achieve a balance between the data representation and the sparsity. So when the data variation mainly lies in one dimension, SPCA and the proposed method are able to achieve the best performance. Moreover, we compute the reconstruction error of different methods on the toy data. The reconstruction errors of PCA, PCAL1, OMRPCA, SPCA, and the proposed method are 2735.8, 20122.0, 720.2, 551.8, and 551.8, respectively. This shows that the reconstruction error of the proposed method is the smallest, and hence the proposed method is effective.

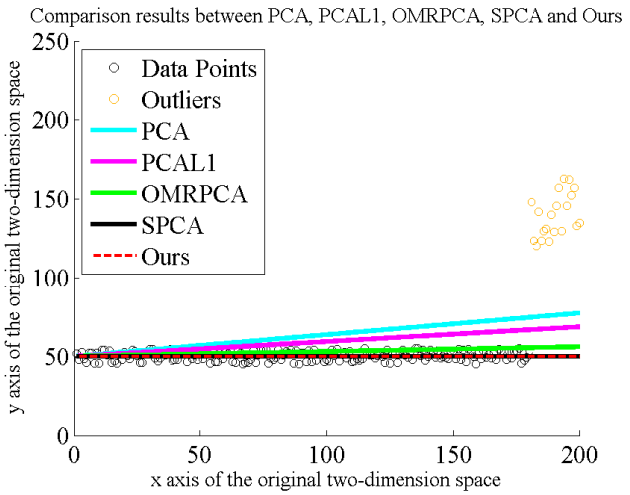

Fig. 2. The principal component axis learned by PCA, PCAL1, OMRPCA, SPCA and the proposed method.

Besides, we applied the proposed method into background-foreground separation experiments. The used dataset, gathered from a static camera over one day, contains 502 images and each image is of size 30×40 pixels. The experimental results are shown in Fig. 3. The first row of Fig. 3 shows

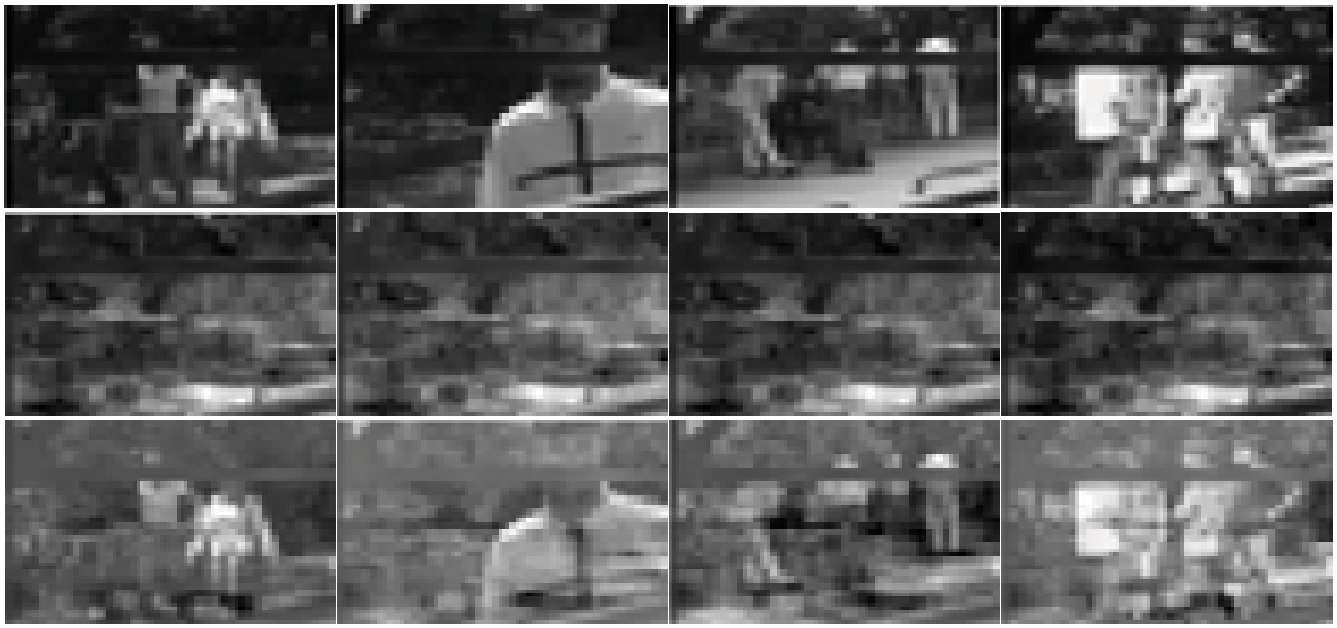

Fig. 3. Visualization of background-foreground separation experiments, where the first row is the original images, the second row is the backgrounds and the third row is the foregrounds.

