Robust Discriminant Regression for Feature Extraction

Zhihui Lai, Dongmei Mo, Wai Keung Wong, Yong Xu, *Member, IEEE*, Duoqian Miao, and David Zhang, *Fellow, IEEE*

***Abstract*—Ridge regression (RR) and its extended versions are widely used as an effective feature extraction method in pattern recognition. However, the RR-based methods are sensitive to the variations of data and can learn only limited number of projections for feature extraction and recognition. To address these problems, we propose a new method called robust discriminant regression (RDR) for feature extraction. In order to enhance the robustness, the $L_{2,1}$-norm is used as the basic metric in the proposed RDR. The designed robust objective function in regression form can be solved by an iterative algorithm containing an eigenfunction, through which the optimal orthogonal projections of RDR can be obtained by eigen decomposition. The convergence analysis and computational complexity are presented. In addition, we also explore the intrinsic connections and differences between the RDR and some previous methods. Experiments on some well-known databases show that RDR is superior to the classical and very recent proposed methods reported in the literature, no matter the L_2-norm or the $L_{2,1}$-norm-based regression methods. The code of this paper can be downloaded from http://www.scholat.com/laizhihui.**

***Index Terms*—Feature extraction, linear regression, reduction, robust dimensionality, subspace learning.**

Manuscript received January 7, 2017; revised April 12, 2017 and June 26, 2017; accepted August 7, 2017. Date of publication October 9, 2017; date of current version July 17, 2018. This work was supported in part by the Natural Science Foundation of China under Grant 61573248, Grant 61375012, Grant 61362031, Grant 61332011, Grant 61370163, Grant 61773328, and Grant 61732011, in part by the Hong Kong Polytechnic University under Project G-UC42 and Project G-YBD9, in part by the Natural Science Foundation of Guangdong Province through the Tensor Presentation Based Sparse Feature Extraction Project, and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20150324141711637 and Grant JCYJ20170302153434048. This paper was recommended by Associate Editor L. Shao. *(Corresponding author: Wai Keung Wong.)*

Z. Lai and D. Mo are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Institute of Textiles and Clothing, Hong Kong Polytechnic University, Hong Kong (e-mail: lai_zhi_hui@163.com; dongmei_mo@qq.com).

W. K. Wong is with the Institute of Textiles and Clothing, Hong Kong Polytechnic University, Hong Kong, and also with Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518055, China (e-mail: calvin.wong@polyu.edu.hk).

Y. Xu is with the Bio-Computing Research Center and the Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China.

D. Miao is with the Department of Computer Science and Technology, College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China.

D. Zhang is with the Biometrics Research Centre, Department of Computing, Hong Kong Polytechnic University, Hong Kong.

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/TCYB.2017.2740949

I. Introduction

LEAST square regression is one of the most simple and effective methods for feature extraction and dimensionality reduction. The classical least square regression method, i.e., ridge regression (RR) [1], has the potential drawbacks, including the sensitivity to the data's variations, and obtaining only limited number of projections, which is equal to the number of classes in the training data. Based on the minimum mean squared error, many regression methods were proposed and modified to different applications [2]–[7]. The developments of the regression methods can be generally divided into three categories: 1) sparse regression extension; 2) subspace regression; and 3) robust regression. Each of them represents one development route of the regression methods.

On the first development route, the basic regression, i.e., RR, has been developed to be different sparse regularized regression methods. The representative methods include lasso regression [8], least angle regression [9], and elastic net regression [10]. These regression methods focus on the variable selection or feature selection. Since the focus of this paper is not on the sparse variable selection, the readers are referred to [8]–[10] for more details.

The second development route of the least square regression is the generalized variations, i.e., subspace regression methods. In the past decades, it was shown that many linear dimensionality reduction methods could be reformulated by least square regression techniques. For example, principle component analysis (PCA) [11]–[13] can be rewritten as regression form and developed to be a sparse PCA [14]. Ye [15] represented the linear discriminant analysis (LDA) [16]–[18] to a least square LDA. Recently, De la Torre [19] reformulated many component analysis methods into a unified least squares framework and Zheng *et al.* [20] extended the regularization method for cross-modal hashing. A common property of these subspace regression methods is that the label indicator matrix usually is not directly used in the regression procedures, which is different from the classical RR. However, since these methods used the L_2-norm or Frobenius norm as the metric, they are sensitive to the outliers or the data's variations [21]–[23].

The last development route of the regression methods is the robust extensions of the least square regression, which is the main focus of this paper. In recent years, the robust regression methods have been paid great attention. The robust least square regression methods mainly introduce the $L_{2,1}$-norm as the basic metric in the regression. Nie *et al.* [21] proposed

an efficient iterative algorithm to solve the $L_{2,1}$-norm regression problem, which is called robust feature selection (RFS). It was shown that this regression model is more robust in pattern recognition [24] than the sparse representation classifier [25]. Based on this fast method proposed in [21], many robust regression methods based on $L_{2,1}$-norm were proposed for subspace learning [22], [23], [26]–[29]. Since the $L_{2,1}$-norm regularized regression can derive jointly sparse solutions, it is usually used as a regularizer. Therefore, Gu *et al.* [27] proposed the joint feature selection and subspace learning (FSSL) method based on the locality preserving projections (LPP) [30]. In addition, $L_{2,1}$-norm was widely used in shared subspace learning [28] for face recognition [31] and some other application [22], [23], [29]. These studies indicate that the $L_{2,1}$-norm-based regression methods are more robust than the L_2-norm or Frobenius norm-based regressions on subspace learning.

As the RR uses the label indicator matrix in regression, it can obtain only C projections, where C denotes the number of the classes. However, since the subspace regression methods do not use the label indicator matrix, they usually can obtain more projection than the classical ridge regression. The drawback in ridge regression and subspace regression methods is that they use the L_2-norm or Frobenius as the distance metric, which is sensitive to the outliers and data's variations [21]–[23]. Despite of the robustness indicated in the $L_{2,1}$-norm, the problems in the robust regression methods, replacing the L_2-norm by the $L_{2,1}$-norm in ridge regression and its variations such as those methods in [22], [23], [28], and [29], is that only limited number of projections can be obtained due to the usages of the label indicator matrix. Particularly, when C is small, the performance of the existing robust regression methods will be degraded.

Robustness is critical in computer vision and pattern recognition. Obtaining the robust features is significant in recognition task. In this paper, we take the advantages of the methods in the second and third categories (i.e., subspace regression and robust regression, respectively) to develop a novel framework so as to deal with the robustness in regression learning. The main contributions of this paper are as follows.

1) We propose a novel regression learning method, i.e., robust discriminant regression (RDR), for discriminant subspace learning. Based on the $L_{2,1}$-norm as the robust metric, an iterative method is proposed to solve the regression learning problem.
2) The convergence and computational complexity analyses are presented. The theoretical similarity and differences between RDR and the classical methods are also explored.
3) Extensive experiments show that the proposed RDR performs better than the related methods for robust image feature extraction.

The rest of this paper is organized as follows. Section II summarizes the previous regression methods. In Section II, RDR is proposed for feature extraction. In Section III, theoretical analyses are presented to explore properties of the RDR algorithm and its relationship to other algorithms. In Section IV, experiments on evaluating the proposed method are reported. We give the conclusions in Section V.

II. Background and Motivations

In this section, we first give the notations used in this paper and the definition of $L_{2,1}$-norm, and then the summarize the previous regression methods. At last we present potential problem of the existing regression methods which motivates the proposed RDR algorithm.

A. Notation and Definition

Let matrix $X = [x_1, x_2, \ldots, x_N]^T$ be the data matrix including all the training samples $\{x_i\}_{i=1}^N \in R^m$ in its rows. Let $\{y_i\}_{i=1}^N \in R^C$ denote the label of the training data, where C denotes the number of classes in training set.

