

Generalized Robust Regression for Jointly Sparse Subspace Learning

Zhihui Lai^{ID}, Dongmei Mo^{ID}, Jiajun Wen, Linlin Shen, and Wai Keung Wong

Abstract—Ridge regression is widely used in multiple variable data analysis. However, in very high-dimensional cases such as image feature extraction and recognition, conventional ridge regression or its extensions have the small-class problem, that is, the number of the projections obtained by ridge regression is limited by the number of the classes. In this paper, we proposed a novel method called generalized robust regression (GRR) for jointly sparse subspace learning which can address the problem. GRR not only imposes $L_{2,1}$ -norm penalty on both loss function and regularization term to guarantee the joint sparsity and the robustness to outliers for effective feature selection, but also utilizes $L_{2,1}$ -norm as the measurement to take the intrinsic local geometric structure of the data into consideration to improve the performance. Moreover, by incorporating the elastic factor on the loss function, GRR can enhance the robustness to obtain more projections for feature selection or classification. To obtain the optimal solution of GRR, an iterative algorithm was proposed and the convergence was also proved. Experiments on six well-known data sets demonstrate the merits of the proposed method. The result indicates that GRR is a robust and efficient regression method for face recognition.

Index Terms—Ridge regression, face recognition, feature selection, subspace learning, small-class problem.

I. INTRODUCTION

AS THE widely-used statistical analysis technique, Least Squares Regression (LSR) [1] has been utilized in many practical applications, such as face recognition [2], video-based gait recognition [3]. However, in the multicollinearity problem, the estimate of LSR is unbiased, this would possibly

Manuscript received October 14, 2017; accepted February 20, 2018. This work was supported in part by the Natural Science Foundation of China under Grant 61573248, Grant 61773328, Grant 061773328, and Grant 61703283, in part by the Research Grant of The Hong Kong Polytechnic University under Project G-UA2B, in part by the China Postdoctoral Science Foundation under Project 2016M590812 and Project 2017T100645, in part by the Guangdong Natural Science Foundation under Project 2017A030313367 and Project 2017A030310067, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20170302153434048. This paper was recommended by Associate Editor N. Boulgouris. (*Corresponding author: Wai Keung Wong*.)

Z. Lai, D. Mo, 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 and Clothing, The Hong Kong Polytechnic University, Hong Kong (e-mail: lai_zhi_hui@163.com; dongmei_mo@qq.com; wenjiajun.hit@gmail.com).

L. Shen is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: llshen@szu.edu.cn).

W. K. Wong is with the Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong, and also with the Shenzhen Research Institute, The Hong Kong Polytechnic University, Shenzhen 518057, China (e-mail: calvin.wong@polyu.edu.hk).

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

Digital Object Identifier 10.1109/TCSVT.2018.2812802

make the result far from the true value when their variances are large [4]. The ridge regression (RR) is a regularized least square method which adds a bias term in the conventional LSR to reduce the standard errors to improve the performance of regression estimates [4]. Many extensions based on LSR or RR are applied to dimensionality reduction and feature extraction.

The classical dimensionality reduction method is Principle Component Analysis (PCA) which solves the eigen decomposition problem to obtain the optimal vectors for dimensionality reduction [5]. Similar to PCA, Linear Discriminant Analysis (LDA) is another famous dimensionality reduction method. Different from PCA, LDA is a supervised method which uses label information in the computational procedure to learn an optimal projection matrix. The projections learned from LDA not only maximize the between-class distance but also minimize the within-class distance in the feature space so as to improve the performance for pattern recognition [6]. Though PCA and LDA are the famous dimensionality reduction techniques, they still have some disadvantages. Firstly, they just take the global structure of the data set into consideration and ignore the local geometric information. This would affect the performance as the locality is of fundamental importance in dimensionality reduction or feature selection [7]. Secondly, as the L_2 -norm based methods, traditional PCA and LDA are sensitive to outliers because the squared residual in L_2 -norm would lead to the undesirable tendency of over-emphasizing the noise and outliers in computing the projection matrix. In addition, for LDA or the LDA-based methods, the number of the projections is limited by the between-class scatter matrix (i.e. the Small Sample Size (SSS) problem) [8]. This limitation would degrade the performance in feature extraction and classification.

To alleviate the first problem mentioned above, many locality (i.e. neighborhood preserving property) based methods were proposed to promote the effectiveness in computer vision and pattern recognition [9]. Among them, the representative nonlinear dimensionality reduction algorithms include Locally Linear Embedding (LLE) [10], Laplacian Eigenmap [11], Isomap [12] and so on. The well-known liner version of locality based methods include Locality Preserving Projection (LPP) [13], Laplacianfaces [14] and linear versions of LLE (i.e. Neighborhood Preserving Projection (NPP) [15], Neighborhood Preserving Embedding (NPE) [16], etc.) and so on.

For the second problem mentioned above, since the methods using L_2 -norm on the loss function are sensitive to outliers,

an alternative resolution is to use L_1 -norm to take place of L_2 -norm and lots of experimental results have proved that L_1 -norm is more robust to outliers than L_2 -norm [17], [18]. Some L_1 -norm methods based on PCA have been proposed to enhance the robustness to outliers for face recognition. One of the popular methods is Robust Principal Component Analysis (RPCA) [19] which can recover both low-rank and sparse components by solving Principal Component Pursuit problem using an augmented Lagrange multiplier algorithm [20]. Other methods include L1-PCA [21], R1-PCA [22], PCA-L1 [23], etc. The proposed L1-PCA and R1-PCA are complicated and computationally expensive while PCA-L1 is fast and robust [6]. Motivated by R1-PCA, a LDA-based method called Linear Discriminant Analysis using rotational invariant L_1 -norm (LDA-R1) [24] has been proposed to improve the robustness for dimensionality reduction and feature selection. Motivated by the PCA-L1, another LDA-based method called Linear Discriminant Analysis based on L_1 -norm maximization (LDA-L1) [6] has been proposed to overcome the drawback in LDA-R1 that it takes too much time to achieve convergence in the high dimension case [6].

Recently, sparse learning methods have attracted much attention in image processing, face recognition [25], [26] and pattern recognition [27]. The $L_{2,1}$ -norm is recently becoming popular because it is an efficient way to obtain the jointly sparse projection for discriminative feature selection or extraction. By imposing joint $L_{2,1}$ -norm minimization on both the loss function and the regularization term, Nie et al. proposed a method called efficient and Robust Feature Selection via joint $L_{2,1}$ -norms minimization (RFS) [28] for jointly sparse feature selection. Motivated by [28], the $L_{2,1}$ -norm minimization is also extended to the study of sparse regression, subspace learning [29], [30]. Another jointly sparse method called $L_{2,1}$ -norm regularized discriminative feature selection for unsupervised learning (UDFS) [31] has also been proposed to improve the performance for unsupervised learning. Though many $L_{2,1}$ -norm based methods have been proposed to deal with different cases, their theoretical relationship with the sparse regression and subspace learning is still unclear. In addition, there are still some drawbacks in the current $L_{2,1}$ -norm based methods. For example, RFS utilizes the $L_{2,1}$ -norm on both loss function and regularization term to obtain joint sparsity for feature selection. However, it just focuses on global structure of the data set and meanwhile ignores the locality of the data. Moreover, the small-class problem in RFS remains unsolved. For some other methods, though they take the local geometric structure of the data into consideration, the learned locality preserving projections are not robust to outliers as the conventional LPP based terms incorporate the L_2 -norm as the measurement on the objective function. Besides, the robustness of these methods is not guaranteed in different cases, especially when the data set is corrupted by strong noise. Therefore, a more robust and efficient method is necessary to be proposed to improve the performance for face recognition or other applications.

In this paper we propose a novel method which integrates the properties of LPP, RR and $L_{2,1}$ -norm to alleviate the potential problems in RR and its extensions. This method is

capable to address the small-class problem in the regression-based methods and simultaneously guarantee the robustness and effectiveness for face recognition. The main contributions of this work are described as below:

1. Unlike the ridge regression based methods, the proposed method can break out the small-class problem that the number of the projections is limited by the number of class. Thus, our method can obtain more projections to perform feature extraction and get higher accuracy than previous methods.
2. The proposed method not only imposes the joint $L_{2,1}$ -norm minimization on both the locality term and the loss function to guarantee the robustness to outliers, but also use $L_{2,1}$ -norm as the penalty on the regularization term to ensure the joint sparsity of the learned projection for discriminative feature selection. It makes the proposed model differ from all the existing LPP-based methods, ridge-regression based methods or the previous $L_{2,1}$ -norm based methods.
3. The proposed method incorporates the elastic factor to the loss function to avoid over-fitting problem and thus can further guarantee the model's stability. Based on the compact model, the proposed method is able to enhance the robustness while dealing with complicated cases, especially when the data set is corrupted by strong noise. Besides, the convergence of the proposed iterative algorithm is also proved.

The rest of the paper is organized as follows. We first present some notations and then propose the novel method and its corresponding iterative algorithm in Section II. Section III shows theoretical analyses, including the convergence of the proposed method and its computational complexity. In Section IV, we perform a series of experiments to evaluate the performance of the proposed method and then draw a conclusion in Section V.

II. GENERALIZED ROBUST REGRESSION

In this section, a model called Generalized Robust Regression (GRR) for jointly sparse subspace learning will be presented and the alternatively iterative algorithm is designed to solve the optimization problem.

A. Notations

Scalars are denoted as lowercase italic letters, i.e. i , j , n , d , c , m , etc. while vectors are represented as bold lowercase italic letters, i.e. \mathbf{x} , \mathbf{y} , \mathbf{v} , etc.. Matrices are defined as bold uppercase italic letters, i.e. \mathbf{X} , \mathbf{Y} , \mathbf{A} , \mathbf{B} , etc.

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denotes the training sample matrix, where n is the number of total training samples and d denotes the feature dimension of each sample, each row of \mathbf{X} is a sample \mathbf{x}_i . Let $\mathbf{Y} \in \mathbb{R}^{n \times c}$ denotes the label matrix, where c is the total number of classes. The matrix \mathbf{Y} is defined as a binary label matrix with $Y_{ij} = 1$ while \mathbf{x}_i belongs to the j -th class; $Y_{ij} = 0$, otherwise.

B. The Motivations and the Novel Definitions

There are some obvious disadvantages in Least square regression (LSR) [1] and Ridge Regression (RR) [4].

First, they have small-class problem. That is, when the number of the class in the training data is too small, only c projections can be obtained to perform feature extraction and selection. Second, since LSR and RR are the L_2 -norm based methods, the square operation on the objective function will lead to sensitivity to outliers. Third, since the learned projection from traditional ridge regression is not sparse, the projections learned from LSR or RR have no sparse property for feature selection. Thus, a more robust loss function is demanded for feature selection. Nie *et al.* proposed a method called Efficient and Robust Feature Selection via joint $L_{2,1}$ -norms minimization (RFS) [28]. The objective function of RFS is

$$\min_{\mathbf{P}} \|\mathbf{X}\mathbf{P} - \mathbf{Y}\|_{2,1} + \gamma \|\mathbf{P}\|_{2,1} \quad (1)$$

By utilizing $L_{2,1}$ -norm on both loss function and regularization term, RFS is able to release the drawback in ridge regression for its sensitivity to outliers. In addition, the regularization term in RFS guarantees the joint sparsity to improve the performance for feature selection and face recognition. But there are still some drawbacks in RFS. Firstly, it still has the small-class problem. Secondly, since the projections learned by RFS are just the liner combinations of the global structure of the data points, the local geometry of the data set is ignored. However, lots of experimental results indicate that preserving the locality tends to improve the performance in feature extraction and classification [14]. Therefore, new technique is needed to deal with these problems.

In this paper we propose a generalized robust regression method for jointly sparse subspace learning. This method not only inherits the advantages in RFS, but also integrates the property of LPP, RR to further improve the performance for feature selection. Namely, it utilizes $L_{2,1}$ -norm on the loss function to minimize the squared operation errors and simultaneously use $L_{2,1}$ -norm minimization on the regularization term to guarantee the joint sparsity for discriminative feature selection. Moreover, it releases the small-class problem and at the same time takes the local geometric structure into consideration. It also imposes the $L_{2,1}$ -norm penalty on the locality preserving projection term to ensure the robustness to outliers. In addition, to improve the robustness of the proposed method, the elastic factor is incorporated to the loss function for jointly sparse subspace learning.

