Low rank embedding for robust image feature extraction

Wai Keung Wong, Zhihui Lai, Jiajun Wen, Xiaozhao Fang, Yuwu Lu

***Abstract*—Robustness to noises, outliers and corruptions is an important issue in linear dimensionality reduction. Since the sample-specific corruptions and outliers exist, the class-special structure or the local geometric structure is destroyed and thus many existing methods, including the popular manifold learning based linear dimensionality methods, fail to achieve good performance in recognition tasks. In this paper, we focus on the unsupervised robust linear dimensionality reduction on corrupted data by introducing the robust low rank representation. Thus a robust linear dimensionality reduction technique termed low rank embedding (LRE) is proposed in this paper, which provides a robust image representation to uncover the potential relationship among the images to reduce the negative influence from the occlusion and corruption so as to enhance the algorithm's robustness in image feature extraction. LRE searches the optimal low rank representation and optimal subspace simultaneously. The model of LRE can be solved by alternatively iterating the argument Lagrangian multiplier method and the eigen decomposition. The theoretical analysis, including convergence analysis and computational complexity, of the algorithms are presented. Experiments on some well-known databases with different corruptions show that LRE is superior to the previous methods of feature extraction and therefore indicates the robustness of the proposed method. The code of this paper can be downloaded from *http://www.scholat.com/laizhihui*.**

***Index Terms*—Robust linear dimensionality reduction, image feature extraction, subspace learning, low rank representation**

I. Introduction

PROJECTION based linear feature-extraction methods are the most simple and effective techniques in computer vision and pattern recognition. In the past decades, many techniques were developed. The most well-known methods include the Principal Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [2] and their extended verisons such as scatter difference discriminant criterion [3], bilateral PCA [4], model-based discriminant analysis [5], kernelized LDA [6], regularized LDA [7], and so on. Since in some cases the label information may be difficult or expensive to obtain and we may have no prior knowledge for new scientific problems [8][9], in this paper, we focus on the unsupervised learning on linear feature extraction.

In the past decade, with the development of manifold learning theory and technique, a large number of local geometry based methods were proposed. For example, He *et al*. proposed the locality preserving projection (LPP) [10] and the orthogonal LPP [11] for face recognition. In [12], [13] and [14], the authors proposed some representative methods for the modified version of LPP, which was further developed for moving object detection [15]. By using the local neighborhood reconstruction information, neighborhood preserving embedding (NPE) [16], orthogonal neighborhood preserving projection [17] and it sparse extension [18]were also proposed for feature extraction, which is essentially to approximate to the Local Linear Embedding (LLE) [19] with a linear projection. Usually, the performance of these methods in image feature extraction and classification were improved when the images of one object lie on a manifold. The key problem of these methods is on how to construct a graph to model the data's structure. However, due to the existence of noise, the variations in illumination and the corruption, the tasks of constructing a robust graph for feature extraction and classification on these images is still a challenging problem. The main reason is that when there are grossly corruptions on the images, the distance between two images will be seriously affected by the corruptions. Thus the graph constructed in this way in this case cannot reflect the real geometric relationship or the potential manifold structure of the image data set. Therefore, the performance of the methods mentioned above will be greatly degraded.

Robust methods on image feature extraction, recognition and clustering have attracted a great deal of attention. One can introduce the robust norms as a metric or regularization to construct more robust model to enhance the performance on the image feature extraction and classification. With the L_1-norm sparse regularized regression, the sparse representation classifier [20] was proposed for robust face recognition. Based on the same idea, the sparsity preserving projection (SPP) [21] was proposed for face image feature extraction. SPP is different from the NPE since it uses the L_1-norm sparse regression to construct the graph. As indicated in [22], the graph construct in this way has been proved to be robust to a

M anuscript received on Aug. 11, 2015. This work was supported by the Natural Science Foundation of China (Grant Nos. 61573248, 61375012, 61362031, 61602270, 61672358), the China Postdoctoral Science Foundation (Grant Nos. 2016M590812, 2016M590100), research grant of The Hong Kong Polytechnic University (Project Grant G-UC42) and Shenzhen Municipal Science and Technology Innovation Council (Grant Nos. JCYJ20150324141711637, JCYJ20160429182058044).

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

Z. Lai (corresponding author) and J. Wen are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China, and also with the Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hong Kong. (e-mail: lai_zhi_hui@163.com, enjoy_world@163.com).

X. Fang is with School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, 510006, China

Y. Lu is with the Graduate School at Shenzhen,Tsinghua University, Shenzhen 518055, China.

TABLE I
COMPARISON OF THE RELATED METHODS MENTIONED IN THIS PAPER

METHOD	ADVANTAGE	LIMITATION	Measurement
LLE	Preserve the local linear reconstruction relationship with nonlinear mapping (unsupervised)	Sensitive to the outliers	L_2 norm
NPE	Preserve the local linear reconstruction relationship with linear mapping (unsupervised)	Sensitive to the outliers	L_2 norm
LPP	Preserve the local neighborhood relationship (unsupervised)	Sensitive to the outliers	L_2 norm
SPP	Preserve the sparse reconstruction relationship (unsupervised)	More robust than NPE and LPP	L_1 and L_2 norm
RPCA	Preserve the maximum covariance with robustness(unsupervised)	Robust to noise and corruption	Nuclear and L_1 norm
IRPCA	Preserve the maximum covariance with robustness(unsupervised)	Robust to noise and corruption	Nuclear and L_1 norm
CTDA	Preserve the discriminant property with robustness (supervised)	Robust to noise and corruption	Nuclear and L_1 norm
LRE	Preserve the low rank reconstruction relationship with robustness (unsupervised)	Robust to noise and corruption	Nuclear and $L_{2,1}$ norm

certain extent. Thus these methods may perform better than the locality based methods such as LPP and NPE when the data contains noise and gross corruption.

Recently, there is another route of study, i.e. low rank representation, to construct the robust graph for data clustering [23]. It is assumed that the data points lie on a low-dimensional subspace, and then the representation matrix of the data points is low rank. In order to address the error correction problem in the classical PCA for uncovering the true low-dimensional subspace structure from the noise data, Wright *et al*. established the so called robust PCA (RPCA) method in [23]. Then, Liu *et al*. [24] [25] extended the single subspace clustering problem [23] into multi-subspace clustering and proposed the low rank representation (LRR) for noisy data clustering. By introducing the manifold structure as the regularized term, the Laplacian regularized LRR [26] was also proposed for clustering the data which lie on the manifold. It is well known that the standard LRR can only work well under the assumption that all the subspaces are independent, which limits its applications. To solve this problem, Tang *et al*. proposed the structure constrained low rank representation method for disjoint subspace segmentation problem [27]. Some nonlinear clustering methods were proposed in [28][29][30] by using the low rank property.

Despite of the subspace clustering and segmentation, low rank property has been widely used for image and video processing. For example, Bhardwaj and Raman [31] used the low rank method for image composition, and Yao and Kwok [32] proposed the image colorization by patch-based local low rank matrix completion. For the applications in the video processing, low-rank property was used in foreground detection [33], and was integrated with sparsity for motion saliency detection [34] and video restoration [35]. A review for comparative evaluation in video surveillance based on RPCA can be found in [36].

However, the above mentioned low rank learning methods only focus on learning a representation low rank matrix to construct the graph for clustering. These methods are essentially transductive methods and therefore cannot handle the new samples which are not involved in the training procedure. In order to solve this problem, inductive RPCA (IRPCA) [37] was proposed to learn a robust projection for feature extraction. In [38], the authors proposed a supervised method termed corruptions tolerant discriminant analysis (CTDA), which integrates the label information, local neighborhood graph and the low rankness of the data together for feature extraction. Comprehensive comparisons for the related algorithms are shown in Table I. Although the robustness of these low-rank based projection learning methods are shown in feature extraction, IRPCA and CTDA lose the function of dimensionality reduction. To address this problem, we propose a new method, called Low Rank Embedding (LRE), for robust linear dimensionality reduction.

One of the most challenges on robust linear dimensionality reduction from high-dimensional image data is to avoid the negative effects brought by the occlusions and corruptions. The motivation of this work is to use the strong robustness of the low rankness to the noise, corruptions and occlusions (the detailed motivations are presented in Section III-A) for developing a novel robust image feature extraction framework. We integrate the optimal low rank representation and projection learning into one model so as to enhance the robustness of the low rankness to deal with the occlusive and corrupted image data in linear dimensionality reduction. The main contributions of this paper are as follows:

(1) We propose a novel robust projection learning method, i.e. Low Rank Embedding (LRE), for unsupervised subspace learning. Based on the low rank constraint and $L_{2,1}$-norm as the robust metric, an iterative method is proposed to solve the regression learning problem.