Since the feature dimension m is often very high, we need to find a projection matrix $Q = (q_1, q_2, \ldots, q_d) \in R^{m\times d}$ to map the sample $x \in R^m$ into $\tilde{x} \in R^d$ ($d << m$) by using

$$\tilde{x} = xQ \in R^d. \tag{1}$$

In this paper, we need the definition of $L_{2,1}$-norm for clarity, the definition of $L_{2,1}$-norm of a matrix $A \in R^{n\times m}$ is defined as follows:

$$\|A\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} a_{ij}^2} = \sum_{i=1}^{n} \left\|a^i\right\|_2 \tag{2}$$

where a^i denotes the ith row vector of A.

B. Summary of Regularization Regression Methods

The most representative regularization regression method is the RR [1] and its variations [15], [22], [27]. They can be concluded by the following optimization problem:

$$\min_f \sum_i \text{loss}(f(x_i), y_i) + \alpha R(f) \tag{3}$$

where loss$(\cdot)$ is a loss function and $R(f)$ is a regularization function on f with α as a regularization parameter.

However, the RR can only obtain C projections for feature extraction. A tractable method for solving this problem is to give up the usage of the label indicator matrix in the regression model. Therefore, we obtain the modified regularization regression model as follows:

$$\min_{f,g} \sum_i \sum_j \text{loss}\big(f\big(g\big(x_j\big)\big), x_i\big) + \alpha R(f) \tag{4}$$

where g is another function of the dataset. Previous works such as [15], [23], and [28] can be concluded in this regression model.

C. Problem of Previous Regression Methods and the Motivation of RDR

As indicated in previous research, the high-dimensional data, such as face images and handwriting digital images, lie on the ambient low-dimensional manifold embedded in high-dimensional space [32]–[35]. Thus the local geometric structure plays an important role in learning the low-dimensional

subspace. However, no matter the methods summarized in (3) use or do not use the local geometric structure information, they are sensitive to the outliers since the L_2-norm is used as the metric. Although more robust norm is used in (4), the methods concluded in (4) usually ignore the local geometric structure information, which is very important to improve the algorithms' performance [20], [36]–[39]. Therefore, how to integrate the local geometric structure information into the regression model and at the same time to improve the robustness is an important issue.

Previous research [32]–[34] showed that geometric structure or class-specific structure is very important in dimensionality reduction. This information can help us find the meaningful low-dimensional representations and enhance the representative ability. For example, local linear embedding [33] and the neighborhood preserving embedding (NPE) [40] can preserve the local reconstruction relationship among the data. The Laplacian eigenmaps [34], LPP, and its variations [36]–[39] can preserve the local nearest neighborhood relationship among the data points. By introducing the discriminant information, the modifications of NPE and LPP learn the low-dimensional subspace for supervised feature extraction, which have been proved to have stronger discriminant ability than the classical LDA in face recognition and objective recognition [41]. The key idea of these methods is to define a graph and then preserve the information characterized by the graph in low-dimensional subspace. We also follow this idea and introduce the graph in the proposed RDR model.

In order to address the locality preserving problem among the data points and improve the robustness in feature extraction, we integrate the local geometric structure into the regression model and use a robust norm as the main measurement. Therefore, we proposed the novel generalized extension of the regression model by introducing the locality of the dataset

$$\min_{f,g} \sum_i \sum_j \text{loss}\big(f\big(g\big(x_j\big)\big), x_i, W_{ij}\big) + \alpha R(f) \tag{5}$$

where matrix W characterizes the local manifold structure of the dataset. The user can use different regularization methods and append different constraints on the regularized framework (5) to design effective methods. In this paper, we focus on developing a representative method with orthogonal constraint on one of the function f based on the $L_{2,1}$-norm as the metric for robustness feature extraction.

III. Robust Discriminant Regression

In this section, the objective function of the proposed RDR and its formulation are first given. Then we propose an alternatively iterative algorithm to compute the optimal solution.

A. Objective Function of RDR and Its Formulation

As indicated in Section II-C, the manifold learning-based linear dimensionality reduction methods aim to minimize the weighted L_2-norm objective function, in which the objective function is sensitive to the outliers or the data's variations. However, in the proposed RDR, the local geometric structure is preserved by minimizing the weighted $L_{2,1}$-norm loss function to enhance the robustness of the algorithm. From the generalized regression model (5), it is supposed that the functions f and g are linear projections. Then we propose the following concrete objective function of RDR derived from the generalized extension of the regression model (5) with orthogonal constraint:

$$\begin{aligned}\min_{P,Q} J(P,Q) &= \min_{P,Q} \sum_i \sum_j \left\|x_i - x_j QP\right\|_2 W_{ij} + \alpha \|P\|_F^2 \\ &= \min_{P,Q} \sum_i \sum_j \left\|x_i - x_j QP\right\|_{2,1} W_{ij} + \alpha \|P\|_F^2 \end{aligned} \tag{6}$$

$$\text{s.t. } Q^T Q = I \tag{7}$$

where $P \in R^{d\times m}$ is a matrix and α is the regularization parameter. Note that for any row vector x, $\|x\|_2 = \|x\|_{2,1}$, it is found that the significant difference between (6) and other manifold learning-based methods such as LPP and NPE is that (6) uses the $L_{2,1}$-norm as the basic metric, in which the element $\|x_i - x_j QP\|_2$ is not squared. Previous research [21]–[23] indicated that using the $L_{2,1}$-norm as the basic metric makes the outliers with less level of importance than the squared term $||x_i - x_j QP||_2^2$. Minimizing the first term of (6) indicates that the data $x_i$is expected to be represented by the $x_j QP$ and keep the data similarity or geometric structure defined by W. In other words, if x_i and x_j is close to each other, then x_i and $x_j QP$ should also be closed to each other under the learned mapping QP. The regularized term $\alpha\|P\|_F^2$ is used for strengthening the algorithm's stability in computing the optimal solution and improving the model's generalization in feature extraction.

From (6), we have the following formulation using the same technique proposed in [21]:

$$\begin{aligned} J(P,Q) &= \sum_i \sum_j \left\|x_i - x_j QP\right\|_{2,1} W_{ij} + \alpha \|P\|_F^2 \\ &= \sum_i \sum_j \text{tr}\big[(x_i - x_j QP)^T G_{ij} W_{ij} \big(x_i - x_j QP\big)\big] \\ &\quad + \alpha \|P\|_F^2 \\ &= \sum_i \sum_j \text{tr}\Big[x_i^T G_{ij} W_{ij} x_i - 2P^T Q^T x_i^T G_{ij} W_{ij} x_j \\ &\qquad + P^T Q^T x_j^T G_{ij} W_{ij} x_j QP\Big] + \alpha \|P\|_F^2 \end{aligned} \tag{8}$$

where

$$G_{ij} = \frac{1}{2\left\|x_i - x_j QP\right\|_2}. \tag{9}$$

We define

$$F = G \odot W \tag{10}$$

where $\odot$ denotes the matrix element wise multiplication, and diagonal matrix D with its diagonal elements is computed as

$$D_{ii} = \sum_j F_{ij}. \tag{11}$$

With the above notations, we obtain the following formulations from (8):

$$\begin{aligned}\min_{P,Q} J(P,Q) &= \min_{P,Q} \sum_i \sum_j \left\| x_i - x_j QP \right\|_{2,1} W_{ij} + \alpha \|P\|_F^2 \\ &= \min_{P,Q} \operatorname{tr}\big(X^T DX - 2P^T Q^T X^T (G \odot W) X \\ &\qquad + P^T Q^T X^T DXQP + \alpha P^T P\big) \\ &= \min_{P,Q} \operatorname{tr}\big(X^T DX - 2P^T Q^T X^T FX \\ &\qquad + P^T Q^T X^T DXQP + \alpha P^T P\big) \\ \text{s.t. } & Q^T Q = I. \end{aligned} \tag{12}$$

Thus, the optimization problem becomes a trace minimization problem with the orthogonal constraint $Q^T Q = I$. In the following section, we will show how to solve the optimization problem.