C. The Objective Function of GRR

The objective function of GRR is to minimize the $L_{2,1}$ -norm based optimization problem with some constraints:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}, \mathbf{h}} & \sum_i \sum_j \|\mathbf{x}_i \mathbf{B} \mathbf{A}^T - \mathbf{x}_j \mathbf{B} \mathbf{A}^T\|_2 \mathbf{W}_{ij} + \beta \|\mathbf{B}\|_{2,1} \\ & + \gamma \|\mathbf{X} \mathbf{B} \mathbf{A}^T + \mathbf{1} \mathbf{h}^T - \mathbf{Y}\|_{2,1} + \lambda \|\mathbf{h}\|_2^2 \\ \text{s.t. } & \mathbf{A}^T \mathbf{A} = \mathbf{I} \end{aligned} \quad (2)$$

where $\mathbf{B} \in \mathbb{R}^{d \times k}$ is the projection matrix, $\mathbf{A} \in \mathbb{R}^{c \times k}$ is an orthogonal matrix, d and k is the number of matrix size while c is the number of class. $\mathbf{W} \in \mathbb{R}^{n \times n}$ is the similarity matrix as defined in LPP. $\mathbf{1} \in \mathbb{R}^{n \times 1}$ is the vector with all elements equaling to 1. The vector $\mathbf{h} \in \mathbb{R}^{c \times 1}$ is the bias term and the three

coefficients β , γ and λ are parameters to balance different terms. Note that the bias term \mathbf{h} was used in some previous semi-supervised algorithms, i.e. LapRLS/L [32], FME [33], etc.. The proposed method extends this bias term as the elastic factor in the generalized regression to improve the robustness for face recognition, especially when the data points are corrupted by strong noise.

In (2), the first part $\sum_i \sum_j \|\mathbf{x}_i \mathbf{B} \mathbf{A}^T - \mathbf{x}_j \mathbf{B} \mathbf{A}^T\|_2 \mathbf{W}_{ij}$ aims at locality preserving property [13]. Instead of computing the Euclidean distance between each training sample \mathbf{x}_i ($i = 1, 2, \dots, n$) and \mathbf{x}_j ($j = 1, 2, \dots, n$) which is sensitive to outlier while preserving local information, the proposed method uses $L_{2,1}$ -norm as the measurement to enhance robustness on the locality preserving ability. By inheriting the locality preserving property, the proposed GRR not only preserves the intrinsic local geometric structure of the data [13], but also guarantees the robustness to outliers.

The second part in (2) is the regularization term $\beta \|\mathbf{B}\|_{2,1}$, which guarantees that the learned projection matrix \mathbf{B} is jointly sparse for discriminative feature selection [28], [31]. The joint sparsity ensures that most elements of the learned projections are zero and the important features are selected for feature extraction.

In (2), the third part $\gamma \|\mathbf{X} \mathbf{B} \mathbf{A}^T + \mathbf{1} \mathbf{h}^T - \mathbf{Y}\|_{2,1}$ is the loss function as in classical RR and the fourth part $\lambda \|\mathbf{h}\|_2^2$ serves as the bias term to guarantee the stability of the whole model. Comparing with RR, (2) using $L_{2,1}$ -norm minimization on the loss function makes the model more robust to outliers [28]. Another potential reason of the robustness of GRR is that the elastic factor \mathbf{h} on the loss function can avoid the overfitting problem in practice. On the loss function of RFS, the matrix \mathbf{P} must always be fitting for \mathbf{Y} so as to ensure that the error between the matrix $\mathbf{X}\mathbf{P}$ and the label matrix \mathbf{Y} can be minimized, which would lead to the potential risk of the overfitting problem. However, GRR imposes the elastic factor \mathbf{h} as the supplement term on the loss function and the matrix $\mathbf{X} \mathbf{B} \mathbf{A}^T$ is not strictly needed to fit the matrix \mathbf{Y} so as to release the overfitting problem to guarantee the strong generalization ability for feature selection or extraction, especially in the case when the images are corrupted by block subtraction or noise.

Moreover, by using the matrix $\mathbf{B} \mathbf{A}^T$ on the loss function instead of the \mathbf{P} in (1), (2) is designed to address the small-class problem in the LSR, RR and RFS. That is, the size of \mathbf{P} is $d \times c$ while the size of \mathbf{B} , \mathbf{A} is $d \times k$ and $c \times k$ respectively, then the size of $\mathbf{B} \mathbf{A}^T$ is $d \times c$ (i.e. $\mathbf{B} \mathbf{A}^T$ has the same size with \mathbf{P}). In LSR, RR and RFS, the projection matrix is \mathbf{P} and the number of the learned projections is c (i.e. the number of the class). However, in GRR, the learned projection matrix is \mathbf{B} with the size $d \times k$ and k can be set as any integer to obtain enough projections to perform face recognition. Therefore, the number of the projection in the proposed GRR is not limited by the number of class and the small-class problem in RR can be addressed.

D. The Optimal Solution

According to the definition of the $L_{2,1}$ -norm on projection matrix \mathbf{B} , a diagonal matrix $\tilde{\mathbf{D}}$ with the i -th diagonal element

299 can be defined as [28]:

$$300 \quad \tilde{D}_{ii} = \frac{1}{2 \|\mathbf{B}^i\|_2} \quad (3)$$

301 where \mathbf{B}^i denotes the i -th row of matrix \mathbf{B} . Thus the second
302 part in (2) is rewritten as

$$303 \quad \|\mathbf{B}\|_{2,1} = \text{tr}(\mathbf{B}^T \tilde{\mathbf{D}} \mathbf{B}) \quad (4)$$

304 Similarly, the third part in (2) is rewritten as

$$305 \quad \|\mathbf{XBA}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y}\|_{2,1} \\ 306 \quad = \text{tr}((\mathbf{XBA}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})^T \hat{\mathbf{D}} (\mathbf{XBA}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})) \quad (5)$$

307 where $\hat{\mathbf{D}}$ is also a diagonal matrix with the i -th diagonal
308 element as

$$309 \quad \hat{D}_{ii} = \frac{1}{2 \|(XBA^T + 1h^T - Y)^i\|_2} \quad (6)$$

310 For the first part in (2), since we change the square of
311 Euclidean norm to be the $L_{2,1}$ -norm, thus in order to utilize
312 the property of LPP, we reformulate it as follow:

$$313 \quad \sum_i \sum_j \|\mathbf{x}_i \mathbf{BA}^T - \mathbf{x}_j \mathbf{BA}^T\|_2 W_{ij} \\ 314 \quad = \sum_i \sum_j \frac{\|\mathbf{x}_i \mathbf{BA}^T - \mathbf{x}_j \mathbf{BA}^T\|_2^2}{\|\mathbf{x}_i \mathbf{BA}^T - \mathbf{x}_j \mathbf{BA}^T\|_2} W_{ij} \\ 315 \quad = \sum_i \sum_j \|\mathbf{x}_i \mathbf{BA}^T - \mathbf{x}_j \mathbf{BA}^T\|_2^2 W_{ij} / G_{ij} \quad (7)$$

316 where $G_{ij} = \|\mathbf{x}_i \mathbf{BA}^T - \mathbf{x}_j \mathbf{BA}^T\|_2$. Thus, we have

$$317 \quad \sum_i \sum_j \|\mathbf{x}_i \mathbf{BA}^T - \mathbf{x}_j \mathbf{BA}^T\|_2 W_{ij} \\ 318 \quad = \text{tr}(\mathbf{B}^T X^T (\mathbf{D} - \mathbf{W} \emptyset \mathbf{G}) \mathbf{X} \mathbf{B}) \quad (8)$$

319 where \emptyset is the element-wise deviation of matrices and \mathbf{D} is
320 a diagonal matrix and its elements are row (or column) sum
321 of $\mathbf{W} \emptyset \mathbf{G}$, namely, $D_{ii} = \sum_i (\mathbf{W} \emptyset \mathbf{G})_{ij}$.

322 From (4), (5) and (8), the objective function (2) is equal to
323 the following function:

$$324 \quad \min_{\mathbf{A}, \mathbf{B}, \mathbf{h}} [\text{tr}(\mathbf{B}^T X^T (\mathbf{D} - \mathbf{W} \emptyset \mathbf{G}) \mathbf{X} \mathbf{B}) + \beta \text{tr}(\mathbf{B}^T \tilde{\mathbf{D}} \mathbf{B}) \\ 325 \quad + \gamma \text{tr}((\mathbf{XBA}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})^T \hat{\mathbf{D}} (\mathbf{XBA}^T + \mathbf{1}\mathbf{h}^T - \mathbf{Y})) \\ 326 \quad + \lambda \text{tr}(\mathbf{h}^T \mathbf{h})] \\ 327 \quad \text{s.t. } \mathbf{A}^T \mathbf{A} = \mathbf{I} \quad (9)$$

328 In order to obtain the local optimal solution of GRR, we fix
329 the two variables \mathbf{A} , \mathbf{B} and set the derivatives of (9) with
330 respect to \mathbf{h} equaling to zero, then we have

$$331 \quad \mathbf{h} = \frac{1}{s} (\mathbf{Y}^T \hat{\mathbf{D}} \mathbf{1} - \mathbf{A} \mathbf{B}^T X^T \hat{\mathbf{D}} \mathbf{1}) \quad (10)$$

332 where $s = \mathbf{1}^T \hat{\mathbf{D}} \mathbf{1} + \lambda I$.

333 Similarly, for fixed \mathbf{A} and \mathbf{h} , we set the derivatives of (9)
334 with respect to \mathbf{B} equaling to zero, then (9) is minimized by

$$335 \quad \mathbf{B} = \gamma [\beta \tilde{\mathbf{D}} + X^T ((\mathbf{D} - \mathbf{W} \emptyset \mathbf{G}) + \gamma \hat{\mathbf{D}}) \mathbf{X}]^{-1} X^T \hat{\mathbf{D}} (\mathbf{Y} - \mathbf{1}\mathbf{h}^T) \mathbf{A} \\ 336 \quad (11)$$

In (9), when the two variables \mathbf{B} and \mathbf{h} are fixed, the following maximization problem about \mathbf{A} provides the optimal solution.

$$340 \quad \max_{\mathbf{A}} \mathbf{A}^T (\mathbf{h} \mathbf{1}^T - \mathbf{Y}^T) \hat{\mathbf{D}} \mathbf{X} \mathbf{B} \\ 341 \quad \text{s.t. } \mathbf{A}^T \mathbf{A} = \mathbf{I} \quad (12)$$

The optimal solution in (12) can be obtained from the following theorem:

Theorem 1 [5]: Let \mathbf{S} and \mathbf{ZZ} be $c \times k$ matrices and \mathbf{ZZ} has rank k . Given the following optimization problem

$$346 \quad \hat{\mathbf{S}} = \arg \max_{\mathbf{S}} \text{Tr}(\mathbf{S}^T \mathbf{Z}) \text{ s.t. } \mathbf{S}^T \mathbf{S} = \mathbf{I}_k$$

Suppose the SVD of \mathbf{ZZ} is $\mathbf{Z} = \check{\mathbf{U}} \check{\mathbf{D}} \check{\mathbf{V}}^T$, then $\hat{\mathbf{S}} = \check{\mathbf{U}} \check{\mathbf{V}}^T$.

From Theorem 1, we can know that for given \mathbf{B} and \mathbf{h} in (12), suppose the SVD of $(\mathbf{h} \mathbf{1}^T - \mathbf{Y}^T) \hat{\mathbf{D}} \mathbf{X} \mathbf{B}$ is $(\mathbf{h} \mathbf{1}^T - \mathbf{Y}^T) \hat{\mathbf{D}} \mathbf{X} \mathbf{B} = \mathbf{U} \check{\mathbf{D}} \mathbf{V}^T$, then

$$351 \quad \mathbf{A} = \mathbf{U} \mathbf{V}^T \quad (13)$$

To obtain the local optimal solution of the objective function, details of the iterative algorithm are represented in Table I. The convergence of the proposed algorithm will be proved in the next section.