(2) The theoretical analysis is also explored for the proposed model and the convergence and computational

TABLE II
THE NOTATIONS USED IN THIS PAPER

$\mathbf{X}$	Training data matrix
$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$	Training data
$\mathbf{P}$	Projection matrix
$\mathbf{W}$, $\mathbf{W}^{NPE}$	Neighborhood matrix of LPP and NPE
$\mathbf{Z}$	Low rank matrix
$\mathbf{E}$	Error matrix
$\mathbf{J}$	Auxiliary variable used in LRE
$\mathbf{Y}_1, \mathbf{Y}_2$	Multipliers used in LRE
μ , ρ , λ	Parameters used in LRE

complexity are presented.

(3) Extensive experiments show that the proposed LRE performs better than the previous methods in most cases when the data are corrupted or contain different noises.

The rest of the paper is organized as follows. In Section II, the related works are reviewed. In Section III, the LRE model and algorithm are presented. In Section IV, theoretical analyses are performed on the proposed method and some important properties are obtained. Experiments are carried out in Section V to evaluate the proposed subspace learning algorithms when the databases have different kinds of noise or occlusion. The conclusions are given in Section VI.

II. RELATED WORKS

In this section, we briefly review some related works, including LPP, NPE, and LRR. In this paper, lowercase and uppercase italic letters, i.e. i, n, t, λ, T etc., denote scalars, bold lowercase letters, i.e. $\mathbf{x}, \mathbf{p}$ etc., denote vectors, and bold uppercase letters, i.e. $\mathbf{X}, \mathbf{D}, \mathbf{P}$ etc., denote the matrices. Let matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ be the data matrix including all training samples $\{\mathbf{x}_i\}_{i=1}^{n} \in R^m$ in its columns. In practice, the feature dimension m is often very high. The goal of the dimensionality reduction methods is to find an optimal projection $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_d) \in R^{d \times m}$ so as to transform the data onto a d dimensional subspace, where $d \ll m$. Table II lists the notations used in this paper.

A. Locality Preserving Projection

LPP also is an unsupervised dimensionality reduction approach which aims to preserve the local structure of the data. The optimal projection $\mathbf{P}^*$ derived by the objective function of LPP is defined as follows:

$$\begin{aligned}\mathbf{P}^* &= \arg\min_{\mathbf{P}} \frac{1}{2} \sum\nolimits_i \sum\nolimits_j \left\| \mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j \right\|^2 \mathbf{W}_{ij} \\ &= \arg\min_{\mathbf{P}} tr(\mathbf{P}^T \mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T \mathbf{P})\end{aligned} \quad (1)$$

where $\mathbf{D}$ is a diagonal matrix with its entries being the row sums of $\mathbf{W}$, i.e. $\mathbf{D}_{ii} = \sum\nolimits_j \mathbf{W}_{ij}$, $\mathbf{L} = \mathbf{D}\text{-}\mathbf{W}$ and the weight $\mathbf{W}$ defined as follows:

$$\mathbf{W}_{ij} = \begin{cases} 1, & if \ \mathbf{x}_i \in N_k(\mathbf{x}_j) \ \ or \ \mathbf{x}_j \in N_k(\mathbf{x}_i) \\ 0, & otherwise \end{cases}$$

where $N_k(\mathbf{x}_i)$ denotes the k nearest neighbors of $\mathbf{x}_i$.

Minimizing (1) ensures that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close then $\mathbf{P}^T \mathbf{x}_i$ and $\mathbf{P}^T \mathbf{x}_j$ should be close as well. That is objective function (1) can preserve the neighborhood relationship among the data points and the optimal projection is the linear approximation to the Laplace Beltrami operator for a manifold generated by the Riemannian metric [39]. By imposing a constraint $\mathbf{P}^T \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{P} = \mathbf{I}$, the optimal projections of LPP is given by the minimum eigenvalue solution to the following generalized eigenvalue problem:

$$\mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T \mathbf{p} = \alpha \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{p} \quad (2)$$

where $\mathbf{p}$ is a eigenvector corresponding to eigenvalue α , which constructs the column of $\mathbf{P}$. The optimal projections for the LPP are the eigenvectors corresponding to the smaller non-zero eigenvalues [10].

B. Neighborhood Preserving Embedding

Similar to LPP, NPE aims to keep the local neighborhood structure of the data in the low dimensional space. The local approximation error in NPE is measured by minimizing the cost function:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_i \left\| \mathbf{x}_i - \sum_{j \in \pi_k(\mathbf{x}_i)} \mathbf{W}_{ij}^{NPE} \mathbf{x}_j \right\|^2 \ \ (j = 1, 2, ...n) \quad (3)$$

where $\pi_k(\mathbf{x}_i)$ denotes the index set of k nearest neighbors of $\mathbf{x}_i$ and $\mathbf{W}_{ij}^{NPE}$ is the optimal local least square reconstruction coefficients. The criterion for choosing an optimal projection $\mathbf{p}$ is to minimize the cost function:

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} \sum_i \left\| \mathbf{p}^T \mathbf{x}_i - \sum_{j \in \pi_k(\mathbf{x}_i)} \mathbf{W}_{ij}^{NPE} \mathbf{p}^T \mathbf{x}_j \right\|^2 \quad (4)$$

By removing an arbitrary scaling factor, the optimal projections of NPE are the eigenvectors corresponding to the minimum eigenvalue of the following generalized eigenvalue problem:

$$\mathbf{X}(\mathbf{I}-\mathbf{W}^{NPE})^T(\mathbf{I}-\mathbf{W}^{NPE})\mathbf{X}^T \mathbf{p} = \alpha \mathbf{X} \mathbf{X}^T \mathbf{p} \quad (5)$$

C. Low Rank Representation

As a subspace clustering method, the basic idea of low rank representation (LRR) is to capture the lowest rank representation in the combination of the bases in the given dataset. The problem can be formulated as the rank minimization problem

$$\mathbf{Z}^* = \arg\min_{\mathbf{Z}} rank(\mathbf{Z}), \ \ s.t. \ \mathbf{X} = \mathbf{X}\mathbf{Z} , \quad (6)$$

However, the above problem is NP-hard. Similar to the trick used in the matrix completion, we practically solve the relaxed optimization problem.

$$\mathbf{Z}^* = \arg\min_{\mathbf{Z}} \left\| \mathbf{Z} \right\|_* , \ \ s.t. \ \mathbf{X} = \mathbf{X}\mathbf{Z} \quad (7)$$

where $\left\| \cdot \right\|_*$ represents the nuclear norm of a matrix. In applications, the data are often noisy and even contain outliers, thus, (7) can be rewritten as

$$(\mathbf{Z}^*, \mathbf{E}^*) = \arg\min_{\mathbf{Z},\mathbf{E}} \left\| \mathbf{Z} \right\|_* + \lambda \left\| \mathbf{E} \right\|_l , \ \ s.t. \ \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E} \quad (8)$$

where $\lambda > 0$ is a parameter and $\mathbf{E} \in R^{m \times n}$ is a noise matrix and $\left\| \cdot \right\|_l$ denotes certain norm regularization, such as L_1 and

$L_{2,1}$ norm. Although the LRR is proved to be effective for robust clustering problem, it cannot be used for linear dimensionality reduction task.

III. Low Rank Embedding

In this section, we first present the motivation of the proposed method and then the LRE algorithm and its solutions.

A. *The motivation of LRE*

In the past decade, the representative manifold learning methods [19][40] were proposed and a lot of previous research focus on exploring the geometric structure or class-specific structure for dimensionality reduction. The well-known methods, such as NPE, LPP and its extensions [11] [12][13] [41], can preserve the local relationship among the data to learn an optimal projection for linear dimensionality reduction. The neighborhood relationship plays an important role in the manifold learning methods. The assumption of the manifold learning based methods is that the data lies on a smooth manifold embedded in the latent high dimensional space. This ideal case, however, cannot be guaranteed in the real world applications. Particularly, when there are sample-specific corruptions and outliers, the data points will not distribute on the local neighborhood even if the data points are belonging to the same objective. Thus, the manifold learning based methods may lose its effectiveness in the learned low-dimensional linear subspace since the intrinsic locality is seriously destroyed.

The low rank representation assumes that the data samples are approximately drawn form a subspace or a union of multiple subspaces [27]. It is shown that even if the data contains noise, the LRR still can cluster the samples into their respective subspaces and remove possible outliers [24]. It is proved that under certain conditions LRR can exactly recover the space of the original data and approximately recover the data space with theoretical guarantees [23], which indicates the robustness for subspace clustering. However, the LRR cannot be directly used for linear dimensionality reduction and thus it lacks the function of feature extraction.