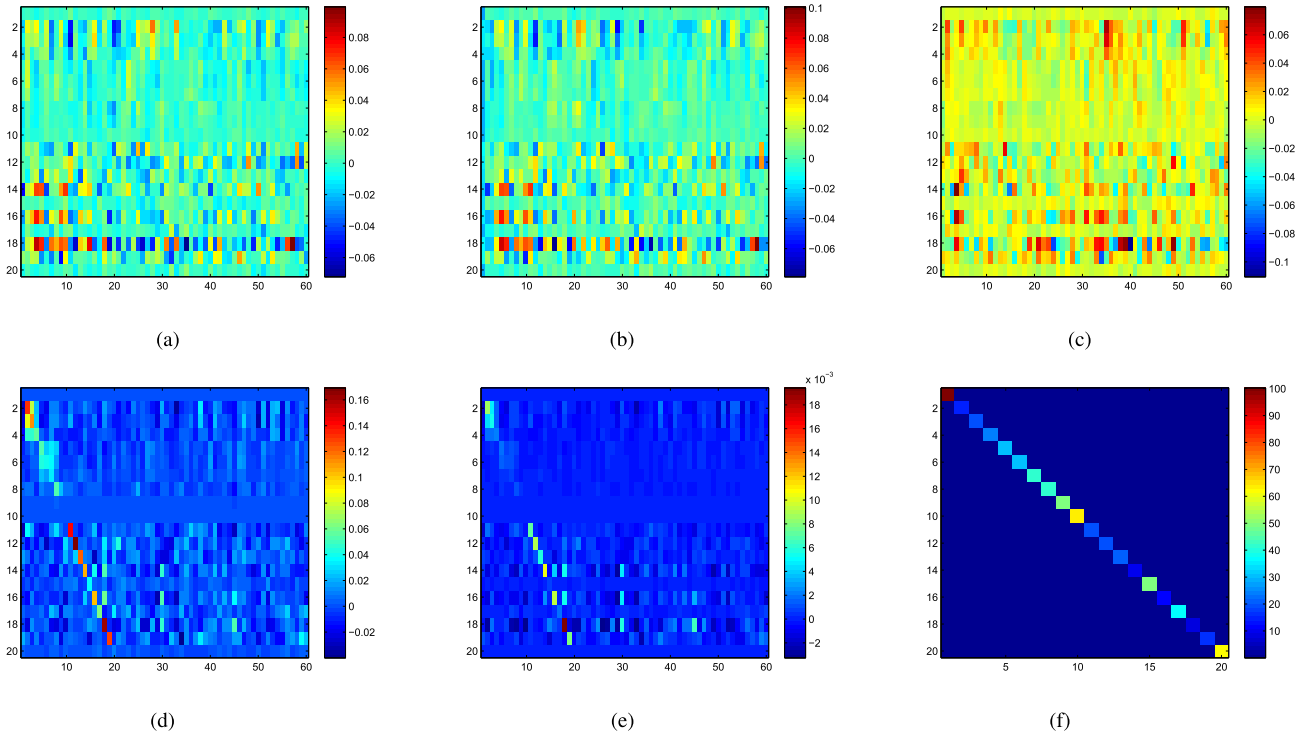

Fig. 4. Visualization of the projection matrix Q. They are obtained by different methods on the LUNG dataset. The first row shows the projection matrix obtained by PCA, SPCA and OMRPCA. The second row shows the projection matrix Q, $\bar{Q}$ and the automatic weight obtained by the proposed method.

the original images with illumination changes of the static background and peoples in various locations. While the peoples often pass through the view of the camera quickly, they sometimes remain relatively still over multiple frames. Here, the peoples (i.e., foreground) can be regarded as outliers and the view of the static camera with illumination changes (i.e., background) can be reconstructed. The second row of Fig. 3 shows the reconstruction result of the proposed method. The outliers (i.e., the peoples) can be clearly seen from the third row of Fig. 3.

2) Interpretation: In general, both PCA and OMRPCA lack a clear interpretation for principal components since the principal components obtained by PCA and OMRPCA are a linear combination of all the original features. SPCA has been therefore developed to interpret the principal components. However, it still can not consistently