E. Comparison and Discussion

From the above subsections, we can conclude the main differences between GRR and previous methods. Comparing with the conventional ridge regression or the ridge-regression-based methods, the proposed method can easily address the small-class problem to obtain enough projections to perform feature selection and extraction. Yang et al. proposed UDFS [31] by imposing the $L_{2,1}$ -norm minimization on the regularization term to obtain discriminative feature subset from the whole feature set for unsupervised learning [31]. Nie et al. proposed RFS [28] by imposing $L_{2,1}$ -norm minimization on both the loss function and the regularization term to guarantee the joint feature selection function and the robustness to outliers. All these previous $L_{2,1}$ -norm based methods have obtained the favorable performance in some degree. However, most of these methods do not take the local geometric structure into consideration for feature selection. In contrast, the proposed method not only uses the joint $L_{2,1}$ -norm minimization on loss function and regularization term as the basic measurement to guarantee the joint sparsity and robustness, but also takes the local geometric structure of the intrinsic data into consideration by incorporating the locality preserving property on the objective function. Additionally, by replacing the L_2 -norm on the locality term with the $L_{2,1}$ -norm, GRR is more robust to outliers than the conventional LPP-based methods. Some other LPP-based methods replace the L_2 -norm with L_1 -norm on the objective function and also obtain good performance in face recognition. Both LPP-L1 [7] and DLPP-L1 [34] use L_1 -norm instead of L_2 -norm in the locality term based on LPP [13] and DLPP [35] respectively. Low-Rank Preserving Projections (LRPP) utilizes $L_{2,1}$ -norm as a sparse constraint on the noise matrix to

TABLE I
GRR ALGORITHM

Input: The training data $\mathbf{X} \in R^{n \times d}$, the training data labels $\mathbf{Y} \in R^{n \times c}$, the symmetric matrix $\mathbf{W} \in R^{n \times n}$, the objective dimension k ($k = 1, 2, \dots, d$), the parameters β , γ and λ , the maximum number of the iteration: $maxStep$.

Output: Low-dimensional discriminative subspace $\mathbf{B} \in R^{d \times k}$, $k = 1, 2, \dots, d$

1: Initialize $\mathbf{A} \in R^{c \times k}$, $\mathbf{B} \in R^{d \times k}$, $\tilde{\mathbf{D}} \in R^{d \times d}$, $\hat{\mathbf{D}} \in R^{n \times n}$, $\mathbf{h} \in R^{c \times 1}$ randomly, initialize $\mathbf{1} \in R^{n \times 1}$ with each element integer 1.

Compute $\mathbf{G} \in R^{n \times n}$, $\mathbf{D} \in R^{n \times n}$ respectively. Set $step = 1$, $converged = \text{false}$.

2: While $\sim converged$ $\&\&$ $step \leq maxStep$

- Compute \mathbf{h} using $\mathbf{h} = \frac{1}{s}(\mathbf{Y}^T \hat{\mathbf{D}} \mathbf{1} - \mathbf{A} \mathbf{B}^T \mathbf{X}^T \hat{\mathbf{D}} \mathbf{1})$, where $s = \mathbf{1}^T \hat{\mathbf{D}} \mathbf{1} + \lambda \mathbf{I}$.
- Compute \mathbf{B} using $\mathbf{B} = \gamma[\beta \tilde{\mathbf{D}} + \mathbf{X}^T ((\mathbf{D} - \mathbf{Z}) + \gamma \hat{\mathbf{D}}) \mathbf{X}]^{-1} \mathbf{X}^T \hat{\mathbf{D}} (\mathbf{Y} - \mathbf{1} \mathbf{h}^T) \mathbf{A}$.
- Compute \mathbf{A} using $\mathbf{A} = \mathbf{U} \mathbf{V}^T$ in (13).
- Update \mathbf{G} using $\mathbf{G}_{ij} = \left\| \mathbf{x}_i \mathbf{B} \mathbf{A}^T - \mathbf{x}_j \mathbf{B} \mathbf{A}^T \right\|_2$.
- Update \mathbf{D} using $\mathbf{D}_{ii} = \sum_i (\mathbf{W} \otimes \mathbf{G})_{ij}$.
- Update $\tilde{\mathbf{D}}$ using $\tilde{\mathbf{D}}_{ii} = \frac{1}{2 \|\mathbf{B}'\|_2}$.
- Update $\hat{\mathbf{D}}$ using $\hat{\mathbf{D}}_{ii} = \frac{1}{2 \left\| (\mathbf{X} \mathbf{B} \mathbf{A}^T + \mathbf{1} \mathbf{h}^T - \mathbf{Y})' \right\|_2}$.
- Set $step = step + 1$.
- Update $converged = \text{true}$ when \mathbf{B} is approximately changeless.

End

3: Standardize the matrix \mathbf{B} to a final normalized matrix and return it for feature selection.

388 perform dimensionality reduction [36]. Comparing with these
389 methods, the advantage of GRR is that it utilizes $L_{2,1}$ -norm
390 minimization on both loss function and regularization term to
391 guarantee the joint feature selection function and at the same
392 time enhance the robustness to outliers. Another contribution
393 of GRR is that it incorporates the elastic factor to the loss
394 function to improve the robustness when the data sets are
395 corrupted by noise or outliers.

396 In short, GRR is a novel and generalized robust regression
397 method for jointly sparse subspace learning. By incorporating
398 the $L_{2,1}$ -norm penalty on the loss function, regularization term
399 and the locality term, GRR can easily obtain the joint sparsity
400 for discriminative feature selection and meanwhile improve
401 the robustness to outliers. Additionally, GRR also addresses
402 the small-class problem in the conventional regression-based
403 methods. Moreover, GRR improves the robustness for jointly
404 sparse subspace learning by incorporating the elastic factor
405 on the loss function to decrease the negative influence when
406 the data is corrupted by strong noise. Experiments will be
407 presented in section IV to show these advantages.

III. THEORETICAL ANALYSIS

409 In this section, we will further analyze the convergence of
410 the proposed algorithm and its computational complexity.

A. The Convergence

412 We begin with the following Lemmas to verify the conver-
413 gences of the proposed iterative algorithm in Table I:

Lemma 1 [28]: For any two non-zero constants a and b ,
414 we have the following inequality:
415

$$\sqrt{a} - \frac{a}{2\sqrt{b}} \leq \sqrt{b} - \frac{b}{2\sqrt{b}} \quad (14)$$

Lemma 2 [28]: Denoted \mathbf{V} as any nonzero matrix, $\mathbf{V} \in R$,
417 the following inequality holds:
418

$$\sum_i \|\mathbf{v}_t^i\|_2 - \sum_i \frac{\|\mathbf{v}_t^i\|_2^2}{2\|\mathbf{v}_{t-1}^i\|_2} \leq \sum_i \|\mathbf{v}_{t-1}^i\|_2 - \sum_i \frac{\|\mathbf{v}_{t-1}^i\|_2^2}{2\|\mathbf{v}_{t-1}^i\|_2} \quad (15)$$

where \mathbf{v}_t^i , \mathbf{v}_{t-1}^i denote the i -th row of matrix \mathbf{V}_t and \mathbf{V}_{t-1} .
421

With the above Lemma 1 and Lemma 2, we have the
422 following theorem:
423

Theorem 2: Given all the parameters on the objective
424 function except $\mathbf{B}, \mathbf{h}, \mathbf{A}, \mathbf{G}, \mathbf{D}, \tilde{\mathbf{D}}, \hat{\mathbf{D}}$ the iterative approach
425 shown in Table I will monotonically decrease the objective
426 function value of (2) in each iteration and provide a local
427 optimal solution of the objective function.
428

Proof: For simplicity, we denote the objective function
429 of (9) as $F(\mathbf{B}, \mathbf{h}, \mathbf{A}, \mathbf{G}, \mathbf{D}) = F(\mathbf{B}, \mathbf{h}, \mathbf{A}, \mathbf{G}, \mathbf{D}, \tilde{\mathbf{D}}, \hat{\mathbf{D}})$.
430 Suppose for the $(t-1)$ -th iteration, $\mathbf{B}_{t-1}, \mathbf{h}_{t-1}, \mathbf{A}_{t-1}, \mathbf{G}_{t-1},$
431 $\mathbf{D}_{t-1}, \tilde{\mathbf{D}}_{t-1}$ and $\hat{\mathbf{D}}_{t-1}$ were obtained. Then we have the
432 following inequality from (10) and (11):
433

$$\begin{aligned} & F(\mathbf{B}_t, \mathbf{h}_t, \mathbf{A}_{t-1}, \mathbf{G}_{t-1}, \mathbf{D}_{t-1}, \tilde{\mathbf{D}}_{t-1}, \hat{\mathbf{D}}_{t-1}) \\ & \leq F(\mathbf{B}_{t-1}, \mathbf{h}_{t-1}, \mathbf{A}_{t-1}, \mathbf{G}_{t-1}, \mathbf{D}_{t-1}, \tilde{\mathbf{D}}_{t-1}, \hat{\mathbf{D}}_{t-1}) \end{aligned} \quad (16)$$

For \mathbf{A}_t , as its optimal value comes from the SVD decomposition value of $(\mathbf{h}_t \mathbf{1}^T - \mathbf{Y}^T) \hat{\mathbf{D}}_{t-1} X \mathbf{B}_t$, that will further decrease the objective function, we have

$$\begin{aligned} F(\mathbf{B}_t, \mathbf{h}_t, \mathbf{A}_t, \mathbf{G}_{t-1}, \mathbf{D}_{t-1}, \tilde{\mathbf{D}}_{t-1}, \hat{\mathbf{D}}_{t-1}) \\ \leq F(\mathbf{B}_{t-1}, \mathbf{h}_{t-1}, \mathbf{A}_{t-1}, \mathbf{G}_{t-1}, \mathbf{D}_{t-1}, \tilde{\mathbf{D}}_{t-1}, \hat{\mathbf{D}}_{t-1}) \end{aligned} \quad (17)$$

For \mathbf{G}_t , \mathbf{D}_t , since $(\mathbf{G}_{ij})_t = \|\mathbf{x}_i \mathbf{B}_t \mathbf{A}_t^T - \mathbf{x}_j \mathbf{B}_t \mathbf{A}_t^T\|_2$, $(\mathbf{D}_{ii})_t = \sum_i ((W\emptyset \mathbf{G})_{ij})_t$ and \mathbf{B}_t , \mathbf{A}_t were both obtained, then we further have the following inequality:

$$\begin{aligned} F(\mathbf{B}_t, \mathbf{h}_t, \mathbf{A}_t, \mathbf{G}_t, \mathbf{D}_t, \tilde{\mathbf{D}}_{t-1}, \hat{\mathbf{D}}_{t-1}) \\ \leq F(\mathbf{B}_{t-1}, \mathbf{h}_{t-1}, \mathbf{A}_{t-1}, \mathbf{G}_{t-1}, \mathbf{D}_{t-1}, \tilde{\mathbf{D}}_{t-1}, \hat{\mathbf{D}}_{t-1}) \end{aligned} \quad (18)$$

For simplicity, let $\mathbf{Q} = \mathbf{X} \mathbf{B} \mathbf{A}^T + \mathbf{1} \mathbf{h}^T - \mathbf{Y}$ in (9), it goes

$$\begin{aligned} & tr(\mathbf{B}^T \mathbf{X}^T (\mathbf{D} - W\emptyset \mathbf{G}) \mathbf{X} \mathbf{B}) + \beta tr(\mathbf{B}^T \tilde{\mathbf{D}} \mathbf{B}) \\ & + \gamma tr((\mathbf{X} \mathbf{B} \mathbf{A}^T + \mathbf{1} \mathbf{h}^T - \mathbf{Y})^T \tilde{\mathbf{D}} (\mathbf{X} \mathbf{B} \mathbf{A}^T + \mathbf{1} \mathbf{h}^T - \mathbf{Y})) \\ & + \lambda tr(\mathbf{h}^T \mathbf{h}) \\ & = tr(\mathbf{B}^T \mathbf{X}^T (\mathbf{D} - W\emptyset \mathbf{G}) \mathbf{X} \mathbf{B}) + \beta tr(\mathbf{B}^T \tilde{\mathbf{D}} \mathbf{B}) \\ & + \gamma tr(\mathbf{Q}^T \tilde{\mathbf{D}} \mathbf{Q}) + \lambda tr(\mathbf{h}^T \mathbf{h}) \end{aligned} \quad (19)$$