B. Optimal Solution

The optimization problem (12) with the orthogonal constraint $Q^T Q = I$ has two variables. To the best of our knowledge, there has no close form solution. Therefore, we design an iterative algorithm to compute the optimal solutions. The algorithm steps to solve the optimization problems are as follows: first, we fix Q to compute the optimal P, then fix P to compute the optimal Q. Iterating these two steps until the convergence of the objective function will give the optimal solutions of the optimization problem (12). Since in the iterations we always compute the matrix D using the last updated P and Q (which are known), $\operatorname{tr}(X^T DX)$ becomes a constant in (12) at that time and can be ignored in computing the current optimal solution with respect to variable P and Q (please see the updating procedures in Table I). Thus the optimization (12) can be further converted to the following minimization problem:

$$\begin{aligned}\min_{P,Q} \ & \operatorname{tr}\big(-2P^T Q^T X^T FX + P^T Q^T X^T DXQP + \alpha P^T P\big) \\ \text{s.t. } & Q^T Q = I. \end{aligned} \tag{13}$$

For the given Q, we take the partial deviations of (12) with respect to P and set it to be 0, it is easy to have

$$\begin{aligned} P &= \big(Q^T X^T DXQ + \alpha I_d\big)^{-1} Q^T X^T FX \\ &= A^{-1} Q^T X^T FX \end{aligned} \tag{14}$$

where $A = Q^T X^T DXQ + \alpha I_d$.

For the optimal Q, we have the following theorem.

Theorem 1: In each iteration, the optimal Q that solves the optimization problem in (6) and (7) is given by the following trace maximization problem:

$$\max_{Q^T Q = I_d} \operatorname{tr}\Big(\big(Q^T \big(X^T DX + \alpha I\big) Q\big)^{-1} Q^T X^T FXX^T FXQ\Big). \tag{15}$$

Proof: Substituting (14) back to (13), we have the following optimization problem:

$$\begin{aligned}\min_{Q^T Q = I_d} \ & \operatorname{tr}\Big(-2X^T F^T XQA^{-1} Q^T X^T FX \\ &\quad + X^T F^T XQA^{-1} Q^T X^T DXQA^{-1} Q^T X^T FX \\ &\quad + \alpha X^T F^T XQA^{-1} A^{-1} Q^T X^T FX\Big) \\ = \min_{Q^T Q = I_d} \ & \operatorname{tr}\Big(-2X^T F^T XQA^{-1} AA^{-1} Q^T X^T FX \\ &\quad + X^T F^T XQA^{-1} Q^T X^T DXQA^{-1} Q^T X^T FX \\ &\quad + \alpha X^T F^T XQA^{-1} A^{-1} Q^T X^T FX\Big) \\ = \min_{Q^T Q = I_d} \ & \operatorname{tr}\Big(X^T F^T XQA^{-1} \big(-2A + Q^T X^T DXQ + \alpha I_d\big) \\ &\quad \times A^{-1} Q^T X^T FX\Big) \\ = \min_{Q^T Q = I_d} \ & \operatorname{tr}\Big(X^T F^T XQA^{-1} (-2A + A) A^{-1} Q^T X^T FX\Big) \\ = \min_{Q^T Q = I_d} \ & -\operatorname{tr}\Big(X^T F^T XQA^{-1} Q^T X^T FX\Big) \\ = \min_{Q^T Q = I_d} \ & -\operatorname{tr}\Big(A^{-1} Q^T X^T FXX^T F^T XQ\Big) \\ = \min_{Q^T Q = I_d} \ & -\operatorname{tr}\Big(\big(Q^T \big(X^T DX + \alpha I\big) Q\big)^{-1} \\ &\quad \times Q^T X^T FXX^T F^T XQ\Big) \\ = \max_{Q^T Q = I_d} \ & \operatorname{tr}\Big(\big(Q^T \big(X^T DX + \alpha I\big) Q\big)^{-1} \\ &\quad \times Q^T X^T FXX^T F^T XQ\Big). \end{aligned}$$

This minimization problem is equivalent to the maximum problem in (15). ■

The solution of optimization problem (15) can be obtained by solved by the following eigen decomposition [42]:

$$\big(X^T DX + \alpha I\big)^{-1} \big(X^T FXX^T F^T X\big) q = \lambda q \tag{16}$$

where λ is the eigenvalue corresponding to eigenvector q. The first d eigenvectors corresponding to the larger eigenvalues are the optimal solutions for variable Q, which is used for feature extraction. However, since the matrix G in $F = G \odot W$ is related to the unknown Q and P, the optimization problem cannot be directly solved. Therefore, we present the details of the iterative method, which is shown in Table I, to obtain the optimal solutions of the proposed model.

IV. Algorithm Analysis and Comparison

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

A. Convergence Analysis

Since the proposed algorithm is an iterative method, we need to prove the convergence of the algorithm. First, we need the following lemma [21].

Lemma 1: For any nonzero vector a and b, the following inequality holds:

$$\|a\| - \frac{\|a\|^2}{2\|b\|} \le \|b\| - \frac{\|b\|^2}{2\|b\|}. \tag{17}$$

From Lemma 1, it is easy to obtain the following corollary.

TABLE I
RDR ALGORITHM

Input: Training samples $\{x_i \in R^m, i = 1,2,...,N\}$, the numbers of iterations T, dimensions d
Output: Low-dimensional features $\tilde{x}_i$ $(i = 1,2,...,N)$
Step 1: Initialize $G = I$.
Step 2: For $t = 1:T$
-Construct matrix $X^T FXX^T FX$ and $(X^T DX + \alpha I)$.
- Solve the (15) to obtain the optimal Q.
- Compute P using (14).
- Update G, F and D using (9), (10) and (11), respectively.
End
Step 3: Project the samples onto the low-dimensional subspace to obtain $\tilde{x}_i = x_i Q$ for classification.

Corollary 1: For any nonzero vectors $a_j^i, b_j^i (i, j = 1, 2, \ldots, N)$, the following inequality holds:

$$\sum_i \sum_j \left\|a_j^i\right\| - \frac{\left\|a_j^i\right\|^2}{2\left\|b_j^i\right\|} \le \sum_i \sum_j \left\|b_j^i\right\| - \frac{\left\|b_j^i\right\|^2}{2\left\|b_j^i\right\|}. \tag{18}$$

Suppose in the tth iteration, we have the following new notation from (13):

$$J(P_t, Q_t, G_t) = \mathrm{tr}\left(-2P_t^T Q_t^T X^T F_t X + P_t^T Q_t^T X^T D_t X Q_t P_t + \alpha P_t^T P_t\right)$$

where G_t is the intrinsic variable related to F_t and D_t. With above preparations, the following theorem can be obtained.

Theorem 2: The iterative scheme in Algorithm 1 monotonically decreases the objective function value of $J(P_t, Q_t, G_t)$ in each iteration.