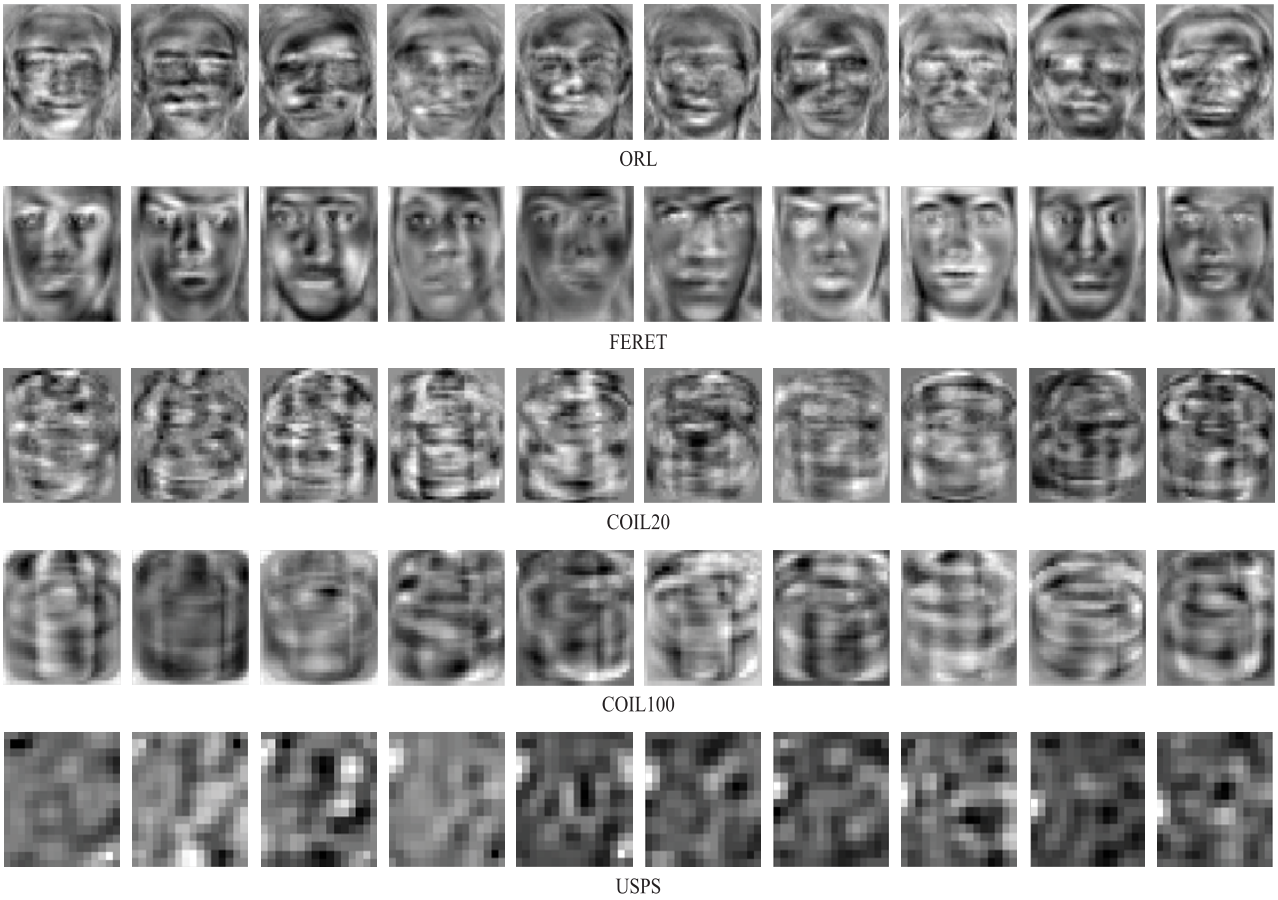

Fig. 5. Basis images of the proposed method on five datasets.

identify features participating in the computation of principal components.

In order to test the ability to interpret the principal components of the proposed method, the first 20 features of the LUNG dataset are used to test PCA, SPCA, OMRPCA and the proposed method. The experimental results are shown in Fig. 4, where it can be seen that the proposed method indicates that it consistently uses all the features except for the first, ninth, tenth and twentieth features to participate in the computation of the principal components, while the other methods do not have this function. Therefore, by comparing the projection matrix Q of the proposed method with the other methods, the proposed method can find the redundant features that contribute less to the principal components while the other methods can not. Moreover, from Fig. 4(f), it can be seen that the redundant features are assigned a larger weight while the remaining feature are assigned a smaller weight. This is because the redundant features have a good reconstruction and the weight D_1 is automatically decided by the reconstruction case. From Fig. 4(d) and Fig. 4(e), we can see that the obtained $\bar{Q}$ is more row-sparse than Q if Q is row-sparse.

Furthermore, the bases images used in our method are shown in Fig. 5, where the bright area indicated the selected principal features. From the ORL and FERET facial datasets in Fig. 5, it can be seen that the bright area is mainly focused on the principal features, such as the eyes, nose, mouth, and contour of each face. From the COIL20, COIL100 and USPS datasets in Fig. 5, it can be seen that the bright area mainly focuses on the contours of an object or digit.