Since we have obtained \mathbf{B}_t , \mathbf{h}_t , \mathbf{A}_t , \mathbf{G}_t , \mathbf{D}_t for (19), then the following inequality holds

$$\begin{aligned} & tr(\mathbf{B}_t^T \mathbf{X}^T (\mathbf{D}_t - W\emptyset \mathbf{G}_t) \mathbf{X} \mathbf{B}_t) + \beta tr(\mathbf{B}_t^T \tilde{\mathbf{D}}_{t-1} \mathbf{B}_t) \\ & + \gamma tr(\mathbf{Q}_t^T \tilde{\mathbf{D}}_{t-1} \mathbf{Q}_t) + \lambda tr(\mathbf{h}_t^T \mathbf{h}_t) \\ & \leq tr(\mathbf{B}_{t-1}^T \mathbf{X}^T (\mathbf{D}_{t-1} - W\emptyset \mathbf{G}_{t-1}) \mathbf{X} \mathbf{B}_{t-1}) \\ & + \beta tr(\mathbf{B}_{t-1}^T \tilde{\mathbf{D}}_{t-1} \mathbf{B}_{t-1}) \\ & + \gamma tr(\mathbf{Q}_{t-1}^T \tilde{\mathbf{D}}_{t-1} \mathbf{Q}_{t-1}) + \lambda tr(\mathbf{h}_{t-1}^T \mathbf{h}_{t-1}) \end{aligned} \quad (20)$$

This is

$$\begin{aligned} & tr(\mathbf{B}_t^T \mathbf{X}^T (\mathbf{D}_t - W\emptyset \mathbf{G}_t) \mathbf{X} \mathbf{B}_t) + \lambda tr(\mathbf{h}_t^T \mathbf{h}_t) \\ & + \beta \sum_i \frac{\|\mathbf{B}_t^i\|_2^2}{2\|\mathbf{B}_{t-1}^i\|_2} + \gamma \sum_i \frac{\|\mathbf{Q}_t^i\|_2^2}{2\|\mathbf{Q}_{t-1}^i\|_2} \\ & \leq tr(\mathbf{B}_{t-1}^T \mathbf{X}^T (\mathbf{D}_{t-1} - W\emptyset \mathbf{G}_{t-1}) \mathbf{X} \mathbf{B}_{t-1}) + \lambda tr(\mathbf{h}_{t-1}^T \mathbf{h}_{t-1}) \\ & + \beta \sum_i \frac{\|\mathbf{B}_{t-1}^i\|_2^2}{2\|\mathbf{B}_{t-1}^i\|_2} + \gamma \sum_i \frac{\|\mathbf{Q}_{t-1}^i\|_2^2}{2\|\mathbf{Q}_{t-1}^i\|_2} \end{aligned} \quad (21)$$

Then, we have

$$\begin{aligned} & tr(\mathbf{B}_t^T \mathbf{X}^T (\mathbf{D}_t - W\emptyset \mathbf{G}_t) \mathbf{X} \mathbf{B}_t) + \lambda tr(\mathbf{h}_t^T \mathbf{h}_t) \\ & + \beta \sum_i \|\mathbf{B}_t^i\|_2 - \beta \left(\sum_i \|\mathbf{B}_t^i\|_2 - \sum_i \frac{\|\mathbf{B}_t^i\|_2^2}{2\|\mathbf{B}_{t-1}^i\|_2} \right) \\ & + \gamma \sum_i \|\mathbf{Q}_t^i\|_2 - \gamma \left(\sum_i \|\mathbf{Q}_t^i\|_2 - \sum_i \frac{\|\mathbf{Q}_t^i\|_2^2}{2\|\mathbf{Q}_{t-1}^i\|_2} \right) \\ & \leq tr(\mathbf{B}_{t-1}^T \mathbf{X}^T (\mathbf{D}_{t-1} - W\emptyset \mathbf{G}_{t-1}) \mathbf{X} \mathbf{B}_{t-1}) + \lambda tr(\mathbf{h}_{t-1}^T \mathbf{h}_{t-1}) \\ & + \beta \sum_i \|\mathbf{B}_{t-1}^i\|_2 - \beta \left(\sum_i \|\mathbf{B}_{t-1}^i\|_2 - \sum_i \frac{\|\mathbf{B}_{t-1}^i\|_2^2}{2\|\mathbf{B}_{t-1}^i\|_2} \right) \\ & + \gamma \sum_i \|\mathbf{Q}_{t-1}^i\|_2 - \gamma \left(\sum_i \|\mathbf{Q}_{t-1}^i\|_2 - \sum_i \frac{\|\mathbf{Q}_{t-1}^i\|_2^2}{2\|\mathbf{Q}_{t-1}^i\|_2} \right) \end{aligned} \quad (22)$$

From Lemma 2, we further have

$$\begin{aligned} & tr(\mathbf{B}_t^T \mathbf{X}^T (\mathbf{D}_t - W\emptyset \mathbf{G}_t) \mathbf{X} \mathbf{B}_t) + \lambda tr(\mathbf{h}_t^T \mathbf{h}_t) \\ & + \beta \sum_i \|\mathbf{B}_t^i\|_2 + \gamma \sum_i \|\mathbf{Q}_t^i\|_2 \\ & \leq tr(\mathbf{B}_{t-1}^T \mathbf{X}^T (\mathbf{D}_{t-1} - W\emptyset \mathbf{G}_{t-1}) \mathbf{X} \mathbf{B}_{t-1}) + \lambda tr(\mathbf{h}_{t-1}^T \mathbf{h}_{t-1}) \\ & + \beta \sum_i \|\mathbf{B}_{t-1}^i\|_2 + \gamma \sum_i \|\mathbf{Q}_{t-1}^i\|_2 \end{aligned} \quad (23)$$

With the definition of $L_{2,1}$ -norm, we finally arrive at

$$\begin{aligned} F(\mathbf{B}_t, \mathbf{h}_t, \mathbf{A}_t, \mathbf{G}_t, \mathbf{D}_t, \tilde{\mathbf{D}}_t, \hat{\mathbf{D}}_t) \\ \leq F(\mathbf{B}_{t-1}, \mathbf{h}_{t-1}, \mathbf{A}_{t-1}, \mathbf{G}_{t-1}, \mathbf{D}_{t-1}, \tilde{\mathbf{D}}_{t-1}, \hat{\mathbf{D}}_{t-1}) \end{aligned} \quad (24)$$

It is easy to draw the conclusion from (24) that according to the updating rule in Table I, the proposed objective function monotonically decreases and the corresponding iterative algorithm will finally converge to the local optimal solution. \square

B. Computational Complexity Analysis

For simplicity, we suppose that the dimension of the training data X is d . The proposed algorithm aims to obtain the local optimal projection matrix \mathbf{B} for further feature selection or classification. During the computing procedure, the algorithm needs to compute seven variables (i.e. \mathbf{B} , \mathbf{h} , \mathbf{A} , \mathbf{G} , \mathbf{D} , $\tilde{\mathbf{D}}$, $\hat{\mathbf{D}}$). Computing \mathbf{h} in (10) is up to $O(4d^2)$ and \mathbf{B} in (11) is up to $O(d^3)$. The matrix \mathbf{A} is obtained from the SVD of $(\mathbf{h} \mathbf{1}^T - \mathbf{Y}^T) \hat{\mathbf{D}} X \mathbf{B}$, then its computational complexity is also $O(d^3)$ at most. Computing the variable \mathbf{G} costs $O(nK(k^2 + kd))$, where K denotes the number of neighbors. The computational complexity of the variable \mathbf{D} , $\hat{\mathbf{D}}$ is the same, i.e. $O(Kd)$ while computing $\tilde{\mathbf{D}}$ needs $O(ndK + nK^2c + nc)$. To sum up, the total computational complexity of the proposed algorithm is $O(Td^3)$ by ignoring some constants since these constant are small compared with the dimension of the training data, where T is the iteration steps.

IV. EXPERIMENTS

In this section, a set of experiments are presented to evaluate the performance of the proposed Generalized Robust Regression for jointly sparse subspace learning (GRR) for face recognition. For comparison, several different methods were also tested on the six databases. The methods include the dimensionality reduction methods, i.e. Sparse Principal Component Analysis (SPCA) [5], Locality Preserving Projections (LPP) [13], the traditional Ridge regression (RR) [4], the L_1 -norm based dimensionality reduction methods, i.e. Principal Component Analysis based on L_1 -norm maximization (PCA-L1) [23], Linear Discriminant Analysis based on L_1 -norm maximization (LDA-L1) [6] and Outlier-resisting graph embedding (LPP-L1) [7], the $L_{2,1}$ -norm regularization method (i.e. $L_{2,1}$ -norm regularized discriminative feature selection for unsupervised learning (UDFS) [31]), the nonlinear kernel-based method (KPCA) and the classical sparse learning method (i.e. robust face recognition via sparse representation (SRC-L1S) [25]).

The Yale face database was used to evaluate the performance of GRR while there are variations in facial expression



Fig. 1. Sample images of one person on FERET face database.

and lighting conditions. The AR database was used to explore the robustness of GRR in frontal views of faces with variational facial expressions, illumination and occlusions. The FERET and ORL databases were used to explore robustness of GRR with the variations of face images in facial expression and pose. Besides, the Char74K_15 database was used to test the performance of GRR in the English character images with excessive occlusion, low resolution or noise. The LFW database was used to evaluate the effectiveness of the proposed GRR and other competing methods based on deep learning situation.

In addition, the AR database was also used to test the robustness of the proposed GRR against the compared methods in three challenging situations (when training images are with random block corruption, disguise and illumination variation).

539 A. Experiments on FERET Face Database

540 The FERET face database is a result of the FERET program,
 541 which was sponsored by the USD department of Defense
 542 through the DARPA Program [37]. It has become a standard
 543 database for testing and evaluating state-of-the-art face recog-
 544 nition algorithms. The proposed method was tested on a subset
 545 of the FERET database. This subset included 1,400 images
 546 of 200 individuals (each individual has seven images) and
 547 involved variations in facial expression, illumination, and pose.
 548 In the experiment, the facial portion of each original image
 549 was automatically cropped based on the location of the eyes,
 550 and the cropped images were resized to 40×40 pixels. The
 551 sample images of one person are shown in Fig.1.

552 *1) Experimental Setting:* The performance of the feature
 553 selection methods (i.e. the proposed GRR and SPCA, LPP,
 554 RR, PCA-L1, LDA-L1, LPP-L1, UDFS, SRC-L1LS, KPCA)
 555 are measured by the recognition rate with selected features on
 556 the testing data set. PCA was first used as the pre-processing to
 557 reduce the dimensionality of the original data. The experiments
 558 were independently performed 10 times on the six databases.
 559 The nearest neighbor classifier is used for classification. The
 560 average recognition rates and the corresponding dimensions
 561 and the standard deviations of each method were listed on the
 562 tables. Besides, the comparison results were also shown in the
 563 figures when several images of each individual were randomly
 564 selected for training while the remaining images were used for
 565 testing. The real dimension used in the experiment is the same
 566 with the number marked on the horizontal axis on the six data
 567 sets and all the images are cropped and aligned automatically.
 568 Usually, the initial value of the proposed algorithm has little
 569 effect on its performance, therefore the variables in GRR are
 570 randomly initialized in our experiments.

571 *2) Exploration of the Performance of the Parameters:*
 572 Since the value of the three parameters β , γ and λ affect
 573 the performance of the proposed GRR in some degree, thus
 574 we need to explore the optimal parameter values of GRR.
 575 For the other methods, since UDFS was introduced with the

parameter lying on the area of $[10^{-3}, 10^3]$, then we used this area for UDFS to perform feature selection and presented the corresponding experimental results. The parameters of other methods were selected according to the related introduction in the original paper.