Due to the LRR's robustness to the noise, corruptions and occlusions, it is expected to take full use of these properties in learning a robust subspace for linear dimensionality reduction. It is known that LRR can obtain the robust representation matrix $\mathbf{Z}$ among the data set, which explores the intrinsic low rank representation relationship [23] [24] [25]. Thus we integrate the low rank representation and low-dimensional subspace learning (i.e. to learn a projection matrix $\mathbf{P}$) together to address the robust subspace learning problem and at the same time to solve the problem of obtaining the optimal projection derived from the LRR for linear dimensionality reduction.

B. *Problem formulation for LRE*

In this paper, we focus on the robust linear dimensionality reduction on the data with sample-specific corruptions and outliers. In order to increase the robustness of the proposed method, we use the $L_{2,1}$ norm as the basic metric to measure the reconstructive errors. It is also assumed that the data can be approximately reconstructed by a low rank matrix $\mathbf{Z}$ on the orthogonal subspace $\mathbf{P}$. Thus we obtain the following objective function for LRPE:

$$(\mathbf{Z}^*,\mathbf{P}^*)=\arg\min_{\mathbf{Z},\mathbf{P}} rank(\mathbf{Z})+\lambda\left\|\mathbf{P}^T\mathbf{X}-\mathbf{P}^T\mathbf{X}\mathbf{Z}\right\|_{2,1}, \quad s.t.\ \mathbf{P}^T\mathbf{P}=\mathbf{I} \tag{9}$$

The above objective function aims to find the optimal low rank reconstructive matrix and the projection. The reconstructive property is measured by the term $\left\|\mathbf{P}^T\mathbf{X}-\mathbf{P}^T\mathbf{X}\mathbf{Z}\right\|_{2,1}$. There are two main reasons to use the $L_{2,1}$ norm in the proposed model. The first one is that the $L_{2,1}$ norm is more robust than the Frobenious norm in characterizing the error term [42]. The second is we tend to model the sample-specific corruptions (and outliers) as in [24].

Since the rank minimization problem is NP-hard, as suggested by [23], the problem can be effectively solved by replacing the rank function with its convex lower bound $\|\cdot\|_*$. Thus we have the nuclear norm minimization problem:

$$(\mathbf{Z}^*,\mathbf{P}^*)=\arg\min_{Z,P}\|\mathbf{Z}\|_*+\lambda\left\|\mathbf{P}^T\mathbf{X}-\mathbf{P}^T\mathbf{X}\mathbf{Z}\right\|_{2,1}, \quad s.t.\ \mathbf{P}^T\mathbf{P}=\mathbf{I} \tag{10}$$

Since it is known that the $L_{2,1}$ norm is more robust than the Frobenious norm, it is expected that this objective function is more robust than the Frobenious norm when there are outliers or the data are with sample-specific corruptions.

C. *The optimal solution*

In this subsection, we show how to solve the optimization problem. Since it is impossible to simultaneously obtain the optimal solutions of the two variables in the model, we design an alternatively iterative algorithm to solve this optimization problem. The idea of the iterative algorithm contains two steps: we first fix the variable $\mathbf{Z}$ to compute the optimal $\mathbf{P}$, and then fix $\mathbf{P}$ to compute the optimal $\mathbf{Z}$.

Step 1: Fix $\mathbf{P}$ to compute the optimal $\mathbf{Z}$ and $\mathbf{E}$

First of all, we convert the optimization problem (10) to an equivalent problem. Let

$$\mathbf{E}=\mathbf{P}^T\mathbf{X}-\mathbf{P}^T\mathbf{X}\mathbf{Z} \tag{11}$$

Then (10) can be converted to the following optimization problem

$$(\mathbf{Z}^*,\mathbf{E}^*,\mathbf{P}^*)=\arg\min_{\mathbf{Z},\mathbf{E},\mathbf{P}}\|\mathbf{Z}\|_*+\lambda\|\mathbf{E}\|_{2,1}, \quad s.t.\ \mathbf{P}^T\mathbf{X}=\mathbf{P}^T\mathbf{X}\mathbf{Z}+\mathbf{E} \tag{12}$$

It can be found that (12) is the modified LRR problem, thus it can be solved by the LRR algorithm in which P^TX is viewed as the data matrix used for low rank decomposition. In fact, optimization problem (12) can be solved by the ALM algorithm, which is stated as follows.

At first, we convert (12) to the following equivalent problem:

$$(\mathbf{Z}^*,\mathbf{E}^*,\mathbf{P}^*,\mathbf{J}^*)=\arg\min_{\mathbf{Z},\mathbf{E}}\|\mathbf{J}\|_*+\lambda\|\mathbf{E}\|_{2,1}, \quad s.t.\ \mathbf{P}^T\mathbf{X}=\mathbf{P}^T\mathbf{X}\mathbf{Z}+\mathbf{E},\ \mathbf{Z}=\mathbf{J} \tag{13}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2691543, IEEE Transactions on Image Processing

ALGORITHM I
LRE ALGORITHM

```
Input: Training samples ${\mathbf{x}_i \in R^m, i = 1,2,...,N}$, the numbers of iterations $T$, dimensions $d$
Output: Low-dimensional features $\tilde{\mathbf{x}}_i$（$i = 1,2,...,N$） and the optimal projection $\mathbf{P}$
Step 1: Initialize $\mathbf{G} = \mathbf{I}$ and $\mathbf{P}$ as the matrix with orthogonal column vectors.
Step 2: For $t = 1:T$
        - Step 2.1:
            -Initialize: $\mathbf{Z} = \mathbf{J} = \mathbf{0}$, $\mathbf{E} = \mathbf{0}$, $\mathbf{Y}_1 = \mathbf{Y}_2 = \mathbf{0}, \mu = 10^{-6}, \mu_{maix} = 10^6, \rho = 1.1$
            -Update the variables using (15)-(18) until the iteration converges and obtain the optimal $\mathbf{Z}$.
        - Step 2.2: Update $\mathbf{G}$ using (20).
        - Step 2.3: Solve the eigenfunction (22) to obtain the optimal $\mathbf{P}$.
    End
Step 3: Project the samples onto the low-dimensional subspace to obtain $\tilde{\mathbf{x}}_i = \mathbf{P}^T\mathbf{x}_i$ for classification.
```

The above problem can be solved by the ALM method, which aims to minimize the following augmented Lagrangian function:

$$\mathcal{L} = \|\mathbf{J}\|_* + \lambda\|\mathbf{E}\|_{2,1} + tr(\mathbf{Y}_1^T(\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z} - \mathbf{E})) + tr(\mathbf{Y}_2^T(\mathbf{Z} - \mathbf{J})) + \mu/2(\|\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2) \tag{14}$$

where $\mu > 0$ is a penalty parameter, and $\|\cdot\|_F$ denotes the Frobenious norm of a matrix, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are the Lagrangian multipliers. The above problem is unconstrained and it can be minimized one by one with respect to variables $\mathbf{Z}$, $\mathbf{J}$, $\mathbf{E}$ by fixing the other variables, respectively, and then updating the Lagrangian multipliers $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

The optimal $\mathbf{J}^*$ can be computed and updated by

$$\mathbf{J}^* = \arg\min_{\mathbf{J}} 1/\mu\|\mathbf{J}\|_* + 1/2\|\mathbf{J} - (\mathbf{Z} + \mathbf{Y}_2/\mu)\|_F^2 \tag{15}$$

which can be solved by the Singular Value Thresholding operator [43].

The optimal $\mathbf{Z}^*$ can be computed and updated by

$$\mathbf{Z}^* = (\mathbf{I} + \mathbf{X}^T\mathbf{P}\mathbf{P}^T\mathbf{X})^{-1}[\mathbf{X}^T\mathbf{P}(\mathbf{P}^T\mathbf{X} - \mathbf{E})) + \mathbf{J} + (\mathbf{X}^T\mathbf{P}\mathbf{Y}_1 - \mathbf{Y}_2)/\mu] \tag{16}$$

And the optimal $\mathbf{E}^*$ can be updated by

$$\mathbf{E}^* = \arg\min_{\mathbf{E}} \lambda/\mu\|\mathbf{E}\|_{2,1} + 1/2\|\mathbf{E} - (\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z} + \mathbf{Y}_1/\mu)\|_F^2 \tag{17}$$

which can be solved by the algorithm proposed in [44].