3) Parameter Settings and Convergence Analyses: There are two parameters in our objective function, λ and β. To demonstrate the effects of these two parameters in experiments, different combinations of these values, selected from a reasonable discrete set $\{1e^{-6}, 1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1e^{0}, 1e^{1}\}$, are evaluated on each dataset, where each parameter combination corresponds to a reconstruction error. When the reconstruction error is minimized, the optimal parameters are obtained. In order to find the smallest reconstruction error easily, we take opposite of the obtained reconstruction error. The reconstruction error of each parameter combination is shown in Fig. 6, from where it can be seen that the reconstruction performance is almost same over a wide range of parameters. This shows that the reconstruction performance of the proposed method is very robust to the parameters. The selection of a suitable parameter combination is therefore straightforward. In this paper, we use $\lambda = 10^{-4}$ and $\beta = 10^{-2}$ for the ORL dataset, $\lambda = 10^{-3}$ and $\beta = 10^{-2}$ for the FERET dataset, $\lambda = 10^{-3}$ and $\beta = 10^{-2}$ for the COIL20 dataset, $\lambda = 10^{-3}$ and $\beta = 10^{-6}$ for the COIL100 dataset, $\lambda = 10^{0}$ and $\beta = 10^{-2}$ for the USPS dataset, and $\lambda = 10^{-4}$ and $\beta = 10^{-2}$ for the LUNG dataset.

The convergence curves of the proposed method are visualized in Fig. 7. As can be observed, the proposed method achieves a fast convergence (within five iterations) for every

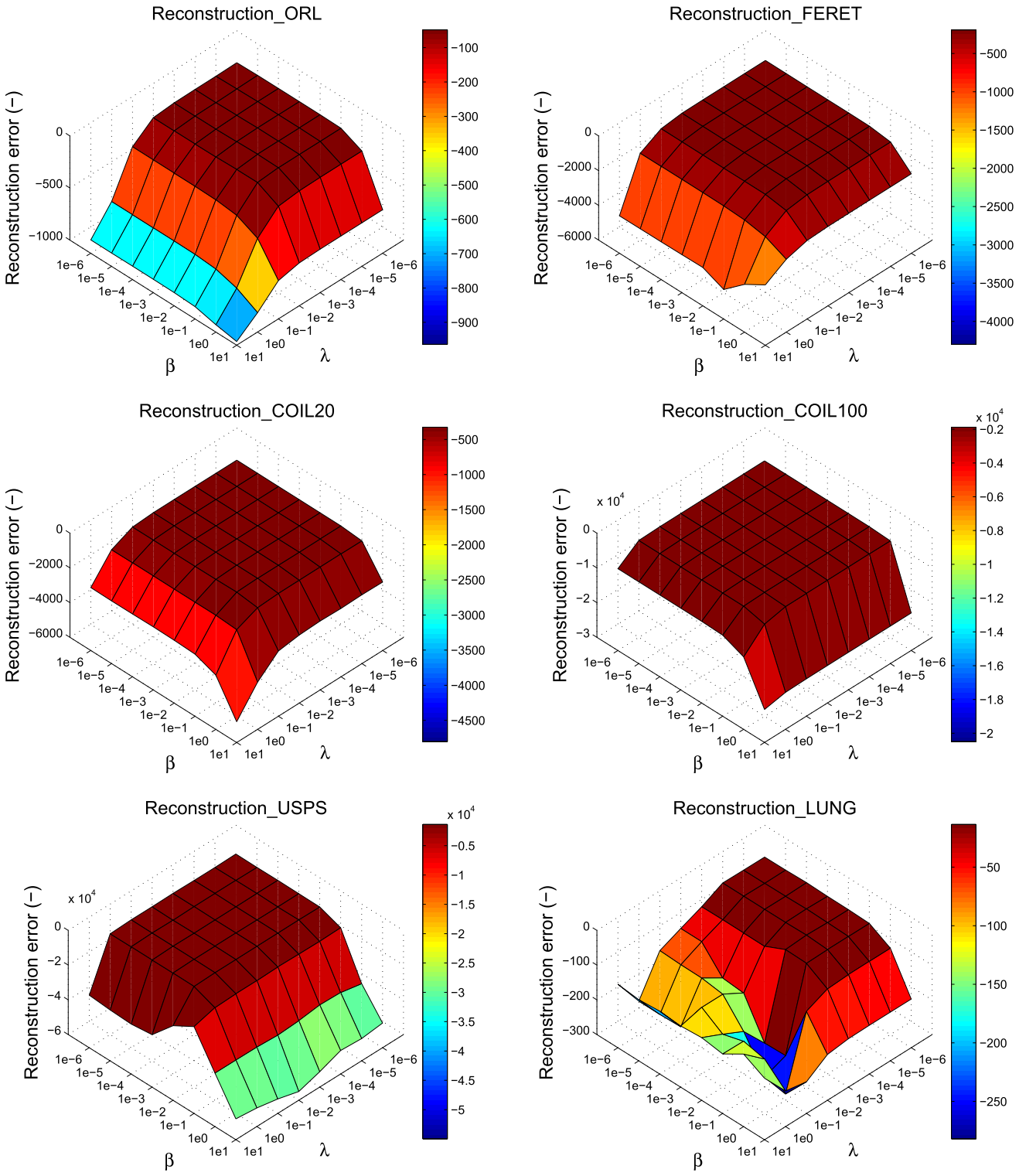

Fig. 6. Parameter settings of the proposed method on the ORL, FERET, COIL20, COIL100, USPS, and LUNG datasets.