In FERET database, we first analyze the optimal values of β , γ , λ and then use the values to obtain the best performance for GRR. Table II shows the best average recognition rates, the corresponding dimensions and the standard deviations of different methods with different dimensions form 5 to 150 based on 10 times experiments. Fig. 2 (a) shows the recognition rates when the two parameters β and γ change from 10^{-9} to 10^9 . Fig. 2 (b) presents the performance of the parameters λ varies in the area of $[10^{-9}, 10^9]$ on all databases. It is obvious that, the value of λ does not affect the performance when it lies in the area of $[10^{-9}, 10^0]$. For simplicity, we choose $\lambda = [10^{-9}, 10^0]$ in all experiments. Fig. 4 (a) shows the average recognition rates versus various dimensions of different methods when 6 images were selected for the training while the remaining images were used for testing.

From the result showed in Fig. 2 (a), we obtain the optimal area for parameters β and γ are $[10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}]$ and $[10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}]$ respectively. Namely, GRR obtains the best performance when the two parameters lie on these areas. Otherwise, other parameter values would cause the large decline of the face recognition rate of GRR. Table II shows that SRC-L1LS and GRR perform better than other methods. Fig. 4 (a) indicates when 6 images of each person are used for training, GRR gives full play to its advantages and outperforms other methods with about 15.75% to 41.05% of the recognition rate. Moreover, GRR keeps going up smoothly and quickly achieves the best performance for feature selection.

609 B. Experiments on ORL Face Database

610 The ORL dataset consists of 10 face images from 40 subjects
 611 for a total of 400 images, with some variation in pose,
 612 facial expression and details. The resolution of the images is
 613 112×92 , with 256 gray-levels. Before the experiments, we
 614 scale the pixel gray level to the range $[0, 1]$. Fig. 3 depicts
 615 some sample images of a typical subset in the ORL dataset.

616 In this experiment, l ($l = 3, 4, 5$) images of each individual
 617 were randomly selected for training, and the rest of the
 618 images in the data set were used for testing. The optimal
 619 areas of parameter β and γ were set in $[10^6, 10^7, 10^8]$
 620 and $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$ respectively. Fig. 4 (b) shows
 621 the average testing recognition rates. Table III listed the
 622 performance of different methods. It is obvious that GRR
 623 outperforms other methods again.

624 C. Experiments on Yale Face Database

625 The Yale face database [43] was constructed at the Yale
 626 Center for Computational Vision and Control. It contains
 627 165 grayscale images of 15 individuals. The images demon-
 628 strate variations in lighting condition (left-light, center-light,
 629 right-light), facial expression (normal, happy, sad, sleepy,
 630 surprised, and wink), and with/without glasses. Fig. 5 shows
 631 the sample images from this database.

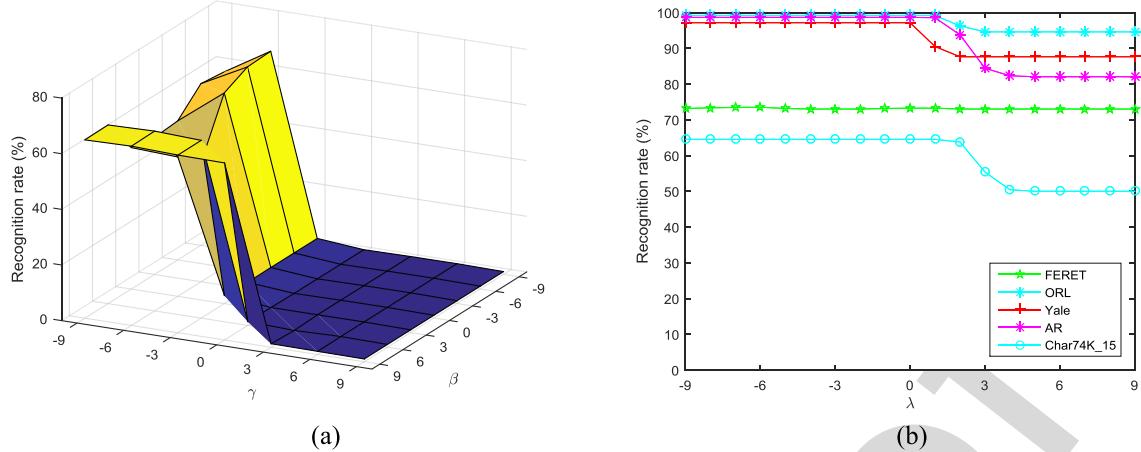


Fig. 2. (a) The recognition rate versus the parameters β and γ on the FERET face database. (b) The recognition rate versus the parameters γ on the FERET, ORL, Yale, AR, Char74K_15 database, respectively.

TABLE II
THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF DIFFERENT METHODS ON FERET FACE DATABASE

Training samples	SPCA	LPP	RR	PCA-L1	LDA-L1	LPP-L1	UDFS	KPCA	SRC-L1LS	GRR
4	58.47±7.76 150	52.47±7.59 150	63.17±9.50 150	58.48±7.94 150	31.12±6.99 30	36.85±6.12 150	58.62±8.08 140	59.28±9.49 145	73.18±4.31 150	70.45±11.33 135
	61.58±9.64 140	55.50±7.24 150	69.47±6.76 150	61.52±9.62 150	33.17±10.08 30	39.58±9.81 150	61.73±9.96 145	65.63±6.95 145	74.95±5.48 130	74.20±15.33 100
6	68.85±9.20 150	63.40±5.36 150	70.90±7.81 150	68.95±9.28 130	45.50±9.50 50	49.15±8.00 150	69.95±10.31 110	68.80±7.41 145	74.45±11.90 125	90.20±6.11 140

TABLE III
THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF DIFFERENT METHODS ON THE ORL FACE DATABASE

Training samples	SPCA	LPP	RR	PCA-L1	LDA-L1	LPP-L1	UDFS	KPCA	SRC-L1LS	GRR
3	89.46±1.76 95	76.82±1.72 60	87.32±1.88 40	89.46±1.62 95	71.46±3.59 15	83.50±3.10 100	90.14±1.88 55	85.64±2.91 95	90.89±1.79 100	99.64±0.00 25
	92.79±1.73 100	85.79±2.30 90	90.87±1.75 40	92.83±1.88 100	82.00±1.74 25	88.00±2.49 90	92.92±1.64 60	90.58±2.16 95	93.42±1.79 75	99.58±0.00 35
5	94.80±1.46 55	90.00±1.59 65	92.50±1.37 40	94.70±1.41 85	87.30±2.74 50	90.85±1.67 100	95.00±1.56 60	92.55±2.54 95	95.10±1.73 60	100.00±0.00 10



Fig. 3. Sample images of one person on ORL face database.

In this experiment, l ($l = 4, 5, 6$) images of each individual were randomly selected for training, and the rest of the images in the data set were used for testing. The optimal areas of parameter β and γ were $[10^7, 10^8, 10^9]$ and $[10^{-7}, 10^{-6}]$, respectively. The performances of different methods are shown in Table IV. Fig. 6 (a) shows the average testing recognition rates when 6 images of each people were used for training. It clearly indicates that GRR obtains outstanding results.

D. Experiments on AR Face Database

The AR face database [38] contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of 120 individuals (65 men and 55 women) were selected and divided into two sessions (separated by two weeks) and each session contains 13 color images. 20 images of these 120 individuals were selected and used in our experiments. The face portion of each image was manually cropped and then normalized to 50×40 pixels. The sample images of one person are shown in Fig. 7 (a). These images vary as follows: neutral expression, smiling, angry, screaming, left light on, right light on, all sides

640
641
642
643
644
645
646
647
648
649
650
651
652

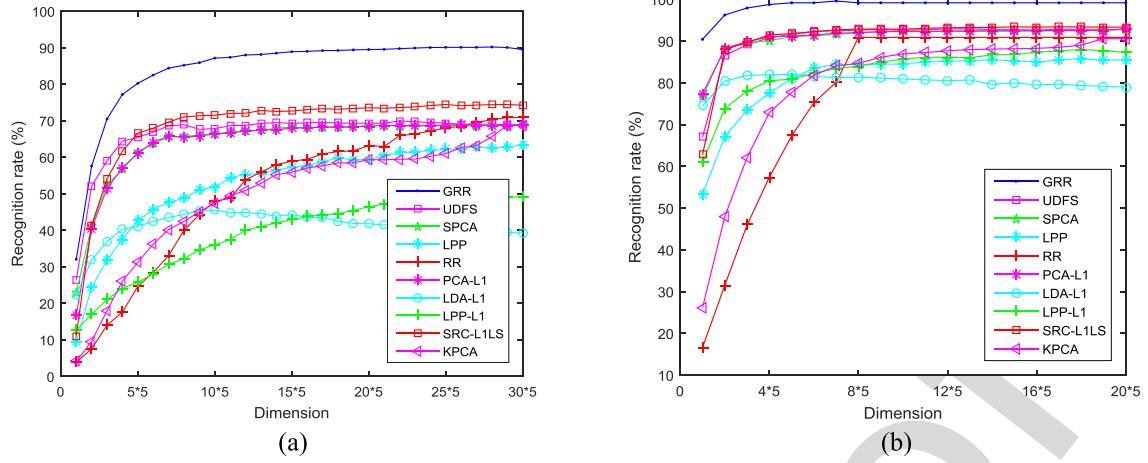


Fig. 4. (a) The recognition rates (%) versus the dimensions of different methods on the FERET face database. (b) The recognition rates (%) versus the dimensions of different methods on the ORL face database.



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

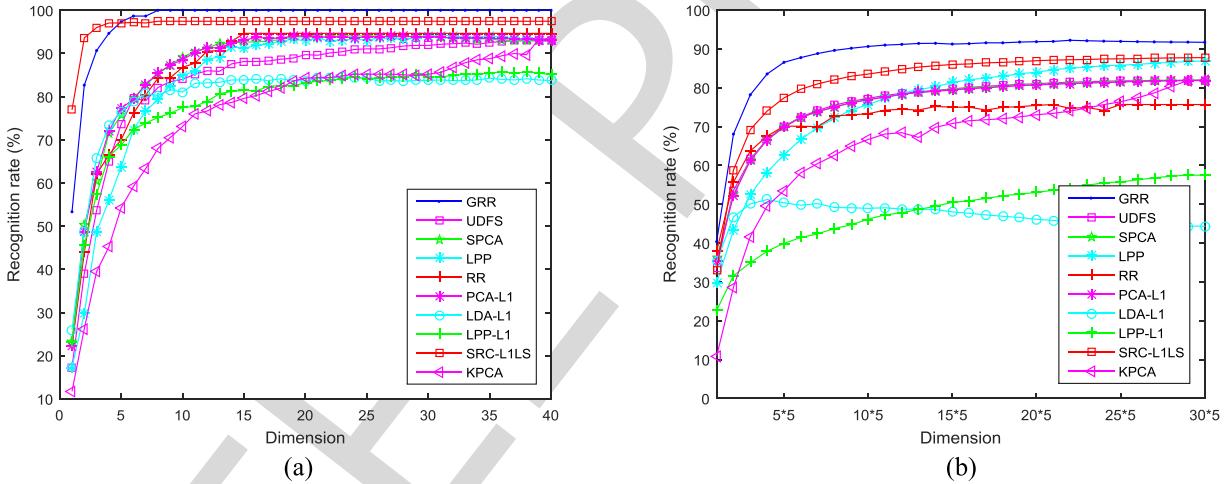


Fig. 6. (a) The recognition rates (%) versus the dimensions of different methods on the Yale face database. (b) The recognition rates (%) versus the dimensions of different methods on the AR face database.

653 light on, wearing sun glasses, wearing sun glasses and left light
654 on, wearing sun glasses and right light on.

655 In this experiment, we randomly selected l ($l = 4, 5, 6$)
656 images of each individual for training, and the rest of the
657 images in the data set were used for testing. The optimal
658 areas of parameter β and γ were in $[10^7, 10^8, 10^9, 10^{10}]$ and
659 $[10^{-2}, 10^{-1}]$ respectively. Fig. 6 (b) shows the average testing
660 recognition rates with 4 images of each object used as training
661 set. Table V lists the performance of different methods. We can
662 know from both Fig. 6 (b) and Table V that GRR outperforms
663 the other methods.