Proof: When Q_t and G_t are given, we know that $P_{t+1} = A_t^{-1} Q_t^T X^T D_t X Q_t + \alpha I_d$ minimizes the objective function value $J(P_t, Q_t, G_t)$. Thus we have

$$J(P_{t+1}, Q_t, G_t) \le J(P_t, Q_t, G_t). \tag{19}$$

On the other hand, since solving the eigenfunction provides the optimal solution, the objective function value is further reduced. Thus we obtain

$$J(P_{t+1}, Q_{t+1}, G_t) \le J(P_t, Q_t, G_t). \tag{20}$$

At last, we need to prove that

$$J(P_{t+1}, Q_{t+1}, G_{t+1}) \le J(P_t, Q_t, G_t). \tag{21}$$

Since

$$\begin{aligned}&\sum_i \sum_j \mathrm{tr}\left[\left(x_i - x_j Q_{t+1} P_{t+1}\right)^T G_{t,ij} W_{ij} \left(x_i - x_j Q_{t+1} P_{t+1}\right)\right] + \alpha \|P_{t+1}\|_F^2 \\ &\le \sum_i \sum_j \mathrm{tr}\left[\left(x_i - x_j Q_t P_t\right)^T G_{t,ij} W_{ij} \left(x_i - x_j Q_t P_t\right)\right] + \alpha \|P_t\|_F^2.\end{aligned} \tag{22}$$

That is

$$\begin{aligned}&\sum_i \sum_j \frac{\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_{t+1} P_{t+1}\right)\right\|_2^2}{2\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_2^2} + \alpha \|P_{t+1}\|_F^2 \\ &\le \sum_i \sum_j \frac{\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_2^2}{2\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_2^2} + \alpha \|P_t\|_F^2.\end{aligned} \tag{23}$$

From Corollary 1, we have

$$\begin{aligned}&\sum_i \sum_j \left\|\sqrt{W_{ij}}\left(x_i - x_j Q_{t+1} P_{t+1}\right)\right\|_2 - \frac{\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_{t+1} P_{t+1}\right)\right\|_2^2}{2\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_2^2} \\ &\le \sum_i \sum_j \left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_2^2 - \frac{\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_2^2}{2\left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_2^2}.\end{aligned} \tag{24}$$

Combining (23) and (24), we obtain

$$\begin{aligned}&\sum_i \sum_j \left\|\sqrt{W_{ij}}\left(x_i - x_j Q_{t+1} P_{t+1}\right)\right\|_{2,1} + \alpha \|P_{t+1}\|_F^2 \\ &\le \sum_i \sum_j \left\|\sqrt{W_{ij}}\left(x_i - x_j Q_t P_t\right)\right\|_{2,1} + \alpha \|P_t\|_F^2.\end{aligned} \tag{25}$$

This indicates that

$$\begin{aligned}&\sum_i \sum_j \mathrm{tr}\left[\left(x_i - x_j Q_{t+1} P_{t+1}\right)^T G_{t+1,ij} W_{ij} \left(x_i - x_j Q_{t+1} P_{t+1}\right)\right] + \alpha \|P_{t+1}\|_F^2 \\ &\le \sum_i \sum_j \mathrm{tr}\left[\left(x_i - x_j Q_t P_t\right)^T G_{t,ij} W_{ij} \left(x_i - x_j Q_t P_t\right)\right] + \alpha \|P_t\|_F^2.\end{aligned} \tag{26}$$

Equation (26) indicates that (21) is satisfied. From (19)–(21), we conclude that the iterative algorithm converges. ■

B. Computational Complexity

We can find that the main computational complexity comes from two parts. The first part is the eigen decomposition of (16), which needs $O(m^3)$. If the algorithm converges

within T iteration steps, the total computational complexity is $O(Tm^3)$, which is very large when the dimension of the samples is very high. The second part is to compute the scatter matrices $X^TDX+\alpha I$ and X^TFXX^TFX in the iteration of the RDR algorithm, which is up to $O(mN^2)$ in single iteration and the algorithm will cost a little more time. Thus, in this case, PCA could be used for predimensionality reduction, which will greatly reduce the computational burden. As shown in the experimental section, the algorithm converges very fast such that the total computational burden is acceptable in the learning procedure. In addition, since the algorithm can be run offline in learning, the additional computational cost is not considered as a distinct disadvantage of the proposed method in feature extraction and pattern recognition task.

C. Properties of RDR and Its Connections to Other Methods

In this section, we indicate some properties/advantages of RDR and show the relations between RDR and the related methods.

One can define the graph W in different modes using class information. In this section, we analyze the properties of the graph which is defined as follows:

$$W_{ij} = \begin{cases} 1/N_c, & \text{if } x_i \text{ and } x_j \text{ belong to the } c\text{th class} \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

where N_c denotes the number of the training samples in cth class. With this graph, due to the high dimensionality of the original data X, most of the linear dimensionality reduction methods can obtain $\text{rank}(X^TWX) = \text{rank}(W) = C$ projections at most. However, in this case we have the following property which is different from previous methods.

Proposition 1: The number of the optimal projection of RDR is at least C. The upper bound of the number of the optimal projection for RDR is N.

Proof: Since $\text{rank}(W) = C$ and G is usually a random symmetric matrix, which is the full-rank matrix without loss of generality, we have

$$\begin{aligned} \text{rank}\left(X^TFXX^TFX\right) &= \text{rank}(F) = \text{rank}(G \odot W) \\ &\le \min\{\text{rank}(G)\} = N. \end{aligned}$$

On the other hand, we have

$$\text{rank}(G \odot W) \ge \text{rank}(W) = C.$$

The "=" satisfies if and only if all the data points in the same class are distributed on a spherical surface and the mapping QP maps them to the corresponding centers of the spheres. ■

Proposition 1 indicates that the proposed regression method at least has C projections for feature extraction. In other words, the number of projections obtained by RDR algorithm exceeds the classical RR and its $L_{2,1}$-norm-based RFS [21], as well as the class LDA. Using the same graph as defined in (27), LPP can only obtain $\text{rank}(W)$ projections in high-dimensional small sample size problem. However, the RDR can learn more projections than LPP in this case, which can be derived by Proposition 1. This is another advantage of RDR. The RDR also has the similar function that can preserve a certain geometric structure related to the graph W. From equation $(X^TDX + \alpha I)^{-1}(X^TFXX^TFX)q = \lambda q$, we can define a new graph or neighborhood matrix as follows:

$$\tilde{W} = FXX^TF = [(G \odot W)X][(G \odot W)X]^T. \tag{28}$$

This graph is not only related to the geometric structure of the predefined graph W, but also related to the weights in G derived by the $L_{2,1}$-norm, as well as the data matrix itself. Therefore, RDR optimally preserves a novel geometric structure characterized by $\tilde{W}$.

We can also find that RDR is the reweighted version of LDA, in which the weights are derived by the $L_{2,1}$-norm. The robustness comes from the different weights defined in $\tilde{W}$, in which the weight matrix G robustly measures the distance of different data points. Thus RDR can perform better than LDA in feature extraction on noisy data.

D. Model Analysis and Comparison

In this section, we compare the proposed model with the following similar models on their performance in feature extraction and classification:

$$\min_{Q} \sum_i \sum_j \left\| x_iQ - x_jQ \right\|_2 W_{ij} \quad \text{s.t. } Q^TQ = I \tag{29}$$

and

$$\min_{P,Q} \sum_i \sum_j \left\| x_i - x_jQP \right\|_2^2 W_{ij} + \alpha \|P\|_F^2 \quad \text{s.t. } Q^TQ = I. \tag{30}$$

It is easy to find that all of RDR, (29) and (30) aims to preserve the locality defined by W. However, the proposed model (6) is essentially different from models (29) and (30). Model (29) only aims to preserve the geometric structure defined by W using the $L_{2,1}$-norm loss function. Model (30) is very similar to (6) but using L_2-norm loss function, which is sensitive to the outliers. The optimal solution of model (29) can be obtained by iteratively solving the following eigen decomposition:

$$X^T(H \odot W)XQ = Q\Lambda \tag{31}$$

where $H_{ij} = 1/\|x_iQ - x_jQ\|_2$ and Λ is the eigenvalue matrix.