dataset. Such a fast convergence is mainly attributed to the process of solving the optimization variable Q. During solving Q, some rows of Q are forced to approximate to zero and hence only some features, but not all, are selected to participate in the reconstruction of data. As a result, the obtained Q is able to effectively select some useful features to make the reconstruction error as small as possible. Thus, P and Q can fast approximate to their optimal solutions. Furthermore, we compare the proposed method with R1PCA, OMRPCA, and SPCA on the ORL dataset. All of them are under the same termination condition, i.e., $|err^{(t)} - err^{(t-1)}| < \varepsilon$, where $err = \left\| X - QQ^T X \right\|_F$, $err^{(t)}$ is the reconstruction error of

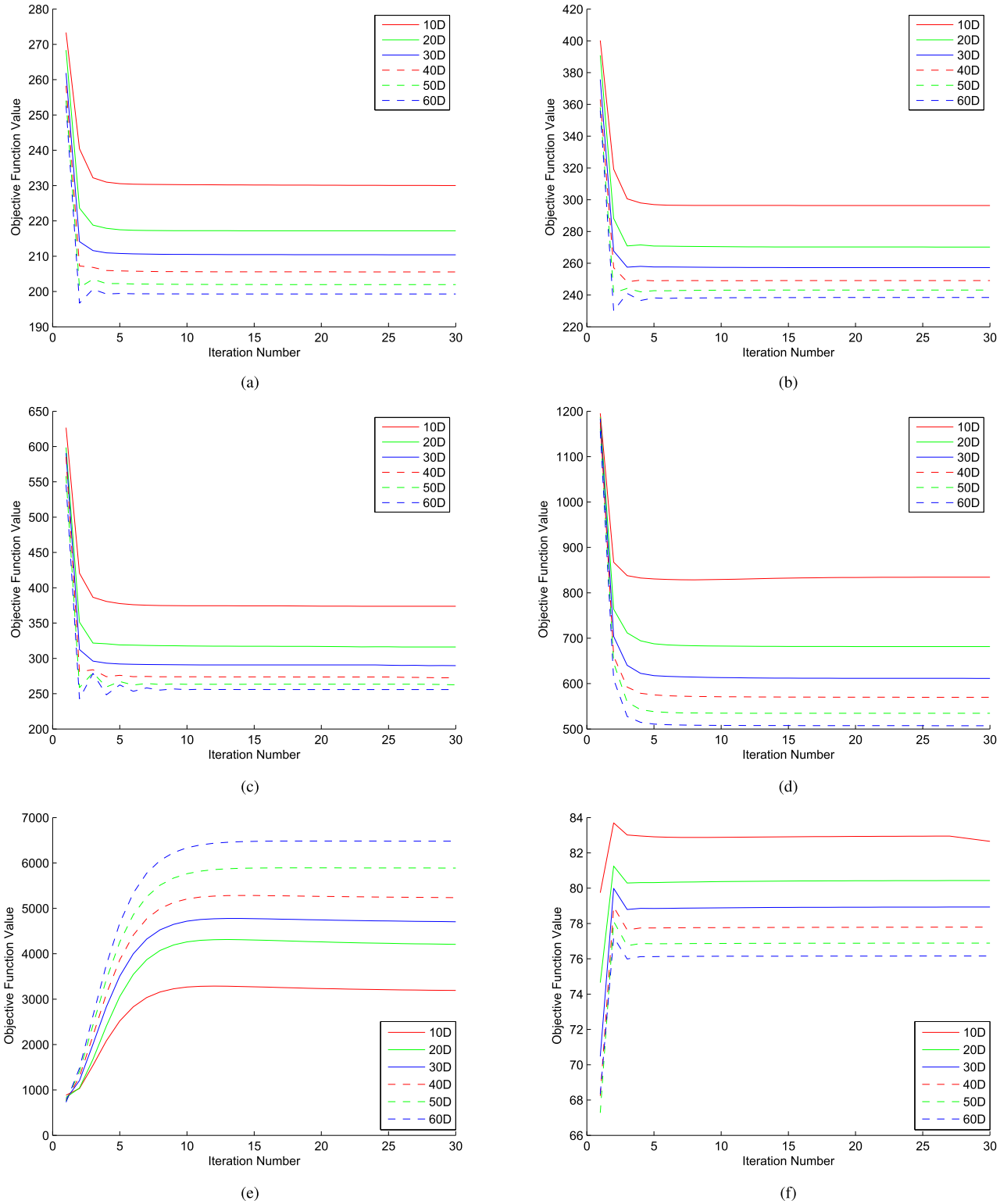

Fig. 7. Convergence curves of the proposed method on the ORL, FERET, COIL20, COIL100, USPS, and LUNG datasets. On each dataset, Convergence curves under six different subspace dimensions are shown.

the t-th iteration and $\varepsilon = 10^{-6}$. These convergence curves are shown in Fig. 8.