664 E. Experiments on Char74K_15 Database

665 The character images of the Char74K dataset [39] are
666 mostly photographed from sign boards, hoardings and

667 advertisements and a few images of products are in super-
668 markets and shops. Fig. 8. (a) depicts sample images of
669 English scene characters from Char74k. As for the Char74K
670 dataset, we only use the English character images cropped
671 from the natural scene images. The English dataset has
672 12503 characters, of which 4798 were labeled as bad images
673 due to excessive occlusion, low resolution or noise. It contains
674 62 character classes. A small subset with a standard partition
675 is used in our experiments, i.e. Char74K-15, which contains
676 15 training samples per class and 15 test samples per class.

677 The optimal areas of β and γ are $[10^1, 10^2, 10^3, 10^4, 10^5]$,
678 $[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$ respectively. Fig. 8 (b) demon-
679 strates the average testing recognition rates versus the dimen-
680 sions of different methods while Table VI shows the best
681 average recognition rates and the corresponding dimensions

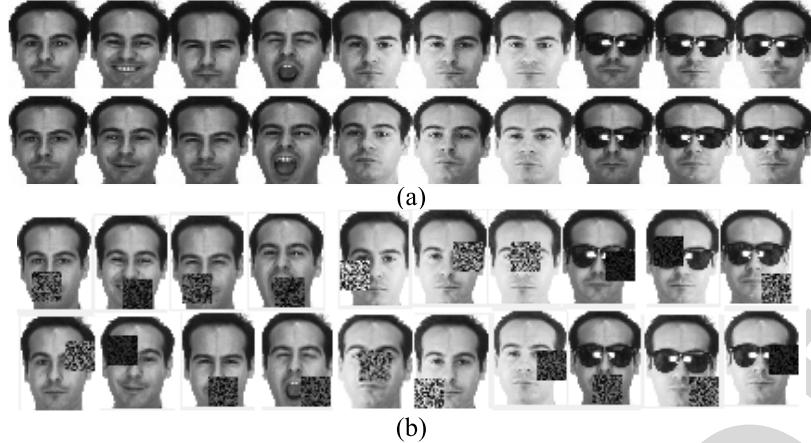


Fig. 7. (a) The sample images of one person form the AR face database. (b) The sample images with block noise of one person in our experiment on the AR face database.

TABLE IV

THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF DIFFERENT METHODS ON THE YALE FACE DATABASE

Training samples	SPCA	LPP	RR	PCA-L1	LDA-L1	LPP-L1	UDFS	KPCA	SRC-L1LS	GRR
4	92.95±4.00 25	86.86±4.38 38	91.43±4.60 15	92.67±3.79 36	82.67±7.25 38	60.67±6.91 6	92.10±4.29 35	91.90±3.98 39	97.24±1.57 8	97.14±0.00 13
5	93.00±4.29 30	89.89±3.56 27	93.00±4.91 15	93.22±4.66 26	83.44±7.98 28	64.00±8.20 7	92.67±4.87 36	92.56±5.08 39	97.67±1.33 7	98.89±0.00 12
6	94.13±4.83 23	94.13±4.09 31	94.67±5.01 15	94.13±5.16 20	84.53±7.03 23	85.73±9.80 36	92.93±5.79 37	93.33±5.50 39	97.47±2.33 8	100.00±0.00 8

TABLE V

THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF DIFFERENT METHODS ON THE AR FACE DATABASE

Training samples	SPCA	LPP	RR	PCA-L1	LDA-L1	LPP-L1	UDFS	KPCA	SRC-L1LS	GRR
4	81.92±4.73 150	86.90±4.97 150	75.67±6.61 105	81.93±4.74 150	51.20±11.94 20	57.48±11.40 145	81.93±4.79 150	81.85±6.12 145	87.75±2.37 150	92.21±4.03 110
5	84.62±4.07 150	89.83±3.84 150	81.49±4.62 105	84.62±4.02 150	54.23±9.22 55	59.72±8.32 150	84.62±4.04 150	86.07±6.85 145	89.08±2.87 150	93.47±3.32 75
6	85.08±4.89 150	91.60±3.65 150	83.95±4.44 105	85.08±4.93 150	58.56±7.47 45	58.83±8.33 150	85.08±4.89 150	86.72±4.20 140	89.62±3.13 150	93.87±3.00 80

and the standard deviations of each method. From the result, we can find that in this dataset, GRR outperforms other methods again.

F. Robustness Evaluation on AR Face Database

To evaluate the performance of the proposed GRR and other compared methods in the case when there are various noises corrupted in the images, we conducted series of experiments including random block corruption and disguise as well as illumination variation.

1) *Images With Random Block Corruption:* To evaluate the robustness of the proposed GRR, the noise was added in the face images in our experiments and the samples image with

block noise of one person are shown in Fig.7 (b). Table VII lists the best average recognition rates, the corresponding dimensions and the standard deviations of different methods based on 10 times experiments on AR database with block size $5*5$, $10*10$, $15*15$, respectively. Fig.9 (a) shows the average recognition rates versus the dimensions of different methods in the case when 4 images of each individual were randomly selected for training from the images that are corrupted by a block size $10*10$. All of the results prove the robustness and effectiveness of the proposed GRR.

2) *Images With Disguise:* To investigate the proposed GRR and other compared methods in the case when training set is corrupted by varying percentage of occlusion in face images,

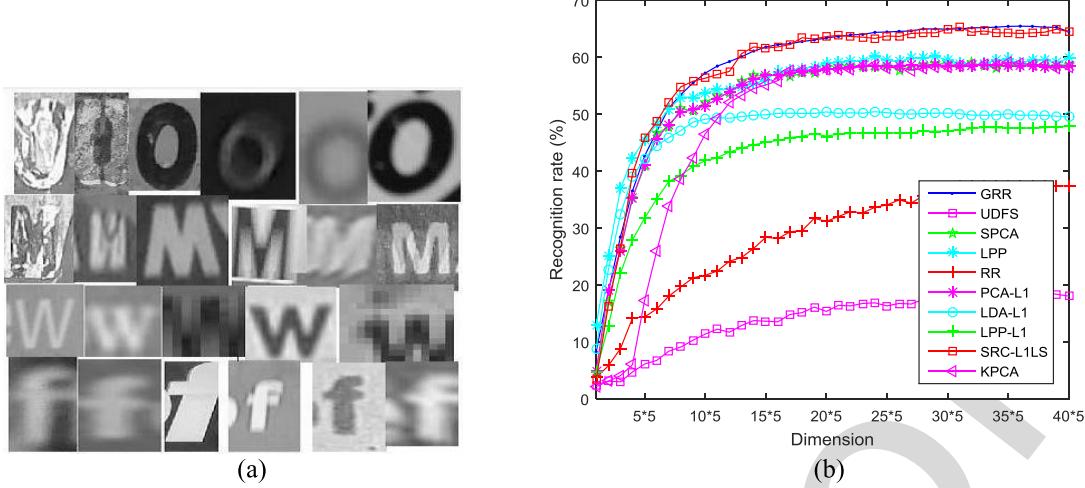


Fig. 8. (a) Sample images of English scene characters from Char74k. (b) The recognition rates (%) versus the dimensions of different methods on the Char74K_15 face databases.

TABLE VI
THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF DIFFERENT METHODS ON THE CHAR74K_15 FACE DATABASE

Method	SPCA	LPP	RR	PCA-L1	LDA-L1	LPP-L1	UDFS	KPCA	SRC-L1LS	GRR
Recognition rate	59.03 ± 0.00	60.22 ± 0.00	37.42 ± 0.00	58.68 ± 0.14	50.45 ± 0.61	47.91 ± 1.01	18.71 ± 0.00	59.14 ± 0.00	65.27 ± 0.00	65.41 ± 0.36
	64	48	62	66	48	80	76	70	62	72

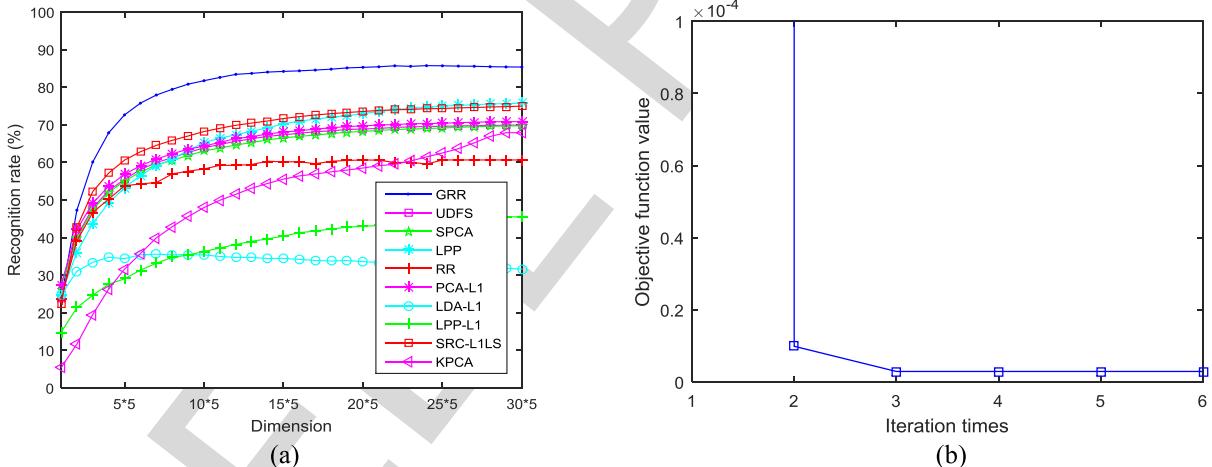


Fig. 9. (a) The recognition rates (%) versus the dimensions of different methods on the AR face databases with noise. (b) An example of the convergence curve of GRR.



Fig. 10. Example images from Session 1 of the AR database.

two protocols are conducted as similar in [40] and [41]. That is, occlusion in training set due to (1) sunglasses, (2) scarf, respectively. Note that scarf accounts for occlusion of about 40% of each face image while the occlusion of sunglasses

amounts to about 20%. Fig.10. shows the sample image of session 1 on AR database in our experiment.

a) *Sunglasses*: For each individual, n_c neutral images and $n_o \in \{0, 1, 2, 3\}$ occluded image(s) from Session 1 are used

TABLE VII
THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF
DIFFERENT METHODS ON THE AR FACE DATABASE WITH NOISE

Block size	Training samples	SPCA	LPP	RR	PCA-L1	LDA-L1	LPP-L1	UDFS	KPCA	SRC-L1LS	GRR
5*5	4	80.47±4.73	83.46±5.46	73.70±6.16	80.22±4.97	47.41±10.88	53.81±10.58	80.64±4.86	83.50±5.45	86.12±2.40	90.83±4.37
		150	150	105	150	40	150	150	145	150	120
	5	83.39±3.91	87.49±4.01	79.19±4.20	83.28±4.10	51.38±8.04	58.78±7.80	83.38±3.95	83.08±5.19	87.65±2.92	92.42±4.00
		150	150	105	150	50	150	145	145	150	110
	6	83.76±4.69	89.47±4.42	81.82±4.21	83.94±4.78	55.32±8.34	58.21±6.31	84.10±4.65	86.55±4.61	88.26±3.22	93.01±3.52
		150	150	105	150	50	150	135	145	150	120
10*10	4	69.74±5.45	75.78±6.52	60.64±4.93	70.85±5.15	35.56±8.43	45.52±10.32	69.92±5.44	67.89±7.70	74.96±2.65	85.73±4.82
		150	150	100	145	35	150	150	145	150	120
	5	73.11±3.55	80.45±5.78	66.28±3.24	74.09±3.90	41.16±4.72	48.97±6.27	73.20±3.83	69.63±5.80	77.32±2.79	88.55±3.65
		150	150	100	150	50	145	150	145	150	130
	6	73.99±4.21	82.91±6.11	69.21±3.77	75.11±4.43	44.76±5.33	49.81±5.83	74.14±4.59	72.12±6.31	78.36±2.83	89.36±3.36
		145	150	105	150	75	150	145	145	150	115
15*15	4	50.82±4.17	56.36±5.86	45.95±3.59	50.66±4.80	25.92±5.18	32.34±6.46	51.06±4.24	50.63±3.61	55.79±2.45	73.34±5.29
		150	150	105	150	35	150	135	145	150	120
	5	54.26±2.83	61.83±6.55	51.97±2.22	54.26±3.46	30.11±3.99	34.39±4.06	54.31±2.86	56.83±2.90	58.76±2.36	78.22±2.66
		150	150	105	150	40	150	140	140	150	120
	6	56.24±2.82	65.81±6.08	55.96±2.96	56.08±3.50	32.36±5.43	36.05±4.92	56.24±2.85	58.45±3.48	60.48±2.82	80.13±2.99
		145	150	105	150	55	150	150	140	150	120

for training, where $n_c + n_o = 7$. Session 2 including 7 neutral images plus 3 occluded image (sunglasses) are used for testing.

b) *Scarf*: The setup of training set and testing set are the same with sunglasses case except using scarf instead of sunglasses as occlusion of the images.