Similar to (6), model (30) can also be solved using the iterative algorithm. The optimal solution

$$P = \left(Q^TX^TDXQ + \alpha I_d\right)^{-1} Q^TX^TWX. \tag{32}$$

The optimal Q can be obtained from the following eigenfunction:

$$\left(X^TDX + \alpha I\right)^{-1}\left(X^TWXX^TWX\right)Q = Q\Lambda. \tag{33}$$

Comparing (32) and (33) with (14) and (15), we can find that the L_2-norm model lost the weight matrix generated by the $L_{2,1}$-norm loss function. Therefore, the proposed model (6) will be more robust than (30).

Fig. 1 shows the performance of PCA, LPP, RDR and model (29) and (30) on FERET face database with block noise. From Fig. 1, we can see that RDR performs significantly better than model (30), which indicates that using the

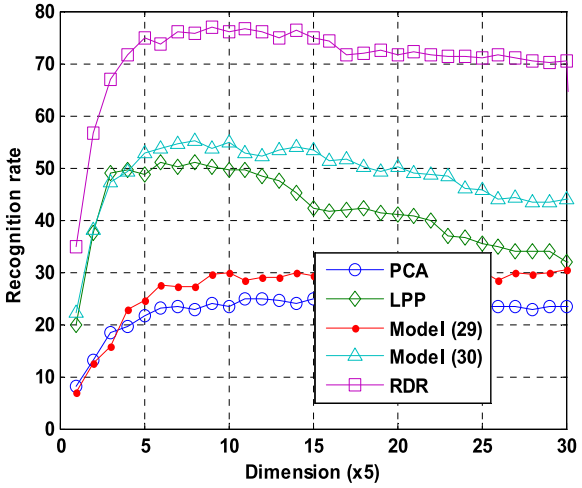


Fig. 1. Performance of different models/methods on FERET face database with 15×15 block noise (the details of the database can be found in Section V-A).

$L_{2,1}$-norm as the distance measurement for the reconstructive term $x_i - x_j QP$ usually performs better than using the L_2-norm. The only reason is that different norms are used in the models. Thus $L_{2,1}$-norm used in the proposed model improves the robustness. Usually, if one model can integrate the reconstructive and discriminant properties together, it will perform better than the ones only using the reconstructive information. Since PCA only uses the reconstructive information, as shown in Fig. 1, it performs poorly in this case. Moreover, even if the discriminant information is used in model (29), the experimental results presented in Fig. 1 shows that RDR still performs significantly better than (29). This indicates that the reconstruction operation P can improve the model generalization in feature extraction to a certain extent.

V. Experiments

Experimental results on six databases will be presented in this section. The code of the proposed RDR can be downloaded from http://www.scholat.com/laizhihui.

A. Details of the Databases

The FERET face database [44] includes 1400 images of 200 individuals (each individual has seven images) with variations in facial expression, illumination, and pose. The images were randomly added with a block with size 5×5, 10×10, and 15×15. Some sample images of one person are shown in Fig. 2(a).

A subset (C29) of the CMU PIE face database [43] containing 1632 images of 68 individuals was used in the experiment. The images were randomly added with the salt and pepper noise with density as 0.03. Fig. 2(b) shows the sample images from this database with different noise densities.

A subset of the AR face database [44] containing 2400 images from 120 individuals were selected and used in our experiments. The pixel values were normalized on 0–255 and no any other preprocessing was performed on the images. The sample images of one person are shown in Fig. 2(c).

PolyU hyperspectral face database [45] (http://www.comp.polyu.edu.hk/~biometrics/hyper_face.htm) is used to test the robustness of the proposed algorithms to the variations of the different hyperspectral images. Fig. 2(d) shows some examples of images of the hyperspectral face. In the experiment, 30 images of 47 individuals (1410 image in total) were used.

Binary alpha digits image database (http://www.cs.nyu.edu/~roweis/data.html) is composed of 1404 binary images of handwritten digits. The resolution of each image is 20×16 pixels and the pixel value is 0 or 1. Some sample images are shown in Fig. 2(e).

PolyU FKP (http://www.comp.polyu.edu.hk/~biometrics/FKP.htm) contains 7920 images from 660 different fingers. The pixel value of the image was normalized to be 0–255. The sample images are shown in Fig. 2(f).

B. Experimental Setup

In the experiments, L images of each individual were randomly selected and used as the training set and the rest for test. We set $L = 3, 4, 5$ for CMU PIE face database and $L = 5$ for FERET, AR, and PolyU hyperspectral face databases, respectively. For binary alpha digits image and PolyU FKP databases, we set $L = 20$ and $L = 2, 4$, respectively.

In the experiments, the proposed method was compared with the L_2 norm-based methods, i.e., PCA, LDA, LPP, RR, and the $L_{2,1}$-norm-based methods, i.e., RFS [21], FSSL [27], semantic analysis via intermediate representation (SAIR) [23], and the recently proposed discriminant elastic-net regularized linear regression (DENLR) [7].

To improve the computational efficiency and avoid the singularity problem, PCA is used as preprocessing method and keep about 98% energy. The subspace dimensions for FERET and AR face databases were varied from 5 to 200 with step 5. For CMU PIE and PolyU hyperspectral face database and FKP database, the numbers of the subspace dimensions were varied from 2 to 150 with step 2. For the binary alpha digits image dataset, the subspace dimensions were varied from 1 to 50.

The neighborhood parameters K were selected from the set $\{1, 2, \ldots, L-1\}$ since the supervised graph was used in the experiments and $K \leq L-1$, and the graph used in the experiments (in LPP, FSSL, and RDR algorithm) was defined as

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \in N_K(x_j) \text{ or } x_j \in N_K(x_i) \\ 0, & \text{otherwise} \end{cases}$$

where $x_i \in N_K(x_j)$ means that x_i is one of the K nearest neighbors of x_j in the same class. The regularization parameter of all the regression methods were selected from $[0.001, 0.01, \ldots, 1000]$.

Tables II–VII show the performances of the algorithms running for ten times. The recognition rates versus the number of the training images or the size of the block occlusions are also shown in Fig. 3.