It is worth noting that on some datasets, such as the LUNG dataset, the objective function value shows a strong vibration. This phenomenon can be interpreted as the consequence of the inexact solution of Eq. (28), that is, the exact solution is permutated a little in our method by adding the Tikhonov regularization ηI to the inverse of the matrix $\lambda D_2 + \beta D_1^{-1} + XX^T$. In this paper, $\eta = 0.001$ is used. In fact, the larger η is, the stronger the oscillation is. However, we eventually observe that the objective function value decreases steadily as the number of iterations continues to increase.

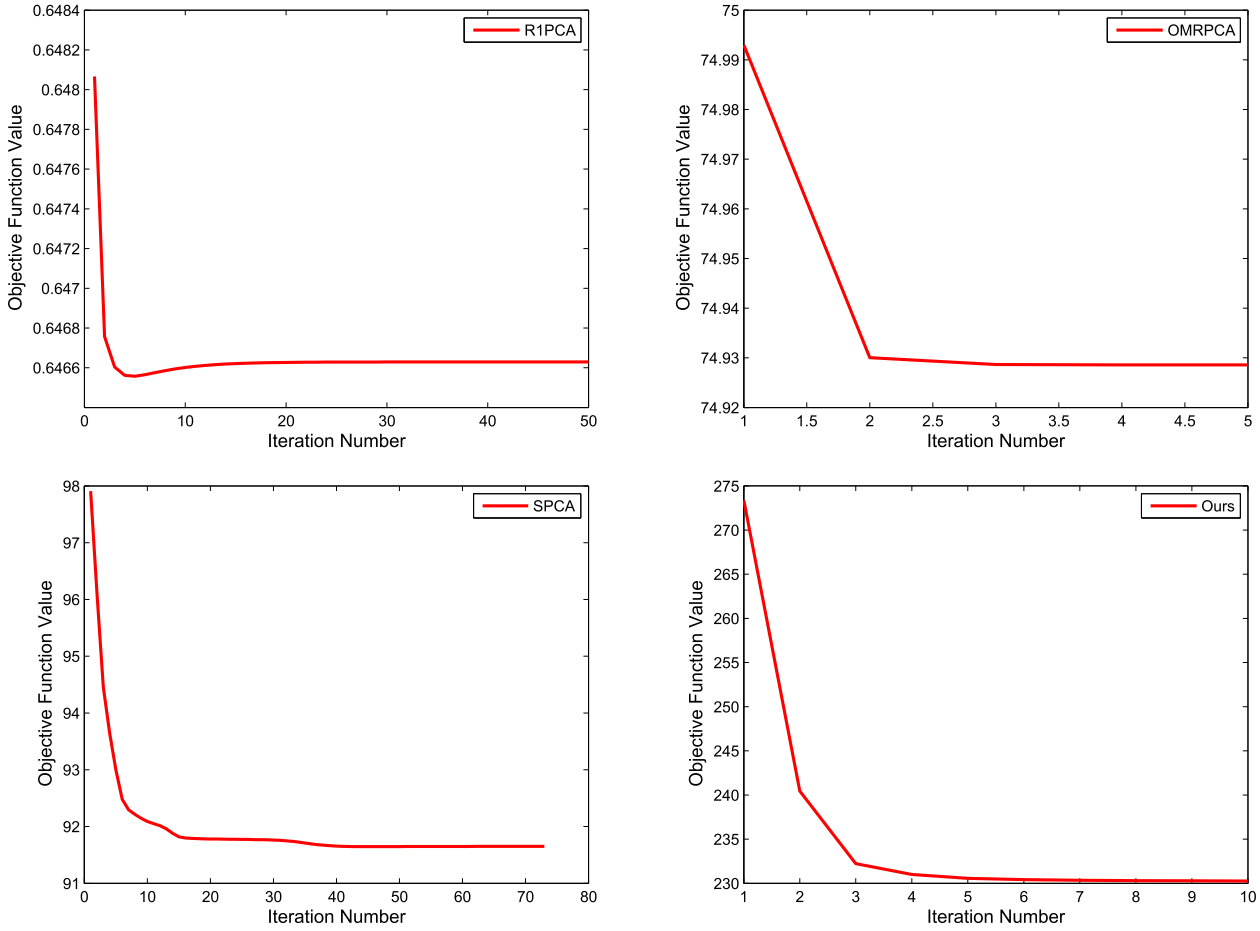

Fig. 8. Convergence curves of four methods on the ORL dataset.