The results of the above two occlusion cases are presented in table VIII. It is obvious that the proposed GRR and SRC always perform better than other methods in these situations. For the sunglasses case, SRC performs a little better than GRR as a whole, yet GRR definitely outperforms SRC under the scarf case. The reason is as follow. SRC assumes the testing image is approximately reconstructed by the training image. However, if the training set include few or even no occluded images, the reconstructed error as classification criteria is not so effective since it varies in large range. In these two occluded cases, especially the scarf case (since the occlusion percentage is much higher), SRC generates large reconstruction errors which bring negative impact on face recognition task. On the contrary, the proposed GRR can guarantee the robustness by designing a more compact and effective model. On one hand, it utilizes $L_{2,1}$ -norm minimization instead of L_2 -norm to alleviate the sensitiveness to outliers. On the other hand, it uses the elastic factor \mathbf{h} to avoid the overfitting problem in this case.

3) *Images With Illumination Variation*: In this part, the first 7 images of each person on the AR database are used to form the training set and testing set. For each individual, the first 4 images varying on facial expression plus $n_{il} \in \{0, 1, 2, 3\}$ image(s) varying on illumination in Session 1 are used to form the training set and the images in Session 2 are used as the testing set. That is, $n_{ni} \in \{4, 5, 6, 7\}$ images of each person are used for training.

Table IX shows the corresponding experimental results. It is obvious that SRC and GRR are comparative and they always outperform other compared methods. Although GRR seems to be a little inferior to SRC at first, it finally catches up and even outperforms SRC.

G. Reconstruction and Subspace Learning

This subsection further performs a set of experiments to explore the learned subspace properties of the proposed GRR and some classical methods, i.e. PCA, RPCA, RR and LPP. In this part, we consider two versions of experiments. For the first version, 7 neutral images of each individual from AR database are trained by LPP, RR, PCA, RPCA and GRR, respectively and the corresponding results are plotted as images. Fig.11 (a) shows the original image of one person in our experiment. Fig.11 (b)-(d) show reconstruction images obtained by RPCA, PCA, GRR, respectively. Note that for RPCA, the learned low-rank approximation is presented in this experiment. For PCA, 50 principle components are used for reconstruction. To explore the properties of the subspaces learned by GRR and other subspace learning methods (i.e. LPP, RR, PCA), we also present the images of the first 2 projections learned by these methods. Fig.11 (e)-(h) show the first 2 projections of the subspace obtained by LPP, RR, PCA, GRR, respectively. In the second version, all the training images are corrupted by random block noise with $50 * 50$ pixels and the results are shown in Fig. 12. According to these experimental results, we have the following interesting conclusion:

RPCA is good at recovering low-rank components of the original data. It can get rid of the corrupted noise in some degree, which can be seen from Fig.12 (b). PCA is an effective

TABLE VIII

COMPARISONS (RECOGNITION RATE STANDARD DEVIATION AND DIMNSION) WTH DIFFEENT PERCENTAGES OF OCCLUDED IMAGES ($n_o/7$) PRESENTED IN THE TRAINING SET. THE FEATURE DIMENSION IS SET AS 150 FOR ALL METHODS

Method	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf
	0% = 0/7		14.29% = 1/7		28.57% = 2/7		42.86% = 3/7	
SPCA	54.70±0.00 145	50.90±0.00 145	55.59±2.76 150	51.18±2.20 150	57.03±3.27 150	49.69±2.44 150	56.63±2.51 150	48.50±2.97 150
LPP	55.00±0.00 140	51.60±0.00 140	58.35±1.86 150	50.66±1.99 150	59.19±2.02 150	51.32±1.74 150	58.71±2.55 150	47.36±2.30 150
RR	51.40±0.00 75	44.50±0.00 100	56.91±1.01 70	49.99±1.64 100	57.52±1.17 70	51.32±2.47 100	57.56±0.41 70	52.85±0.81 100
PCA-L1	54.83±0.11 145	50.94±0.13 145	55.59±2.75 150	51.24±2.22 150	57.10±3.30 150	49.71±2.45 150	56.64±2.58 150	48.52±2.94 150
LDA-L1	34.87±0.79 110	33.49±0.70 75	34.85±3.22 80	33.53±2.03 90	38.15±2.36 90	31.34±3.05 110	38.95±2.57 75	29.64±2.90 95
LPP-L1	7.61±0.85 5	7.39±1.19 5	8.74±2.55 5	9.28±1.44 10	9.20±3.37 5	8.84±0.48 10	10.02±1.91 5	7.71±0.90 5
UDFS	55.00±0.00 125	51.10±0.00 140	55.69±2.77 135	51.38±2.20 125	57.13±3.34 130	49.87±2.51 105	56.74±2.52 135	48.99±3.10 110
KPCA	56.54±1.01 145	51.14±1.07 145	57.56±3.28 145	52.93±3.35 145	58.42±4.12 145	53.42±2.77 145	59.10±2.83 145	51.74±4.39 145
SRC_L1LS	59.40±0.00 150	51.90±0.00 150	60.00±1.73 150	54.67±1.50 150	60.85±1.87 150	55.28±1.83 150	60.01±1.71 150	53.67±1.24 150
GRR	57.24±0.58 150	52.65±0.53 145	58.24±2.51 145	56.30±1.59 145	60.33±2.99 150	57.26±2.49 140	61.48±2.96 150	55.14±3.42 145

TABLE IX

THE PERFORMANCE (RECOGNITION RATE, STANDARD DEVIATION AND DIMENSION) OF DIFFERENT METHODS ON AR DATABASE WITH VARIOUS ILLUMINATION. THE FEATURE DIMENSION IS SET AS 150 FOR ALL METHODS

Method	Illumination	Illumination	Illumination	Illumination
	0% = 0/4	20% = 1/5	33.33% = 2/6	42.86% = 3/7
SPCA	40.86±0.00 145	52.19±0.48 150	61.89±0.87 145	70.57±0.00 145
LPP	44.43±0.00 130	54.51±3.14 145	65.14±3.62 145	68.71±0.00 140
RR	44.29±0.00 65	52.77±1.55 75	61.27±0.92 70	64.14±0.00 70
PCA-L1	40.86±0.10 150	52.27±0.46 145	61.89±0.77 145	70.59±0.12 145
LDA-L1	19.61±0.88 85	29.81±0.98 45	35.09±2.04 70	42.89±1.71 85
LPP-L1	13.26±0.78 5	10.86±1.24 5	37.53±3.02 150	43.77±2.46 150
UDFS	41.14±0.00 140	52.51±0.52 140	62.29±0.60 130	70.71±0.00 125
KPCA	45.14±3.76 145	53.41±2.64 145	64.47±4.07 145	70.54±1.42 145
SRC_L1LS	53.71±0.00 150	60.71±2.39 150	65.34±0.44 150	71.14±0.00 150
GRR	44.83±0.71 85	57.86±0.85 145	65.93±0.72 135	73.29±0.30 125

method for reconstructing the original images. However, GRR does not tend to reconstruct the original images since it focus on discriminant information instead of reconstruction. That is, GRR finds the most discriminative projections to obtain an

optimal subspace for efficient feature selection or extraction. As it can be seen from Fig.12 (d) and Fig.12 (h), the block noise on the face image is not evident. On the contrary, Fig.12 (g) shows that the block noise on PCA-based images



Fig. 11. (a) is original sample images, (b), (c), (d) is reconstruction images obtained by RPCA, PCA, GRR, respectively; (e), (f), (g), (h) is 2 images of the first 2 projections from subspace obtained by LPP, RR, PCA, GRR, respectively.

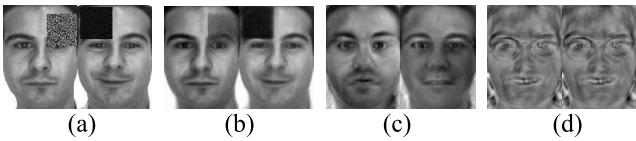


Fig. 12. (a) is corrupted images by block noise with size 50*50, (b), (c), (d) is reconstruction images obtained by RPCA, PCA, GRR, respectively; (e), (f), (g), (h) is 2 images of the first 2 projections from subspace obtained by LPP, RR, PCA, GRR, respectively.

786 is quite obvious. This indicates that GRR can avoid the noise
787 impact more effectively than PCA. Both LPP and RR can
788 obtain the face-like images, but LPP's projections are affected
789 seriously by the block noise.

790 H. Experiments on LFW Database Based on Deep Learning

791 The LFW database [42] is a very challenging dataset which
792 contains images of 5,749 subjects in the uncontrolled environment.
793 The LFW-a dataset is the aligned version of LFW after
794 performing the face alignment. In our experiment, 4324 images
795 of 158 subjects (each subject has more than 10 images) are
796 selected from LFW-a dataset. Note that all images are aligned
797 as well as cropped and resized to 112*96 pixels. Fig.13 shows
798 the sample images in our experiment.

799 The LFW database were used to evaluate the performance
800 of the proposed GRR and other competing methods based
801 on the deep learning techniques. Similar to [43], the deep
802 convolutional neural network (CNN) was used as the feature
803 extractor to obtain the deep features of all samples (the
804 number of the features is 1024). After the deep features
805 were obtained, we further used the subspace learning methods
806 (i.e. SPCA, LPP, RR, PCA-L1, LDA-L1, LPP-L1, UDFS,
807 KPCA, SRC-L1LS and the proposed GRR) to perform feature
808 extraction, and then the nearest neighbor classifier was used
809 for classification.

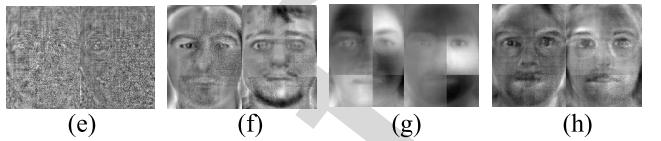
810 In this experiment, we randomly selected l ($l = 4, 5, 6, 7$)
811 images of each subject to form the gallery set and the rest are
812 used as probe set. The optimal areas of parameter β and γ
813 are $[10^7, 10^8, 10^9, 10^{10}]$ and $[10^{-2}, 10^{-1}]$, respectively.

814 The best recognitions rates of deep feature plus the subspace
815 learning methods are shown in Table X. The results indicate
816 the performance of the proposed GRR is still better than other
817 methods. Fig.9 (b) shows an example of the convergence curve
818 of GRR.