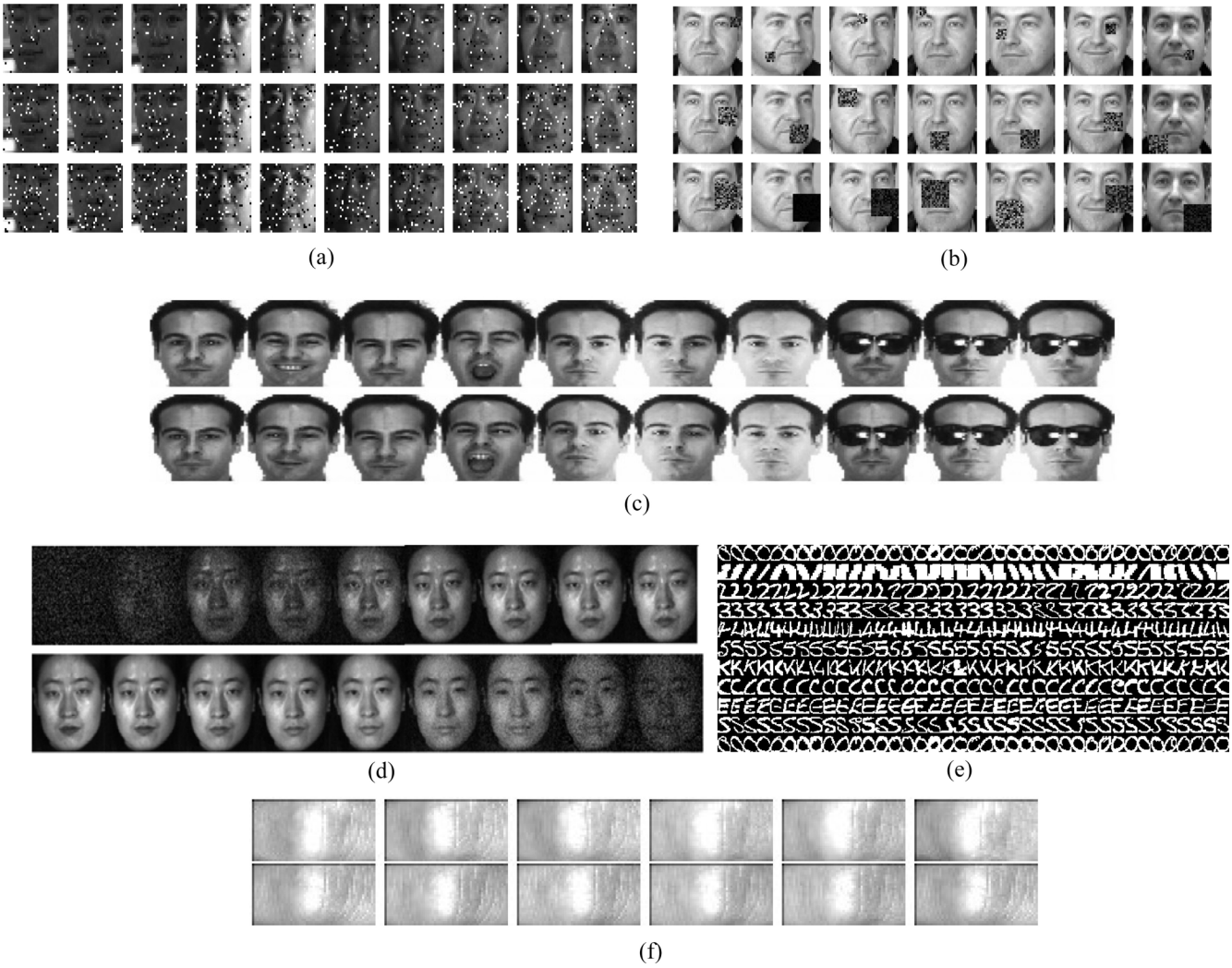


Fig. 2. Image samples used in the experiments. (a) CMU PIE face database. (b) FERET face database. (c) AR face database. (d) PolyU hyperspectral face database. (e) Binary alpha digits image database. (f) PolyU FKP image database.

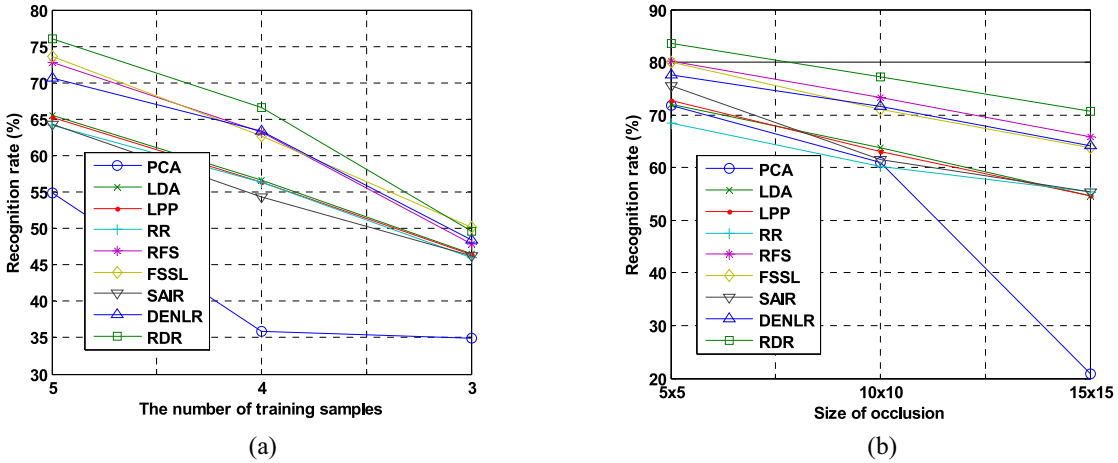


Fig. 3. (a) Average recognition rates (%) versus the variations of the training size on CMU PIE face database. (b) Recognition rate (%) versus the size of occlusion on the FERET face database.

Moreover, we also computed the training time (in seconds) of all the methods on the databases used in the experiments. All the experiments were run on a workstation (CPU: Intel Xeon, 2.53 GHz; RAM: 8 GB; 64-bit operation system).

C. Experimental Result and Analysis

From the experiments, we can obtain some interesting observations and the conclusions from Tables II–VII.

TABLE II
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION AND TRAINING TIME] OF DIFFERENT ALGORITHMS ON CMU PIE DATASET

L	PCA	LDA	LPP	RR	RFS	FSSL	SAIR	DENLR	RDR
5	54.83±5.32 (55) (0.1384s)	65.51±5.46 (40) (0.1536s)	65.20±5.29 (35) (0.1647s)	64.21±4.03 (68) (0.1737s)	72.77±5.41 (68) (0.4467s)	73.55±6.51 (68) (0.1996s)	64.34±7.71 (66) (0.7407s)	70.70±9.54 (66) (1.0845s)	76.06±4.46 (40)(0.9608s)
4	45.92±6.54 (45) (0.0801s)	56.65±7.54 (44) (0.0943s)	56.33±7.74 (50) (0.0964s)	56.41±7.21 (68) (0.0865s)	63.17±4.69 (68) (0.3369s)	62.68±5.34 (68) (0.1136s)	54.26±5.11 (66) (0.8207s)	63.35±9.23 (66) (0.4605s)	66.60±6.54 (45)(0.6800s)
3	34.93±3.95 (35)(0.0407s)	46.46±7.32 (60)(0.0587s)	46.38±7.50 (24)(0.0.0561s)	46.00±8.79 (68)(0.0583s)	47.89±9.39 (68) (0.2569s)	50.14±6.53 (68)(0.0807s)	46.21±7.87 (66)(0.4349s)	48.35±11.78 (66) (0.4151s)	49.64±6.54 (45)(0.4737s)

TABLE III
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION AND TRAINING TIME] OF DIFFERENT ALGORITHMS ON FERET DATASET

size	PCA	LDA	LPP	RR	RFS	FSSL	SAIR	DENLR	RDR
5*5	71.82±3.23 (90)	72.00±4.21 (50)	72.78±4.24 (40)	68.47±4.41 (200)	80.16±3.82 (145)	80.08±4.14 (195)	75.52±3.92 (195)	77.55±7.71 (200)	83.62±4.11 (50)
10*10	60.97±9.49 (165)	63.77±6. 51 (30)	63.07±6.50 (25)	60.26±5.99 (195)	73.36±6.08 (180)	71.03±6.06 (200)	61.61±5.25 (195)	71.59±5.31 (200)	77.19±5.12 (45)
15*15	20.87±4.71 (90)	54.70±10.57 (20)	54.70±11.07 (20)	55.40±8.61 (175)	65.85±9.66 (200)	63.70±10.43 (200)	55.42±9.29 (190)	64.20±7.96 (200)	70.67±9.63 (30)
Time	(2.2032s)	(2.2572s)	(2.6406s)	(2.2190s)	(3.2482)	(2.4984s)	(2.9376s)	(6.0986s)	(8.5680s)

TABLE IV
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION AND TRAINING TIME] OF DIFFERENT ALGORITHMS ON AR DATASET