At each step, we update the multipliers and the parameter as follows:

$$\begin{aligned} &\mathbf{Y}_1 \leftarrow \mathbf{Y}_1 + \mu(\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z} - \mathbf{E}) \\ &\mathbf{Y}_2 \leftarrow \mathbf{Y}_2 + \mu(\mathbf{Z} - \mathbf{J}) \\ &\mu \leftarrow \min(\rho\mu, \mu_{\max}) \end{aligned} \tag{18}$$

where $\rho > 0$ is parameter set by users.

Step 2: Fix $\mathbf{Z}$ to compute the optimal $\mathbf{P}$.

Let us consider the case when Z is given. In this case, the optimization problem becomes

$$\mathbf{P}^* = \arg\min_{\mathbf{P}} \|\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}\|_{2,1}, \quad s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{19}$$

The above problem can be solved using the recently proposed $L_{2,1}$ norm minimization technique, which also includes two steps. The first step is to compute a diagonal matrix $\mathbf{G}$ which is defined as

$$\mathbf{G}_{ii} = \frac{1}{2\|(\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z})^i\|_2} \tag{20}$$

where $(\cdot)^i$ denotes the i-th column of a matrix.

Then, (19) is converted to solve the following equivalent trace minimization problem

$$\mathbf{P}^* = \arg\min_{\mathbf{P}} tr(\mathbf{P}^T\mathbf{X}(\mathbf{I} - \mathbf{Z})\mathbf{G}(\mathbf{I} - \mathbf{Z})^T\mathbf{X}\mathbf{P}), \quad s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{21}$$

The optimal solution of (21) can be obtained from solving the standard eigenfunction:

$$\mathbf{X}(\mathbf{I} - \mathbf{Z})\mathbf{G}(\mathbf{I} - \mathbf{Z})^T\mathbf{X}\mathbf{p} = \alpha\mathbf{p} \tag{22}$$

where α is the eigenvalue and p is the corresponding eigenvector. The optimal solution $\mathbf{P}^*$ contains the eigenvectors corresponding to the smaller none-zero eigenvalues.

The detailed algorithm steps of the iterative method are presented in Algorithm I. and a block diagram of the LRE algorithm is shown in Fig. 1 for ease of understanding.

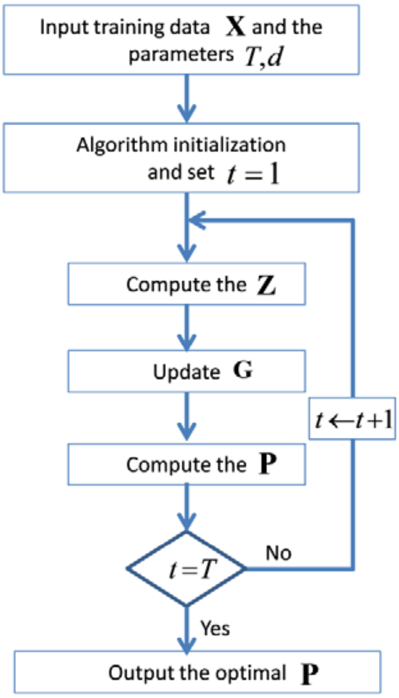

Input training data $\mathbf{X}$ and the parameters T,d
Algorithm initialization and set $t=1$
Compute the $\mathbf{Z}$
Update $\mathbf{G}$
Compute the $\mathbf{P}$
$t \leftarrow t+1$
$t=T$
No
Yes
Output the optimal $\mathbf{P}$

Fig. 1 The diagram of the proposed LRE algorithm.

IV. ALGORITHM ANALYSIS

In this section, we first present the theoretical analysis on the algorithm's convergence. Then the computational complexity is also presented. At last, we give the detailed comparisons between the proposed method and the other most related classical methods.

A. Convergence analysis

Since the proposed algorithm is an iterative method, we need to prove the convergence of the algorithm.

Let

$$J(\mathbf{Z},\mathbf{P}) = \|\mathbf{Z}\|_* + \lambda \left\|\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{X}\mathbf{Z}\right\|_{2,1} \tag{23}$$

We can obtain the following Theorem.

Theorem 1. *The iterative scheme in algorithm 1 monotonically decreases the objective function value of $J(\mathbf{Z},\mathbf{P})$ in each iteration.*

Proof.

Suppose in the t-th iteration, we have the following result from (12):

$$J(\mathbf{Z}_t,\mathbf{P}_t) = \|\mathbf{Z}_t\|_* + \lambda \left\|\mathbf{P}_t^T\mathbf{X} - \mathbf{P}_t^T\mathbf{X}\mathbf{Z}_t\right\|_{2,1} \tag{24}$$

For the given P_t, since the optimization problem (10) or (11) can be solved by the AML algorithm, which reduces the objective function value by solving the equivalent optimization problem, thus we have

$$J(\mathbf{Z}_t,\mathbf{P}_t) \geq J(\mathbf{Z}_{t+1},\mathbf{P}_t) = \|\mathbf{Z}_{t+1}\|_* + \lambda \left\|\mathbf{P}_t^T\mathbf{X} - \mathbf{P}_t^T\mathbf{X}\mathbf{Z}_{t+1}\right\|_{2,1} \tag{25}$$

On the other side, for the given Z_{t+1}, we can define the objective function containing the diagonal matrix as

$$\begin{aligned} & J(\mathbf{Z}_{t+1},\mathbf{P}_t) \triangleq J(\mathbf{Z}_{t+1},\mathbf{P}_t,\mathbf{G}_{t+1}) \\ & = \|\mathbf{Z}_{t+1}\|_* + \lambda \left\|\mathbf{P}_t^T\mathbf{X} - \mathbf{P}_t^T\mathbf{X}\mathbf{Z}_{t+1}\right\|_{2,1} \\ & = \|\mathbf{Z}_{t+1}\|_* + \lambda tr(\mathbf{P}_t^T\mathbf{X}(\mathbf{I}-\mathbf{Z}_{t+1})\mathbf{G}_{t+1}(\mathbf{I}-\mathbf{Z}_{t+1})^T\mathbf{X}\mathbf{P}_t) \end{aligned} \tag{26}$$

where

$$\mathbf{G}_{t+1,ii} = \frac{1}{2\left\|(\mathbf{P}_t^T\mathbf{X} - \mathbf{P}_t^T\mathbf{X}\mathbf{Z})^i\right\|_2}$$

When the $\mathbf{G}_{t+1}$ is obtained, the algorithm solves the standard eigenfuction to obtain $\mathbf{P}_{t+1}$ which further reduces the objective function value, thus we have

$$J(\mathbf{Z}_{t+1},\mathbf{P}_{t+1}) \triangleq J(\mathbf{Z}_{t+1},\mathbf{P}_{t+t},\mathbf{G}_{t+1}) \leq J(\mathbf{Z}_{t+1},\mathbf{P}_t,\mathbf{G}_{t+1}) \tag{27}$$

We conclude from (18) and (19) that $J(\mathbf{Z}_t,\mathbf{P}_t) \geq J(\mathbf{Z}_{t+1},\mathbf{P}_{t+1})$. Thus, the iterative algorithm converges. □

Since the objective function (10) of LRE has positive lower bound, from Theorem 1 we can know that the iterative algorithm will converge to local optimal solution. In fact, we will show in the experimental section that the iterative algorithm will converge very fast. Usually, the outer loop of the algorithm will converge within 3~5 time iterations.

B. Computational complexity

Suppose the dimension of the data is larger than the number of the samples, i.e. $m > n$ .We can find that the main computational complexity comes from the eigen-decomposition of (15) and the ALM method. The eigen-decompostion needs $O(m^3)$ and the main computation in ALM algorithm is the SVD decomposition, which is also at most $O(m^3)$ in each iteration. If the algorithm converges within T iteration steps for its outer loop, the upper bound of the total computational complexity is at most $O(Tm^3 + Ttm^3)$, where t denotes the loops in the inner iteration of ALM algorithm. This is very large when the dimension of the samples and the T are larger numbers. Fortunately, the outer iteration converges very fast and thus the total computational burden is with the same order to the classical PCA method. Moreover, when the dimension of the data is very high, one can also use the KL transformation to compute the standard eigenvectors, which will greatly reduce the computational burdens.

C. Some properties of the LRE

In this section, we explore some properties of the proposed model. At the first glance on the optimization problem (12), we may believe that the optimal low rank matrix $\mathbf{Z}^*$ lies on the subspace spanned by projection matrix P . In fact, it is not true. The following theorem uncovers the relationship between $\mathbf{Z}^*$ and $\mathbf{P}^*$.

Theorem 2. *For any optimal solution $(\mathbf{Z}^*,\mathbf{E}^*,\mathbf{P}^*)$ to optimization problem (12), we have $\mathbf{Z}^* \in span(\mathbf{X}^T)$.*

Proof.