V. Conclusion

In this paper, we propose a simple but effective PCA version. Unlike PCA, the proposed method can clearly identify the redundant features that do not participate in reconstruction, although the proposed method only achieves the reconstruction result similar to those of the other PCA methods. Moreover, this paper mainly provides some theoretical analysis in order to give a thoroughly understanding of the essence of principal components and self-contained regression-type.

References

[1] Q. Peng, Y.-M. Cheung, X. You, and Y. Y. Tang, "A hybrid of local and global saliencies for detecting image salient region and appearance," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 1, pp. 86–97, Jan. 2017.

[2] X.-Y. Jing *et al.*, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1363–1378, Mar. 2017.

[3] Q. Ge *et al.*, "Structure-based low-rank model with graph nuclear norm regularization for noise removal," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3098–3112, Jul. 2017, doi: 10.1109/TIP.2016.2639781.

[4] X.-Y. Jing, F. Wu, X. Zhu, X. Dong, F. Ma, and Z. Li, "Multi-spectral low-rank structured dictionary learning for face recognition," *Pattern Recognit.*, vol. 59, pp. 14–25, Nov. 2016.

[5] F. Wu, X.-Y. Jing, X. You, D. Yue, R. Hu, and J.-Y. Yang, "Multi-view low-rank dictionary learning for image classification," *Pattern Recognit.*, vol. 50, pp. 143–154, Feb. 2016.

[6] W.-S. Chen, Y. Zhao, B. Pan, and B. Chen, "Supervised kernel nonnegative matrix factorization for face recognition," *Neurocomputing*, vol. 205, pp. 165–181, Sep. 2016.

[7] W.-S. Chen, X. Dai, B. Pan, and Y. Y. Tang, "Semi-supervised discriminant analysis method for face recognition," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 13, no. 6, pp. 1550049–1550071, 2015.

[8] L. Chen, L. Liu, and C. L. P. Chen, "A robust bi-sparsity model with non-local regularization for mixed noise reduction," *Inf. Sci.*, vol. 354, no. 1, pp. 101–111, Aug. 2016.

[9] L. Liu, L. Chen, C. L. P. Chen, Y. Y. Tang, and C. M. Pun, "Weighted joint sparse representation for removing mixed noise in image," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 600–611, Mar. 2017.

[10] W.-S. Chen, P. C. Yuen, J. Huang, and B. Fang, "Two-step single parameter regularization FISHER discriminant method for face recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 20, no. 2, pp. 189–207, 2006.

[11] S. Chakraborty, S. Singh, and P. Chakraborty, "Local gradient hexa pattern: A descriptor for face recognition and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[12] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. San Francisco, CA, USA: Academic, 1980.

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

[14] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[15] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1208–1213.

[16] A. J. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*. New York, NY, USA: Springer, 2013, pp. 237–280.

[17] 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, Dec. 2000.

[18] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Neural Inf. Process. Syst.*, 2001, pp. 585–591.

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

[20] 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.

[21] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 281–288.

[22] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1062–1070.

[23] Z. Fan, E. Liu, and B. Xu, "Weighted principal component analysis," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, 2011, pp. 569–574.

[24] F. de la Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 362–369.

[25] F. De la Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, no. 1, pp. 117–142, Aug. 2003.

[26] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[27] X. Shi, Z. Guo, F. Nie, L. Yang, J. You, and D. Tao, "Two-dimensional whitening reconstruction for enhancing robustness of principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2130–2136, Oct. 2016.

[28] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-rank preserving projections," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1900–1913, Aug. 2016.

[29] Z. Lai, Y. Xu, Z. Jin, and D. Zhang, "Human gait recognition via sparse discriminant projection learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1651–1662, Oct. 2014.

[30] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1294–1299.

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

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

[33] K. Sjöstrand, "Matlab implementation of LASSO, LARS, the elastic net and SPCA," Dept. Inf. Math. Model., Tech. Univ. Denmark, Copenhagen, Denmark, Tech. Rep. Version 2.0, 2005.

[34] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2,1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[35] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "ℓ2, 1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.

[36] S. Yi, Z. Lai, Z. He, Y.-M. Cheung, and Y. Liu, "Joint sparse principal component analysis," *Pattern Recognit.*, vol. 61, pp. 524–536, Jan. 2017.

[37] I. Jolliffe, *Principal Component Analysis and Factor Analysis*. New York, NY, USA: Springer, 2002.

[38] X. Fang, Y. Xu, X. Li, Z. Fan, H. Liu, and Y. Chen, "Locality and similarity preserving embedding for feature selection," *Neurocomputing*, vol. 128, pp. 304–315, Mar. 2014.