PCA	LDA	LPP	RR	RFS	FSSL	SAIR	DENLR	RDR
81.84±6.68 (160) (0.5887s)	87.89±5.46 (119) (0.6293s)	89.90±5.73 (120) (0.7470s)	88.97±5.97 (120) (0.6018s)	91.76±4.90 (120) (1.0912s)	93.22±5.40 (125) (0.6924s)	92.08±5.22 (120) (1.1851s)	89.48±7.86 (120) (1.8329s)	94.58±3.43 (125) (2.4632s)

TABLE V
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), STANDARD DEVIATION, DIMENSION AND TRAINING TIME] OF DIFFERENT ALGORITHMS ON HYPERSPECTRAL FACE DATABASE

PCA	LDA	LPP	RR	RFS	FSSL	SAIR	DENLR	RDR
55.75±6.37 (100)(0.0703s)	72.95±3.83 (46) (0.1096s)	72.27±3.81 (45) (0.1196s)	71.77±3.88 (47) (0.0754s)	73.75±4.21 (47) (0.2995s)	74.36±3.87 (105) (0.0885s)	69.91±3.80 (47) (0.7438s)	74.21±5.08 (47)(0.5713)	76.30±3.56 (50)(0.4874s)

TABLE VI
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), TRAINING TIME, DIMENSION] OF DIFFERENT ALGORITHMS ON POLYU FKP DATASET

L	PCA	LDA	LPP	RR	RFS	FSSL	SAIR	DENLR	RDR
2	51.75(0.1522s) (111)	43.57(0.1892s) (125)	42.24(0.2050s) (200)	47.15(0.1577s) (114)	52.55(0.3552s) (102)	60.66(0.2751s) (47)	55.94(0.4056s) (114)	60.74(1.2860s) (120)	61.15(0.6915s) (69)
4	82.96(0.6419s) (108)	84.72(0.7417s) (175)	83.39(0.8417s) (200)	76.12(0.7451s) (120)	89.33(0.9910s) (102)	91.51(1.0265s) (90)	83.64(0.8756s) (117)	88.25 (1.944s) (135)	93.03(2.7267s) (42)

TABLE VII
COMPARISON OF THE PERFORMANCE [RECOGNITION ACCURACY (%), TRAINING TIME, DIMENSION] OF DIFFERENT ALGORITHMS ON BINARY ALPHA DIGITS IMAGE DATASET

L	PCA	LDA	LPP	RR	RFS	FSSL	SAIR	DENLR	RDR
20	71.92(0.0779s) (27)	71.63(0.0811s) (29)	65.05(0.1766s) (35)	71.34(0.0794s) (35)	72.07(01421s) (35)	72.22(0.2210s) (34)	71.78(0.1499s) (33)	72.04(0.2203s) (35)	76.04(1.6654s) (28)

1) When there are strong noises, for the regression methods, $L_{2,1}$ norm-based RFS usually performs better than the L_2 norm-based RR, which can be found on CMU PIE, FERET, and AR databases. RFS and FSSL perform better than the L_2 norm-based LDA and LPP. Among these methods, RDR achieves the best performance. This indicates that $L_{2,1}$ norm-based methods is more robust than the L_2 norm-based methods when the image contains salt and pepper noises.
2) When there is increasing number of occlusion size on the image, the recognition rates of all methods are decreased. However, the $L_{2,1}$ norm-based methods, i.e., RDR, FSSL, and RFS, also perform better than the L_2 norm-based methods. In this case, compared with

TABLE VIII
COMPARISON OF THE PERFORMANCE (IN AVERAGE RANK) OF DIFFERENT ALGORITHMS ON DIFFERENT DATABASES

Dataset	PCA	LDA	LPP	RR	RFS	FSSL	SAIR	DENLR	RDR
CMU PIE	9.00	6.50	5.20	5.70	4.50	4.10	4.50	3.80	1.70
FERET	8.26	5.63	5.60	6.57	3.83	4.10	5.56	4.00	1.43
AR	7.70	6.90	6.30	5.80	5.10	3.60	4.30	3.80	1.50
Hyperspectral face	8.50	5.30	4.90	5.30	4.80	3.50	6.70	4.40	1.60
FKP	6.05	6.10	6.20	6.35	5.50	4.70	5.20	3.50	1.40
Binary Alpha Digits	5.10	6.50	6.90	6.50	4.70	4.60	5.10	4.30	1.30
Total average rank	7.4350	6.1550	5.850	6.0367	4.7383	4.1000	5.2267	3.9667	1.4883

TABLE IX
STATISTICAL SIGNIFICANCE TEST OF RDR AGAINST OTHER METHODS

i	Method	z	p	α / i	Statistical significance
1	PCA	3.7601	0.0002	0.0063	Yes
2	LDA	2.9515	0.0032	0.0071	Yes
3	RR	2.8766	0.0042	0.0083	Yes
4	LPP	2.7586	0.0058	0.0100	Yes
5	SAIR	2.3643	0.0184	0.0125	No
6	RFS	2.0555	0.0404	0.0167	No
7	FSSL	1.6518	0.0990	0.0250	No
8	DENLR	1.5674	0.1164	0.0500	No

other $L_{2,1}$ norm-based methods, SAIR achieves lower recognition rates. The reason maybe that directly using the label indicator matrix for regression without sparse regularization term or the local geometric structure in regression is not robust to the occlusions or the variations of hyperspectral images. The other reason may be that the SAIR can only obtain C projection for feature extraction, which is not enough for achieving high recognition rate in recognition tasks.

3) FSSL can also obtain higher recognition rates on different cases in the experiments due to the fact that it integrates the local geometric structure and sparse feature selection together. However, RDR still performs better than FSSL. The possible reason is that the L_2 norm is still a main distance metric in FSSL's regression steps, which are sensitive to the block occlusions or image's significant variations. For FSSL and RFS, the feature selection ($L_{2,1}$ norm regularization term for sparse feature selection) plays an important role in obtaining the high recognition rates.
4) Despite of comparing with the classical methods (i.e., PCA, LDA, LPP, and RR) and the most related methods (i.e., RFS, FSSL, and SAIR), the proposed method is also compared with the most recently proposed methods DENLR [7]. We find that RDR performs better than the recently proposed DENLR in most cases. This indicates that using the L_2 norm in the loss function with nuclear norm regularization cannot obtain the better performances than RDR, in which the $L_{2,1}$ norm is used as the measurement for the loss function.

D. Statistical Significance Test

After all experimental results are obtained, we also compute the statistical significance of the proposed RDR against the compared methods using the nonparametric test [46]. We first compute the average rankings of each method on different datasets (shown on Table VIII), and then Friedman tests followed by Holm tests are performed.

From the average rankings we obtain the statistic value $F_F = 3.187$. With eight algorithms and six datasets, F_F is distributed according to the F distribution with $6 - 1 = 5$ and $(6 - 1) \times (9 - 1) = 40$ degrees of freedom. The critical value of $F(5, 40)$ for $\alpha = 0.05$ is 2.45, which is smaller than 3.187. Therefore, we reject the null-hypothesis and the Friedman test shows that the performance differences between these algorithms have statistical significance.

Holm test is used as post-hoc test for the Friedman test. We first compute the value $z = (R_i - R_{\mathrm{RDR}})/\sqrt{N_{\mathrm{method}}(N_{\mathrm{method}} + 1)/6N_{\mathrm{data}}}$, where R_{RDR} and R_i denote the total average ranks of RDR and the ith method listed in Table IX, $N_{\mathrm{method}}(= 9)$ and $N_{\mathrm{data}}(= 6)$ denotes the number of methods and databases. Then zs are used to find the corresponding probability from the table of normal distribution to compute the significant p values of these methods. Holm's step-down procedures are used to evaluate the significance. As can be seem from Table IX, there is significance for RDR against PCA, RR, LDA, and LPP since 0.0058 is smaller than 0.0100. However, the significance for RDR against RFS, FSSL, SAIR, and DENLR is not detected since 0.0184 is larger than 0.0125.