Note that Eq.(12) always has feasible solution(s), for instance, the solution $(\mathbf{Z}=\mathbf{0},\mathbf{E}=\mathbf{0},\mathbf{P}=\mathbf{I})$ is feasible. Supposed the optimal solution denoted as $(\mathbf{Z}^*,\mathbf{E}^*,\mathbf{P}^*)$ exists.

Let the skinny SVD of $[\mathbf{P}^T\mathbf{X},\mathbf{P}^T\mathbf{X}-\mathbf{E}] = \mathbf{U}\mathbf{\Xi}\mathbf{V}^T$, and partition $\mathbf{V}=[\mathbf{V}_\mathbf{X};\mathbf{V}_\mathbf{E}]$ such that $\mathbf{P}^T\mathbf{X} = \mathbf{U}\mathbf{\Xi}\mathbf{V}_\mathbf{X}^T$, then $\mathbf{V}_\mathbf{X} = \mathbf{X}^T\mathbf{P}\mathbf{U}\mathbf{\Xi}^{-1}$.

From Theorem 5.1 in [24], we know that the optimal $\mathbf{Z}^*$ can be represented as

$$\mathbf{Z}^* = \mathbf{V}_\mathbf{X}(\mathbf{V}_\mathbf{X}^T\mathbf{V}_\mathbf{X})^{-1}\mathbf{V}_\mathbf{E}^T \tag{28}$$

Substituting $\mathbf{V}_\mathbf{X} = \mathbf{X}^T\mathbf{P}\mathbf{U}\mathbf{\Xi}^{-1}$ in above equation we have

$$\mathbf{Z}^* = \mathbf{X}^T\mathbf{P}^*\mathbf{U}\mathbf{\Xi}^{-1}(\mathbf{V}_\mathbf{X}^T\mathbf{V}_\mathbf{X})^{-1}\mathbf{V}_\mathbf{E}^T \tag{29}$$

This indicates that $\mathbf{Z}^* \in span(\mathbf{X}^T)$. □

The above theorem show that the optimal $\mathbf{Z}^*$ still lies on the row space of the data matrix, which is similar to the LRR problem. The orthogonal projection matrix $\mathbf{P}^*$ is one of the representation factors for the $\mathbf{Z}^*$.

In fact, if the reconstructed error is zero, i.e. for the given $\mathbf{P}$, $\mathbf{P}^T\mathbf{X}-\mathbf{P}^T\mathbf{X}\mathbf{Z}=\mathbf{E}=\mathbf{0}$ in (12), we have the following conclusion.

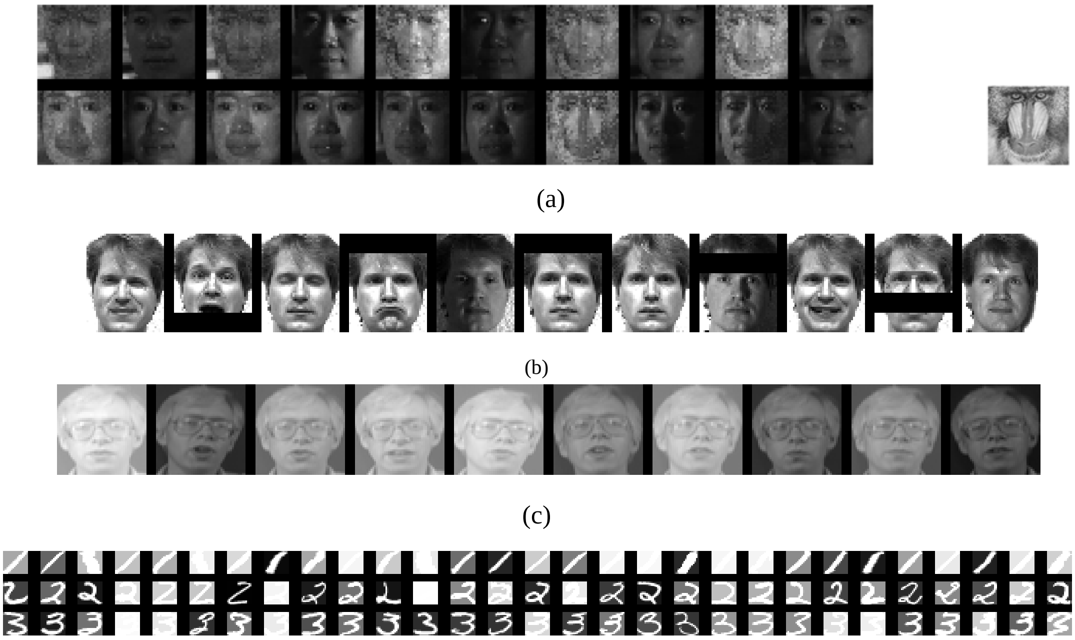
Proposition 1. *Suppose $rank(\mathbf{X}) = n$. For any n-dimensional orthogonal subspace $\mathbf{P}$ spanned by the column of $\mathbf{X}$, if $\mathbf{E}=\mathbf{0}$, the optimization problem*

$$\min_\mathbf{Z} \|\mathbf{Z}\|_* \quad s.t.\ \mathbf{P}^T\mathbf{X} = \mathbf{P}^T\mathbf{X}\mathbf{Z} \tag{30}$$

and (7) has the same solutions.

Proof.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2691543, IEEE Transactions on Image Processing



(a)

(b)

(c)

(d)

Fig. 2. The corrupted image samples used in the experiments. (a) CMU PIE face database image (left) and the baboon face image (right) used for sample-specific-occlusions. (b) Corrupted face images on Yale database. (c) Corrupted images on ORL face database. (d) USPS digital image database corrpted by random gray images.

Let the skinny SVD of $\mathbf{X}=\bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^T$.

For one thing, it is easy to check that the solution of (7) is also the solution of (12). Since the constraint $\mathbf{X}=\mathbf{X}\mathbf{Z}^*$ in (7) always indicates that for any projection $\mathbf{P}$ we always have $\mathbf{P}^T\mathbf{X}\equiv\mathbf{P}^T\mathbf{X}\mathbf{Z}^*$. From the result (theorem 5.1) in [24], the optimal solution of (7) is $\mathbf{V}\mathbf{V}^T$. Thus, $\mathbf{Z}^*=\bar{\mathbf{V}}\bar{\mathbf{V}}^T$ is also the optimal solution of (30).

Furthermore, since orthogonal subspace P is spanned by the column of $\mathbf{X}$, the SVD of $\mathbf{X}=\bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^T\Rightarrow\mathbf{P}^T\mathbf{X}=(\mathbf{P}^T\bar{\mathbf{U}})\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^T$, thus $\bar{\mathbf{V}}\bar{\mathbf{V}}^T$ is the optimal solution of (30), which is still the optimal solution of (7). □

The Proposition 1 indicates that if there is no noise, for any n-dimensional subspace $\mathbf{P}$ spanned by the column of $\mathbf{X}$, the projection has no influence on the representation coefficient matrix. In fact, this ideal case seldom happens since $\mathbf{E}$ is usually not equal to zero (i.e. the data cannot be completely reconstructed) and the number of the projection vector in $\mathbf{P}$ may also not equal to n. Thus, it is necessary to design an iterative algorithm as stated in Table I to obtain the optimal projection matrix.

Similarity and significant difference between the proposed LRE and reinforcement learning exist. The similarity is that both of them use the iterative method to update the optimal strategy or optimal solution. The difference between them is very significant. The proposed LRE simply and mechanically compute the optimal solution in each step. However, the reinforcement learning has different states and actions in each iteration, and the learning procedures interact with its environment so as to achieve the maximum cumulative effect.

V. Experiments

In this section, a set of experiments are presented to show the effectiveness of the proposed subspace learning algorithms for image feature extraction and recognition against the classical subspace learning methods (i.e. PCA), the most related manifold learning based methods (i.e. LPP, NPE and ONPP) and the recently proposed IRPCA. The CMU PIE (Pose29, light and illumination change) face database is used to evaluate the performance of these methods when there are variations in face poses and lighting conditions with occlusions by an image. The Yale face database is used to test the performances of the algorithms when there are block subtraction in the face images. The ORL face database is used to test the robustness of the proposed algorithms when all the images are occluded with different levels of gray images. The USPS handwriting digital image database is used to test the robustness of the proposed algorithm on handwritten digital image recognition when some of the images are occluded. The nearest neighbor classifier with the Euclidean distance is used in all experiments. The code of this paper can be downloaded from *http://www.scholat.com/laizhihui*.