819 I. Experimental Results and Discussions

820 From the experimental results listed in Tables and the
821 figures presented in previous subsections, we can draw the
822 following interesting points:

- 823 1. All the experimental results indicate that GRR outperforms
824 the other methods. This is because GRR integrates



multiple robust factors and locality based on the combination of RR, LPP and the $L_{2,1}$ -norm minimization.
825 Thus, with these advantages, GRR is capable to outperform the conventional RR and LPP as well as the
826 L_1 -norm based methods.
827

- 828 2. From Fig. 4 (b), 6 and 8 (b), we can know that the projection learned from the traditional ridge regression
829 RR is no more than the number of the class in training
830 data while GRR can learn any number of projections
831 and preserve high and stable recognition rate. Especially,
832 Table VI showed that when RR obtained the best recogni-
833 tion rate on the dimension of 62 (i.e. the number of
834 class), GRR broke out this number and obtained the
835 best recognition rate on the dimension of 72. This is
836 because RR has the small-class problem which makes
837 the number of the projections limited by the number of
838 the class in training data. However, GRR can break out
839 this limitation to obtain enough projections to obtain
840 high recognition rate for face recognition, character
841 recognition or other practical applications.
842
- 843 3. GRR obtains quite good performance on these
844 databases when there are variations on pose and
845 face expressions. The reason is that GRR not
846 only uses the $L_{2,1}$ - norm on both the loss function
847 and regularization term, but also takes the
848 local geometric structure into consideration. These
849 techniques guarantee the joint sparsity and local
850 information for GRR to obtain effective features
851 to increase the recognition rates.
852
- 853 4. As demonstrated in Subsection IV-F, when training
854 images are with random block corruption, dis-
855 guise or illumination variation, GRR still outperforms
856 the other methods and obtains the better recogni-
857 tion rate in most cases. Comparing with the other methods,
858 another potential reason for the good performance of
859 GRR is that the elastic factor \hbar on the loss function is
860 capable to decrease the negative influence caused by the
861 block subtraction, noise or disguise.
862
- 863 5. In most cases, the $L_{2,1}$ -norm based methods (i.e. GRR
864 and UDFS) obtain the better performance than
865 the L_1 -norm based methods (i.e. PCA-L1, LPP-L1,



Fig. 13. Example images from LFW database.

TABLE X
THE PERFORMANCE OF DIFFERENT METHODS ON THE LFW FACE DATABASE

Training samples	SPCA	LPP	RR	PCA-L1	LDA-L1	LPP-L1	UDFS	KPCA	SRC-L1LS	GRR
4	95.29	64.68	93.40	95.35	61.16	86.29	94.86	95.83	94.53	95.88
5	95.42	72.38	94.04	95.59	69.20	90.78	94.93	96.13	95.45	96.87
6	95.79	78.94	94.48	95.82	76.82	87.80	95.34	96.49	95.96	97.01
7	95.82	64.14	94.80	95.85	83.19	87.60	95.38	96.48	96.10	97.58

866 LDA-L1). The main reason is that using $L_{2,1}$ -norm on
 867 the regularization term can improve the recognition rates
 868 since it performs the joint sparse feature selection in
 869 extraction step.

870 6. The state-of-the-art recognition rate on FERET, ORL,
 871 Yale, AR and LFW database is 71.25%, 98.90%,
 872 98.80%, 91.23% and 53.39% when 4, 5, 5, 5, 6
 873 images of each subject are used as training, respectively [44]–[46]. The newest maximum recognition rate
 874 on Char74K_15 is 67.00% [47]. From the experimental
 875 results, we can find that the recognition rates of
 876 the proposed GRR are higher than the state-of-the-art
 877 recognition rates on ORL, Yale and AR database and
 878 also competitive on other databases. This indicates that
 879 GRR is an effective method for feature extraction and
 880 classification.

V. CONCLUSION

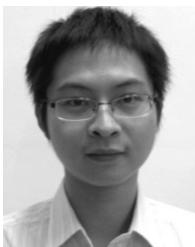
883 Motivated by the robustness and $L_{2,1}$ -norm minimization
 884 on loss function and regularization term, a novel method
 885 called Generalized Robust Regression (GRR) is proposed
 886 for jointly sparse subspace learning. GRR integrates the
 887 advantages of locality preserving projections and $L_{2,1}$ -norm
 888 minimization to select informative features and at the same
 889 time guarantee the robustness in the learning procedure by
 890 introducing an elastic factor. GRR also breaks out the small-
 891 class problem which exists in the traditional ridge regression
 892 or its derivatives so as to obtain enough projections to
 893 perform efficient feature extraction or selection. To optimize
 894 the problem of GRR, an iterative algorithm is proposed
 895 and the convergence is also proved in this paper. Moreover,
 896 we also analyze the computational complexity of the proposed
 897 algorithm. The favorable performance of GRR on six well-
 898 known databases indicates that GRR outperforms the conventional
 899 Sparse Principal Component Analysis (SPCA), Locality
 900 Preserving Projections (LPP), the traditional Ridge regres-
 901 sion (RR), the L_1 -norm based methods PCA-L1, LDA-L1,
 902 LPP-L1, the $L_{2,1}$ -norm regularized discriminative fea-
 903 ture selection method UDFS, the nonlinear kernel-based

method (KPCA) and the classical sparse learning method
 904 (i.e. SRC-L1LS).

905 Since the proposed GRR is an iteration algorithm,
 906 its computational complexity is more than the traditional
 907 LPP-based methods. Also, the case of imbalanced data is not
 908 taken into consideration in this paper. Therefore, how to reduce
 909 the computation cost as well as how to extend the regression
 910 method to deal with the imbalanced data is an interesting
 911 research in the near future.

REFERENCES

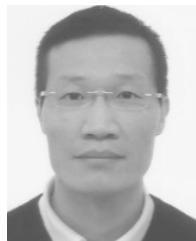
- [1] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, “Discriminative least squares regression for multiclass classification and feature selection,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.
- [2] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang, “A two-phase test sample sparse representation method for use with face recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1255–1262, Sep. 2011.
- [3] Y. Huang, D. Xu, and F. Nie, “Patch distribution compatible semisupervised dimension reduction for face and human gait recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 3, pp. 479–488, Mar. 2012.
- [4] N. R. Draper and R. C. Van Nostrand, “Ridge regression,” *Quality*, vol. 21, no. 4, pp. 451–466, 2011.
- [5] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2004.
- [6] F. Zhong and J. Zhang, “Linear discriminant analysis based on L1-norm maximization,” *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.
- [7] Y. Pang and Y. Yuan, “Outlier-resisting graph embedding,” *Neurocomputing*, vol. 73, nos. 4–6, pp. 968–974, 2010.
- [8] Y. Xu, Z. Zhang, G. Lu, and J. Yang, “Approximately symmetrical face images for image preprocessing in face recognition and sparse representation based classification,” *Pattern Recognit.*, vol. 54, pp. 68–82, Jun. 2016.
- [9] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, “Regularized discriminative spectral regression method for heterogeneous face matching,” *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 353–362, Jan. 2013.
- [10] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [11] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [12] J. B. Tenenbaum, V. de Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Sci.*, vol. 290, no. 5500, pp. 2319–2323, 2000.

- AQ:2
- [13] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.
 - [14] 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.
 - [15] Y. Pang, L. Zhang, Z. Liu, N. Yu, and H. Li, "Neighborhood preserving projections (NPP): A novel linear dimension reduction method," in *Advances in Intelligent Computing*, vol. 3644. 2005, pp. 117–125.
 - [16] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1208–1213.
 - [17] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, Mar. 2012.
 - [18] Y. Pang, X. Li, and Y. Yuan, "Robust tensor analysis with L1-norm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 2, pp. 172–178, Feb. 2010.
 - [19] E. J. Cand, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2009, Art. no. 11.
 - [20] Z. Lin and M. Chen, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Math. Program.*, vol. 9, 2010.
 - [21] Q. Ke and T. Kanade, "Robust L₁ norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 739–746.
 - [22] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: rotational invariant L₁-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 281–288.
 - [23] 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.
 - [24] X. Li, W. Hu, H. Wang, and Z. Zhang, "Linear discriminant analysis using rotational invariant L₁ norm," *Neurocomputing*, vol. 73, nos. 13–15, pp. 2571–2579, 2010.
 - [25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
 - [26] L. Wang, H. Wu, and C. Pan, "Manifold regularized local sparse representation for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 651–659, Apr. 2015.
 - [27] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.
 - [28] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_1, \ell_2, 1$ -norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
 - [29] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proc. Int. Joint. Conf. Artif. Intell.*, 2011, pp. 1324–1329.
 - [30] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1294–1299.
 - [31] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, " $\ell_2, 1$ -Norm regularized discriminative feature selection for unsupervised learning," in *Proc. 22nd Int. Joint. Conf. Artif. Intell.*, 2011, pp. 1589–1594.
 - [32] V. Sindhwani, P. Niyogi, and M. Belkin, "Linear manifold regularization for large scale semi-supervised learning," in *Proc. Int. Conf. Mach. Learn. Workshop Learn. Partially Classified Training Data*, 2005, pp. 80–83.
 - [33] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.
 - [34] 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, Nov. 2014.
 - [35] W. Yu, X. Teng, and C. Liu, "Face recognition using discriminant locality preserving projections," *Image Vis. Comput.*, vol. 24, no. 3, pp. 239–248, 2006.
 - [36] 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.
 - [37] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *Dept. Elect. Comput. Eng., State Univ. New York Buffalo, Amherst, NY, USA*, Tech. Rep. NISTIR 6264, 1997, pp. 137–143.
 - [38] A. A. Martinez and R. Benavente, "The AR face database," *CVC, Barcelona, Spain, Tech. Rep. #24*, 1998.
 - [39] T. E. de Campos, B. R. Babu, and M. Varma, "Character recognition in natural images," in *Proc. 4th Int. Conf. Comput. Vis. Theory Appl.*, 2009, pp. 1–8.
 - [40] C.-P. Wei, C.-F. Chen, and Y.-C. F. Wang, "Robust face recognition with structurally incoherent low-rank matrix decomposition," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3294–3307, Aug. 2014.
 - [41] C. Georgakis, Y. Panagakis, and M. Pantic, "Discriminant incoherent component analysis," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2021–2034, May 2016.
 - [42] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *Univ. Massachusetts, Amherst, MA, USA*, Tech. Rep. 07-49, 2007.
 - [43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
 - [44] A. Nazari and S. B. Shouraki, "A constructive genetic algorithm for LBP in face recognition," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal.*, Apr. 2017, pp. 182–188.
 - [45] T. Alabaidi and W. B. Mikhael, "Two-step feature extraction in a transform domain for face recognition," in *Proc. IEEE Comput. Commun. Workshop Conf.*, Jan. 2017, pp. 1–4.
 - [46] J. Liang, C. Chen, Y. Yi, X. Xu, and M. Ding, "Bilateral two-dimensional neighborhood preserving discriminant embedding for face recognition," *IEEE Access*, vol. 5, pp. 17201–17212, 2017.
 - [47] Z. Zhang, Y. Xu, and C.-L. Liu, "Natural scene character recognition using robust Pca and sparse representation," in *Proc. 12th IAPR Int. Workshop Document Anal. Syst.*, Apr. 2016, pp. 340–345.
- AQ:3
- AQ:4
- Zhihui Lai** received the B.S. degree in mathematics from South China Normal University, in 2002, the M.S. degree from Jinan University, in 2007, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, China, in 2011. He was a Research Associate, a Post-Doctoral Fellow, and a Research Fellow with The Hong Kong Polytechnic University. He has authored over 60 scientific articles. His current research interests include face recognition, image processing, and content based image retrieval, pattern recognition, compressive sense, human vision modulation, and applications in the fields of intelligent robot research. He is currently an Associate Editor of the *International Journal of Machine Learning and Cybernetics*.
- 
- Dongmei Mo** received the B.S. degree from Zhao Qing University. She is currently pursuing the M.S. degree with Shenzhen University. She is currently with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen. Her current research interests include artificial intelligence and pattern recognition.
- 
- Jiajun Wen** received the Ph.D. degree in computer science and technology from the 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 Post-Doctoral Fellow with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include pattern recognition and video analysis.
- 

1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098



Linlin Shen received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 2005. He was a Research Fellow with the Medical School, University of Nottingham, researching brain image processing of magnetic resonance imaging. He is currently a Professor and a Director of the Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His current research interests include Gabor wavelets, face/palmprint recognition, medical image processing, and hyperspectral image classification.



Wai Keung Wong received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 1999. He is currently a Professor with The Hong Kong Polytechnic University. He has authored or co-authored over 100 papers in refereed journals and conferences, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART C: APPLICATIONS AND REVIEWS, *Pattern Recognition*, *CHAOS*, the European Journal of Operational Research, Neural Networks, Applied Soft Computing, Information Science, Decision Support Systems, and among others. His current research interests include pattern recognition and feature extraction.

IEEE Proot