[39] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[40] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[41] S. K. Nayar, S. A. Nene, and H. Murase, "Columbia object image library (COIL-100)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

[42] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

[43] Z. Chen, X. You, B. Zhong, J. Li, and D. Tao, "Dynamically modulated mask sparse tracking," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2577718.

[44] R. Z. Liu, Y. Y. Tang, and B. Fang, "Topological coding and its application in the refinement of SIFT," *IEEE Trans. Cybern.*, vol. 44, no. 11, pp. 2155–2166, Nov. 2014.

[45] J. Qian, B. Fang, W. Yang, X. Luan, and H. Nan, "Accurate tilt sensing with linear model," *IEEE Sensors J.*, vol. 11, no. 10, pp. 2301–2309, Oct. 2011.

[46] X. Li, Q. Liu, Z. He, H. Wang, C. Zhang, and W.-S. Chen, "A multi-view model for visual tracking via correlation filters," *Knowl.-Based Syst.*, vol. 113, pp. 88–99, Dec. 2016.

[47] X. Ma, Q. Liu, Z. He, X. Zhang, and W.-S. Chen, "Visual tracking via exemplar regression model," *Knowl.-Based Syst.*, vol. 106, pp. 26–37, Aug. 2016.

[48] Z. He, X. Li, D. Tao, X. You, and Y. Y. Tang, "Connected component model for multi-object tracking," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3698–3711, Aug. 2016.

[49] S. Yi, Z. He, Y.-M. Cheung, X. You, and Y. Y. Tang, "Robust object tracking via key patch sparse representation," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 354–364, Feb. 2017.

[50] Z. He, X. You, L. Zhou, Y.-M. Cheung, and J. Du, "Writer identification using fractal dimension of wavelet subbands in Gabor domain," *Integr. Comput.-Aided Eng.*, vol. 17, no. 2, pp. 157–165, 2010.

[51] Z. Lai, W. K. Wong, Y. Xu, J. Yang, and D. Zhang, "Approximate orthogonal sparse embedding for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 723–735, Apr. 2016.

[52] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[53] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, "Transfer learning with graph co-regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1805–1818, Jul. 2014.

[54] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2760–2771, Sep. 2015.

[55] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1302–1308.

[56] S. Zhang, X. Lan, Y. Qi, and P. C. Yuen, "Robust visual tracking via basis matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 421–430, Mar. 2017.

[57] J. Zhang, M. Wang, S. Zhang, X. Li, and X. Wu, "Spatiochromatic context modeling for color saliency analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1177–1189, Jun. 2016.

[58] S. Zhang, H. Zhou, F. Jiang, and X. Li, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1749–1760, Nov. 2015.

[59] Z. He, X. You, and Y. Y. Tang, "Writer identification of Chinese handwriting documents using hidden Markov tree model," *Pattern Recognit.*, vol. 41, no. 4, pp. 1295–1307, 2008.

[60] X. Fang, S. Teng, Z. Lai, Z. He, S. Xie, and W. Wong, "Robust latent subspace learning for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

Shuangyan Yi received the M.S. degree from the Mathematics Department, Shenzhen Graduate School, Harbin Institute of Technology, China. She is currently pursuing the Ph.D. degree in computer science and technology with the Shenzhen Graduate School, Harbin Institute of Technology. Her current research interests include object tracking, pattern recognition and machine learning.

Zhenyu He received the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2007. He is currently an Associated Professor with the School of Computer Science and Technology, Shenzhen Graduate School, Harbin Institute of Technology, China. His research interests include sparse representation and its applications, deep learning and its applications, pattern recognition, image processing, and computer vision.

Yiu-Ming Cheung received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong. He is a Full Professor with the Department of Computer Science, Hong Kong Baptist University. His current research interests focus on artificial intelligence, visual computing, pattern recognition, and optimization. He is an IET/IEE fellow, a BCS Fellow, and a IETI Fellow. He is the Founding Chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He is currently serving as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, PATTERN RECOGNITION, KNOWLEDGE AND INFORMATION SYSTEMS, and the *International Journal of Pattern Recognition and Artificial Intelligence*.

Wen-Sheng Chen received the B.S. and Ph.D. degrees in mathematics from Sun Yat-Sen University, Guangzhou, China, in 1989 and 1998, respectively. From 2008 to 2009, he was a Visiting Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. He joined the College of Science, Shenzhen University, Shenzhen, China, in 1999. He is currently a Professor with the College of Mathematics and Statistics and the Head of the Department of Information and Computational Science with Shenzhen University. He has authored over 100 refereed scientific research articles in international journals and international conferences in his research areas. His current research interests include image processing and pattern recognition, kernel-based machine learning, wavelet analysis, and its applications.