A. The description of the databases

The CMU PIE face database [45] contains 68 individual with 41,368 face images as a whole. The face images were captured under varying pose, illumination and expression. In our experiments, we selected a subset (C29) which contains 1632 images of 68 individuals (each individual has 24 images). The C29 subset involves variations in illumination, facial expression and pose. All of these face images are aligned based on eye coordinates and cropped to 32×32. Half of the

TABLE III

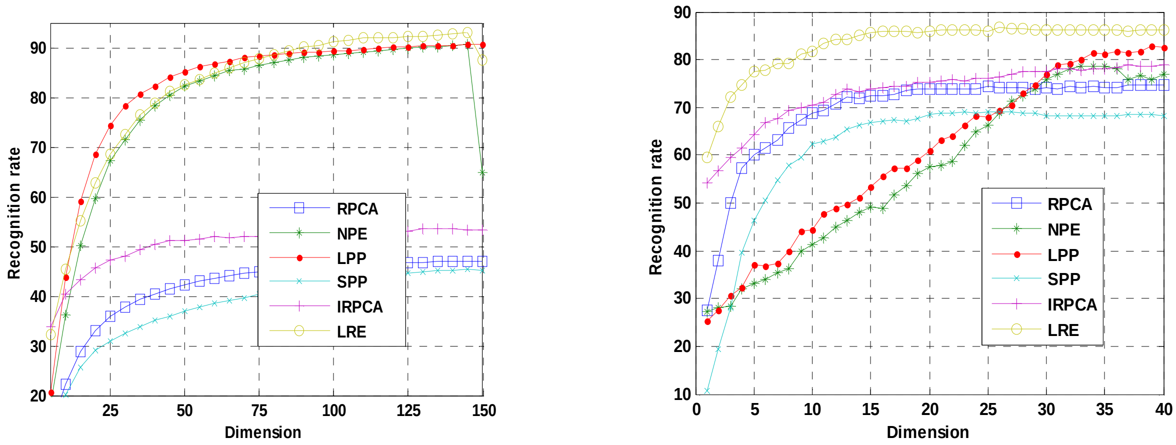
COMPARISON OF THE PERFORMANCE (RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION) OF DIFFERENT ALGORITHMS ON CMU PIE DATA SET

L	RPCA	NPE	LPP	SPP	IRPCA	LRE
6	47.00±7.60 (140)	90.57±4.05 (145)	90.53±3.99 (145)	45.29±8.45 (145)	53.49±3.47 (140)	**93.08**±4.31 (145)
5	41.07±8.43 (145)	83.33±6.40 (135)	83.614±5.08 (130)	39.16±8.52 (140)	45.76±7.20 (140)	**85.65**±4.81 (135)
4	35.87±7.90 (150)	74.51±10.85 (140)	75.05±6.96 (140)	34.47±8.21 (140)	37.53±7.84 (150)	**77.49**±5.17 (130)

TABLE IV

COMPARISON OF THE PERFORMANCE (RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION) OF DIFFERENT ALGORITHMS ON YALE DATA SET

L	RPCA	NPE	LPP	SPP	IRPCA	LRE
6	74.58±5.23 (37)	76.91±4.03 (40)	82.83±5.09 (39)	69.33±7.43 (24)	79.00±4.18 (66)	**86.46**±4.46 (39)
5	66.58±6.86 (38)	72.75±4.03 (37)	77.50±5.25 (39)	63.55±7.87 (31)	77.88±4.64 (39)	**82.66**±4.46 (38)
4	63.83±6.21 (39)	67.33±7.21 (38)	73.00±5.98 (39)	60.09±5.26 (28)	74.57±5.12 (37)	**80.43**±6.54 (37)



(a) (b)

Fig.3. The average recognition rates (%) versus the variations of the dimension of the subspace. (a) On CMU PIE face database. (b) On Yale face database.

images are added by the baboon face image as the continuous occlusion with random intensity. Fig. 2 (a) shows the sample images from this database with different noise densities.

The Yale face database contains 165 images of 15 individuals (each person providing 11 different images) with various facial expressions and lighting conditions. In our experiments, each image was manually cropped and resized to 32×32 pixels. Half of the images in the database were performed as block subtraction, where one fifth image pixel was randomly subtracted. Fig. 2 (b) shows sample images of one person in the Yale database, in which half of the images are the block subtraction image.

The ORL database is used to evaluate the performance of LPE under conditions where the pose, face expression and sample size vary. The ORL face database contains images from 40 individuals, each providing 10 different images. The facial expressions and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there are also some variations in the scale of up to about 10 percent. All images were normalized to a resolution of 32×32. Some occlusion images with different intensity/level (random) gray image are shown in Fig. 2 (c).

The USPS handwriting digital image database includes 10 classes from “0” to “9”. Each class has 1100 examples. In our experiment, we selected a subset from the original database. We cropped each image to be size of 16×16. There are 100 images for each class in the subset and the total number is 1000. Half of the images were added by different level (random) gray images as the sample special corruptions. Fig. 2 (d) displays a subset of the occlusion image from original USPS handwriting digital database.

B. Experiment setup

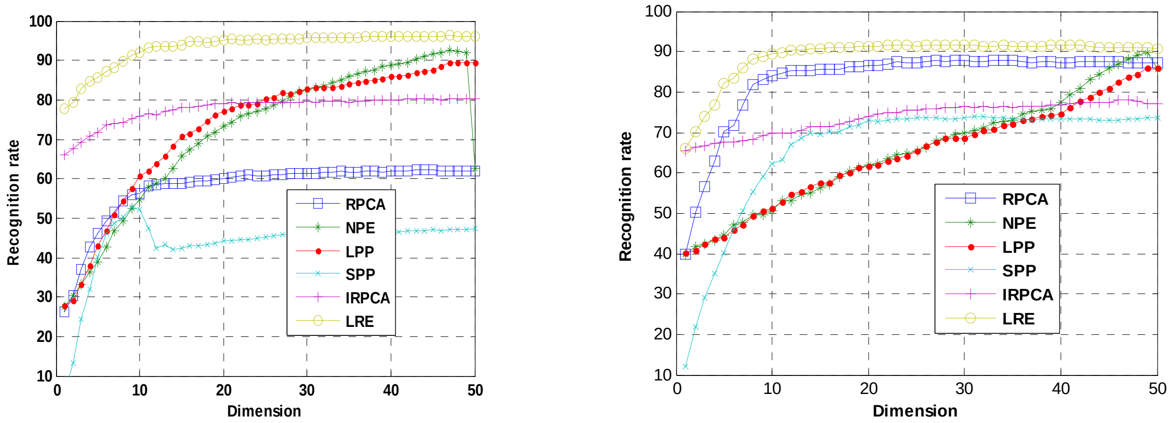
In the experiments, we compare the proposed method with the RPCA, LPP, NPE, SPP and IRPCA. Note that the SPP is the L_1-norm based sparse representation method for feature extraction, which is shown to be robust to some kind of noise in previous research. IRPCA is the low rank based robust feature extraction method. In the experiments, L images of each individual were randomly selected and used as the training set, and one half of the remaining images were used as the validation set and test set, respectively. The best

TABLE V
COMPARISON OF THE PERFORMANCE (RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION) OF DIFFERENT ALGORITHMS ON ORL DATA SET

L	RPCA	NPE	LPP	SPP	IRPCA	LRE
6	62.03±3.99 (44)	92.43±6.21 (47)	89.41±7.96 (50)	47.37±5.17 (50)	80.37±9.96 (47)	**96.25**±2.13 (47)
5	51.5±3.63 (49)	90.65±6.51 (48)	86.46±7.67 (50)	44.80±6.25 (50)	72.40±9.53 (43)	**93.54**±2.04 (43)
4	42.35±2.66 (39)	86.46±4.36 (49)	81.78±7.97 (49)	37.54±3.94 (47)	68.75±4.40 (50)	**89.65**±1.70 (48)

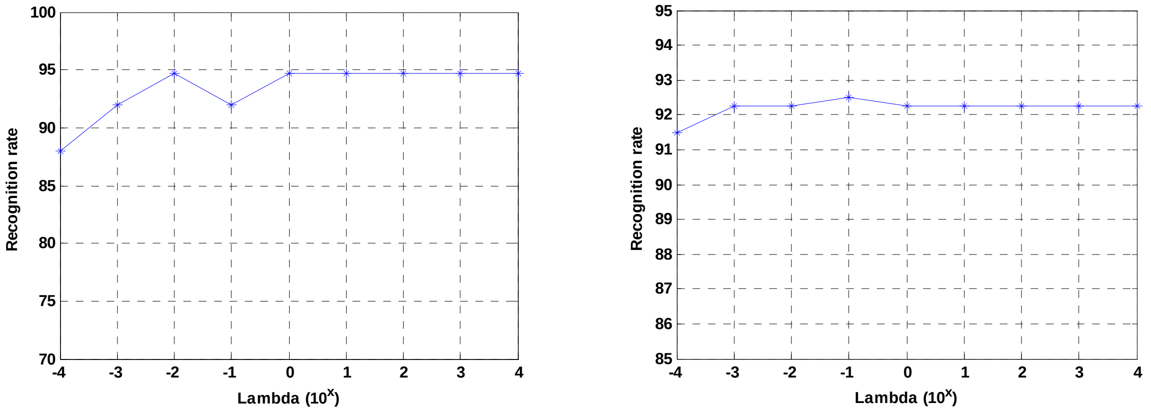
TABLE VI
COMPARISON OF THE PERFORMANCE (RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION) OF DIFFERENT ALGORITHMS ON USPS DATA SET

L	RPCA	NPE	LPP	SPP	IRPCA	LRE
60	87.73±2.13 (27)	89.85±1.83 (49)	86.08±1.72 (50)	73.87±3.71 (31)	77.97±4.44 (47)	**91.72**±2.02 (22)



(a) (b)

Fig.4. (a) The average recognition rates (%) versus the variations of the dimension on (a) ORL face database. (b) USPS database.

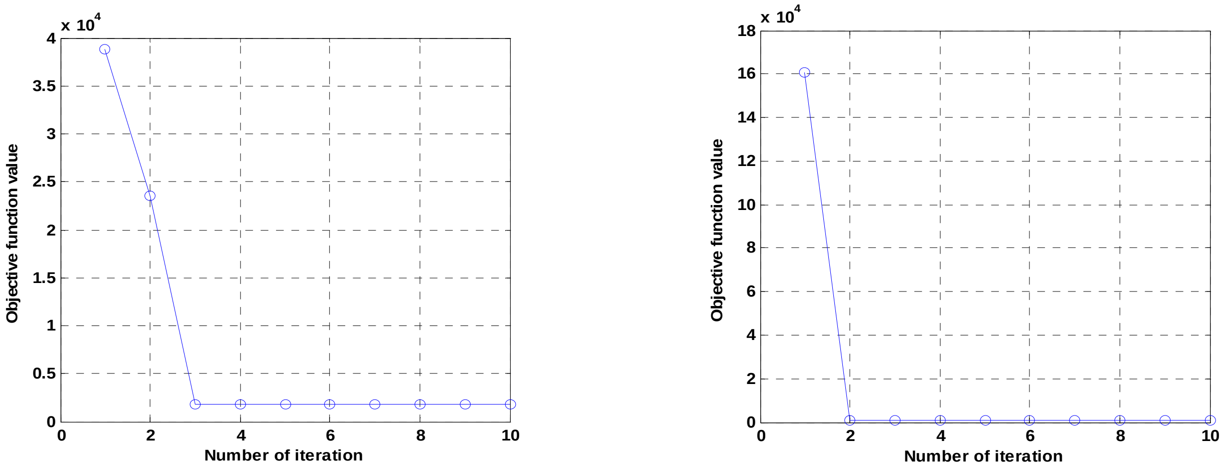


(a) (b)

Fig. 5. The recognition rates vs. the parameter lambda on Yale (a) face database and USPS (b) hand writing digital database of LRE.

parameters determined by the validation set were used for learning the projections for feature extraction and classification. The L was set as different numbers according to the size of each individual/object on the different datasets, i.e., $L = 4,5,6$ for CMU PIE, ORL and Yale face databases, and $L = 60$ for USPS handwriting digital databases, respectively.

For the manifold learning based linear dimensionality reduction methods, i.e. LPP and NPE, since the performance by directly using the occluded data are very poor, we use RPCA as a preprocessing step. The neighborhood parameters in LPP and NPE were selected from the set {1, 2, …,7, 2^3,…, 2^8 } by using the validation set. The numbers of final subspace dimensions for PIE face database were varied from 5 to 150 with step 5. For Yale, ORL and USPS database, the numbers of the final subspace dimensions were varied from 1 to 50 with step 1 since the dimensions corresponding to the best recognition rates are in this range. The regularization parameters of all the regression methods were selected from $[0.001, 0.01, ..., 1000]$ since we use the grid searching strategy to find the optimal value of the parameters in LRE.



(a)

(b)

Fig. 6. The convergence property of the proposed algorithm on (a) Yale database, (b) USPS database

The algorithms were independently run 10 times. In each run, the best parameters determined by the validation set were used to learn the optimal projections for feature extraction. The average recognition rate, standard deviation and the corresponding dimension on the test set are reported in the Tables III-VI. The recognition rates vs. the number of the training images or the size of the block occlusions are also shown in Fig. 3.

C. *Experimental result and analysis*

From the experiments on the occluded databases, we have the following interesting observations and conclusions from the experimental results, including Table III to Table VI.

1. The low rank based methods are robust to a certain extent to the occlusion of the images. LRE has the best performance in different databases with different kinds of occlusion. However, RPCA and IRPCA are not suitable for feature extraction since they have no function of dimensionality reduction. The learned projections of RPCA and IRPCA are the matrix of $m \times m$, where m is the dimension of the data. As we now know, the eigenvectors corresponding to the smaller (or larger) eigenvalues of the eigenfunction are the optimal projection for the comparison methods. However, there is no criterion to decide which projection vector in the $m \times m$ projection matrix P should be used for dimensionality reduction purpose.

2. Although the robust of SPP is reported in previous research, in the four cases investigated in this study, SPP cannot obtain good performance. As shown in Table V, SPP obtains the recognition accuracy as low as 40% to 50%. The key reason, as stated in [46], is that the L_1 norm sparse representation attempts to find the same occlusion image to represent the samples instead of using the images from the same object. In this case, the robustness of the L_1 norm based method is lost. Thus, SPP obtains low recognition rates on the four corruption databases. This indicates that LRE is more robust than SPP and some other compared methods when there are sample specific corruptions on the images.

3. Using the PCA for pre-dimensionality reduction for LPP, NPE and SPP is a good idea to avoid the singularity in the eigenfunction but this cannot obtain good performance when the data are corrupted and there are outliers in each class. However, in the RPCA subspace, LPP and NPE can achieve better performance than in the PCA subspace. This indicates that the RPCA has strong robustness for data representation. In the RPCA subspace, LPP and NPE have similar performance in different occlusion or corruption databases.

4. Different methods have different robustness in the experiments presented in this study. When there are similar image occlusions, the RPCA as a preprocessing step for NPE and LPP has significant influence on their performance, which can be found in Fig. 3 (a). NPE and LPP have similar performances on the RPCA subspace. RPCA has the strong robustness on image occlusion or image subtraction as a pre-processing method for NPE and LPP. The reason is that this step can recover the latent manifold structure such that NPE and LPP can perform well in the RPCA subspace. The proposed LRE essentially integrates the low rank representation and dimensionality reduction together. LRE's strong robustness inherits from the low rank representation so that LER can also learn a robust subspace. This is the key reason for LRE to obtain higher accuracy. The other reason is that LER can simultaneously learn an optimal low rank representation and low-dimensional subspace. However, the combination of RPCA plus LPP and the combination of RPCA plus NPE are not the optimal as a whole. This may potentially degrade their performances.

5. Additional observations in the experiments can be obtained. We find that other unsupervised learning methods, such as k-means or its extensions, achieve very poor results in clustering accuracy. For example, the clustering accuracies of k-means vary from 10% to 20% on these occlusion and corruption datasets. The key reason is that since there are occlusions and corruptions on the images, the k-means method tempts to focus more on the images with the similar occlusions, corruptions and cluster them together. Therefore k-means method obtains the poor results in our experiments.

D. *Parameters sensitivity study*

In the proposed LRE, there are two parameters, namely the parameter d, i.e. the number of dimension, and the balance parameter λ, which are the important parameters in the model. From Fig. 5 (a) and (b), one can find that parameter λ has some effectiveness on different databases. On the Yale, ORL and CMU PIE face databases, the performance of the parameter's variations are very similar, thus we only show the case on Yale face database. One common phenomenon we find in the experiments is that when $\lambda = 1$ the algorithm performs very well in most cases in different databases. Usually, the recognition rate of LRE is very robust to parameter λ when $\lambda \geq 1$, which can be found in Fig. 4 (a). However, the recognition rate of LRE performs best when $\lambda = 0.1$ on USPS database, which can be observed in Fig.5 (b).

E. Convergence study

Theoretical analysis in previous section shows that the objective function of LRE will converge to the local optimum. In fact, in real world applications, it is important for us to show how fast the proposed method will be.

Fig. 6 shows the variation of the objective function value of the proposed LRE on two representative databases, i.e. Yale face database and USPS database. Fig. 6 (a) and (b) show the convergent properties of the LRE on Yale face database and USPS database, respectively. It can be found that LRE converges very fast. Generally, the proposed RDR can converge within 4 to 5 iterations. Thus, to achieve computational effectiveness, one can set the iteration number of the outer loop as 5 or 10.

VI. Conclusion

In this paper, a robust linear dimensionality reduction method named LRE is proposed. LRE uses the low rank representation of the data to explore the intrinsic relationship of the noised data. Since the low rank representation has the strong robustness to the training data's variations or occlusions, the proposed LRE inherits the robustness of the low rank representation to learn an optimal subspace. Since we adopt the alternative iteration strategy, the learned subspace and the low rank representation can be integrated smoothly together so that the whole model is optimal for subspace learning and low rank representation simultaneously. The optimal solutions of the LRE model can be obtained by using the standard eigen-decomposition and the ALM method. Some theoretical analyses are also presented to explore the properties of the proposed method, including the convergence and the low rank coefficient matrix. It is shown that the low rank representation coefficient matrix in the model still lies on the row space of the data and the projection matrix just acts as one of the factor matrices to represent the low rank matrix. Experiments on four well-known databases with strong sample-specific corruptions, occlusions or block subtractions show that the proposed LRE has strong robustness in feature extraction. Similar to [15], it is interesting to point out that the proposed method can also be used for video-based moving object detection, which may take advantage of the robustness from the low rank representation of the video frames.

References

[1] M. Turk, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

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

[3] F. Song, D. Zhang, D. Mei, and Z. Guo, "A multiple maximum scatter difference discriminant criterion for facial feature extraction," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 37, no. 6, pp. 1599–1606, 2007.

[4] W. Zuo, D. Zhang, J. Yang, and K. Wang, "BDPCA plus LDA:a novel fast feature extraction technique for face recognition," *IEEE Trans. Syst. Man, Cybern. Part B, Cybern.*, vol. 36, no. 4, pp. 946–952, 2006.

[5] C. Fraley and Adrian E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *J. Am. Stat. Assoc.*, vol. 97, no. 1, 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, 2004.

[7] Y. Pang, S. Member, S. Wang, Y. Yuan, and S. Member, "Learning regularized LDA by clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2191–2201, 2014.

[8] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, vol. 32, pp. 1062–1070.

[9] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2328–2335.

[10] 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.

[11] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition.," *IEEE Trans. image Process.*, vol. 15, no. 11, pp. 3608–14, Nov. 2006.

[12] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition," *Pattern Recognit.*, vol. 40, no. 1, pp. 339–342, 2007.

[13] W. K. Wong and H. T. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognit.*, vol. 45, no. 1, pp. 186–197, Jan. 2012.

[14] F. Zhong, J. Zhang, and D. Li, "Discriminant locality preserving projections based on L1-norm maximization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2065–2074, 2014.

[15] M. T. G. Krishna, V. N. M. Aradhya, M. Ravishankar, and D. R. R. Babu, "LoPP: Locality preserving projections for moving object detection," *Procedia Technol.*, vol. 4, pp. 624–628, 2012.

[16] X. He, D. Cai, S. Yan, and H. Zhang, "Neighborhood preserving embedding," in *IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 1208–1213.

[17] 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.

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

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

[20] 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.

[21] 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.

[22] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with L1 graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

[23] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 1–37, May 2011.

[24] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–84, Jan. 2013.

[25] G. Liu and Z. Lin, "Robust subspace segmentation by low-rank representation," in *Proceedings of International Conference on Machine Learning*, 2010.

[26] J. Liu, Y. Chen, J. Zhang, and Z. Xu, "Enhancing low-rank subspace clustering by manifold regularization," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4022–4030, Jul. 2014.

[27] K. Tang, R. Liu, Z. Su, and J. Zhang, "Structure-constrained low-rank representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2167–2179, 2014.

[28] C. Li, X. Qi, and J. Guo, "Dimensionality reduction by low-rank embedding," *Lect. Notes Comput. Sci.*, vol. 7751, pp. 181–188, 2013.

[29] B. Kulis, A. C. Surendran, and J. C. Platt, "Fast low-rank semidefinite programming for embedding and clustering," *Int. Conf. Artif. Intell. Stat.*, pp. 235–242, 2007.

[30] V. M. Patel, H. Van Nguyen, and R. Vidal, "Latent space sparse subspace clustering," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 225–232, 2013.

[31] A. Bhardwaj and S. Raman, "Robust PCA-based solution to image composition using augmented Lagrange multiplier (ALM)," *Vis. Comput.*, vol. 32, no. 5, pp. 591–600, 2016.

[32] Q. Yao and J. T. Kwok, "Colorization by patch-based local low-rank matrix completion," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015, pp. 1959–1965.

[33] C. Guyon, T. Bouwmans, and E. H. Zahzah, "Foreground detection via robust low rank matrix decomposition including spatio-temporal constraint," in *International Workshop on Background Model Challenges, ACCV 2012*, 2012, vol. 7728 LNCS, no. PART 1, pp. 315–320.

[34] Y. Xue, X. Guo, and X. Cao, "Motion saliency detection using low-rank and sparse decomposition," in *Spectrum*, 2012, pp. 1485–1488.

[35] H. Ji, S.-B. Huang, Z. Shen, and Y. Xu, "Robust video restoration by joint sparse and low rank matrix approximation," *SIAM J. Imaging Sci.*, vol. 4, no. 4, pp. 1122–1142, 2011.

[36] T. Bouwmans and E. H. Zahzah, "Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance," *Comput. Vis. Image Underst.*, vol. 122, pp. 22–34, 2014.

[37] B.-K. Bao, G. Liu, C. Xu, and S. Yan, "Inductive robust principal component analysis," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3794–3800, Aug. 2012.

[38] B.-K. Bao, G. Liu, R. Hong, S. Yan, and C. Xu, "General subspace learning with corrupted training data via graph embedding," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4380–4393, Nov. 2013.

[39] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14*, 2001, pp. 585–591.

[40] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 12, pp. 2268–2269, 2000.

[41] 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.

[42] F. Nie, H. Huang, X. Cai, and Chris Ding, "Efficient and robust feature selection via joint L2,1 norms minimization," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 1813–1821.

[43] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[44] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multihannel image restoration," *SIAM J. Imaging Sci.*, vol. 2, no. 2, pp. 569–592, 2009.

[45] 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, 2003.

[46] Z. Zheng, M. Yu, J. Jia, H. Liu, D. Xiang, X. Huang, and J. Yang, "Fisher discriminantion based low rank matrix recovery for face recognition," *Pattern Recognit.*, vol. 47, no. 11, pp. 3502–3511, Nov. 2014.

W.K. Wong received his Ph.D. degree from The Hong Kong Polytechnic University. Currently, he is a professor in this university. He has published over ninety scientific articles in refereed journals, including IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Image Processing, IEEE Transactions on Cybernetics, among others. His recent research interests include pattern recognition, feature extraction image retrieval and defect detection.

Zhihui Lai received the B.S. degree in mathematics from South China Normal University, M.S. degree from Jinan University, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2002, 2007 and 2011, respectively. He has been a Research Associate, Postdoctoral Fellow and Research Fellow at The Hong Kong Polytechnic University. His 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. He has published over 60 scientific articles. Now he is an associate editor of International Journal of Machine Learning and Cybernetics. For more information including all papers and related codes, the readers are referred to the website (http://www.scholat.com/laizhihui).

Jiajun Wen received the Ph.D. degree in computer science and technology from Harbin Institute of Technology, China, in 2015. He has been a Research Associate with the Hong Kong Polytechnic University, Hong Kong, since 2013. He is currently a Postdoctoral Fellow with the College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China. His research interests include pattern recognition and video analysis.

Xiaozhao Fang received his M.S. degree in 2008, and the Ph.D. degree in computer science and technology at Shenzhen Graduate School, HIT, Shenzhen (China) in 2016. He is currently with the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include pattern recognition and machine learning.

Yuwu Lu received the B.S. degree in mathematics from the XingTai University in 2008 and the M.S. degree in mathematics from the Inner Mongolia University of Technology in 2011, and the Ph.D. degree in computer science and technology from the Harbin Institute of Technology in 2015. He is a Post-Doctoral Fellow with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University. He is the author of more than 15 scientific papers in pattern recognition and computer vision. His current research interests include pattern recognition and machine learning.