E. Parameters Sensitivity Study

In the proposed RDR method, there are three key parameters, namely the parameter d, i.e., the number of dimension, the neighborhood size K, and the α. From Fig. 4(a) and (b), one can find that parameter K influences the effectiveness of the algorithm on different databases to certain extent. One common phenomenon we find in the experiments is that when $K = 2$ the algorithm performs the best in most cases in different databases. We show the cases on FERET and PIE CMU face databases as two examples in Fig. 4(a) and (b), respectively.

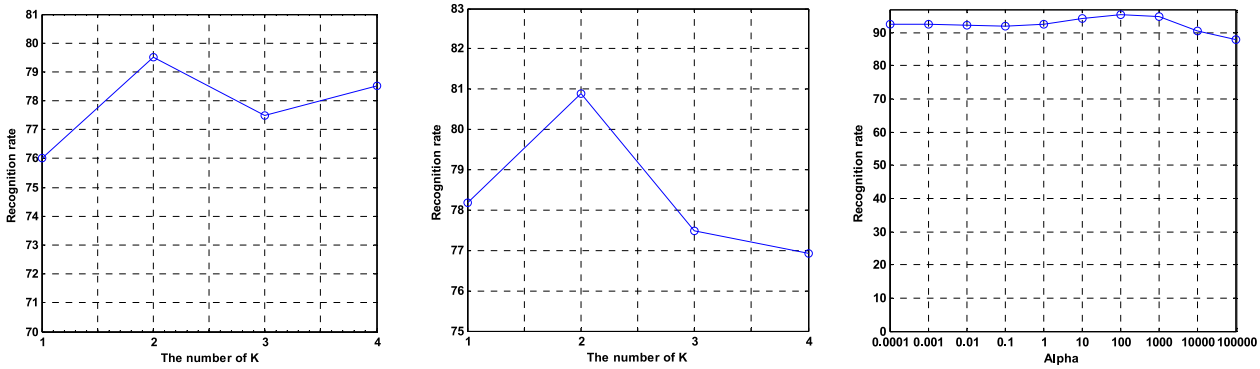


Fig. 4. Recognition rates versus the neighborhood size on (a) FERET and (b) CMU PIE face databases and (c) regularized parameter of RDR.

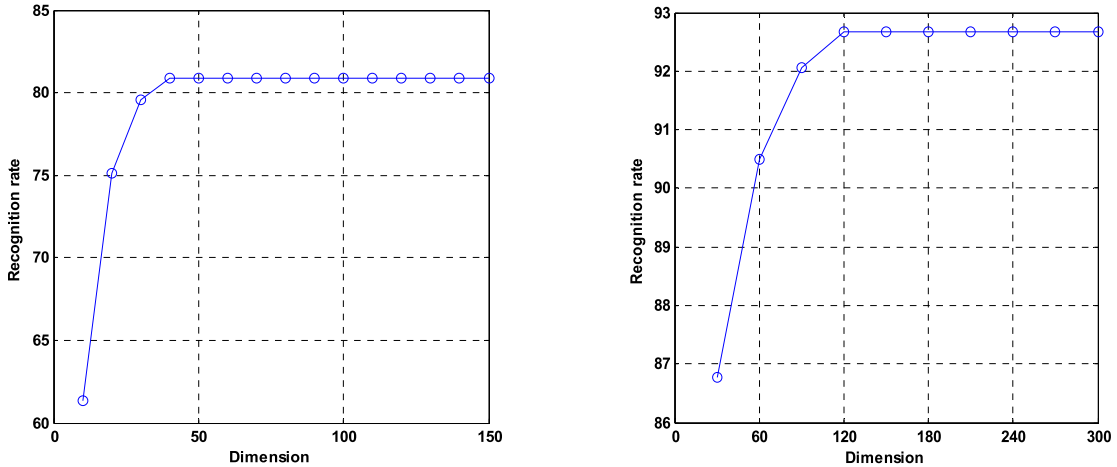


Fig. 5. Recognition rate versus the dimension of the initialized subspace of RDR on (a) CMU PIE and (b) AR face databases.

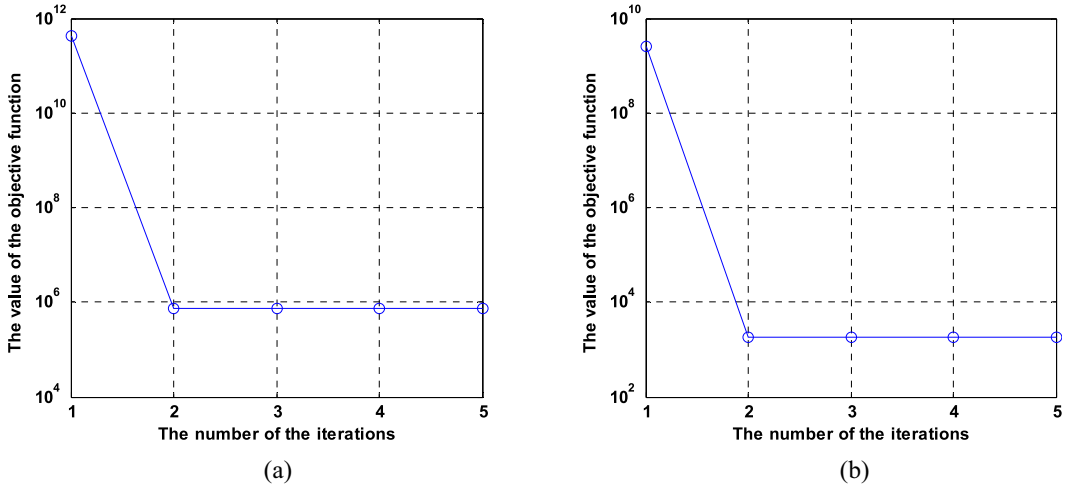


Fig. 6. Convergence on (a) FERET and (b) AR face databases.

The second parameter is the dimension d in RDR. This paper indicates that when d is close to C (the number of class) or slightly larger than C, RDR always achieves its best performance. Furthermore, when a larger $d(d \gg C)$ used in RDR, there is no significant effect on the recognition rate. Fig. 5 shows the recognition rate versus the variations of the dimension on CMU PIE and AR face databases. It can be found from Fig. 5(a) that when the initialized subspace dimension d is 40 (or larger than 40), RDR achieves its best performance. Using a larger $d(d \gg C)$ in RDR cannot obtain higher recognition rate. Similar case can also be found on AR face database. Usually, in order to guarantee that RDR can

achieve its best performance, the dimension of the initialized subspace dimension should be slightly larger than the number of class (this case happens on PolyU hyperspecial face database).

The third one is the regularization parameter α. Usually, the recognition rate of RDR is very robust to α in a big range. When $\alpha \in [10, 1000]$, RDR can achieve its best performance. In Fig. 4(c), we show an example of the recognition rate versus the value α on AR face database. Usually, the proposed method will not achieve its best performance when $\alpha = 0$, which indicates that the regularized term of RDR can improve the discriminative ability and model generalization in feature extraction. Similar property can also be found on other databases.

F. Convergence Study

In Section IV-A, it is proven that the objective function of RDR will converge to the local optimum. In fact, we find that the proposed RDR converges very fast. Generally, the proposed RDR can converge within 3–10 iterations. Fig. 6 shows the convergence curves of RDR with respect to the value of the objective function. Fig. 6(a) and (b) shows the convergent properties of the RDR on FERET and AR face databases, respectively.

VI. Conclusion

In this paper, we present a generalized regression model and propose a novel linear dimensionality reduction method called RDR. RDR uses the robust $L_{2,1}$ norm as the basic metric in the objective function. Since the optimization problem of RDR has no closed form solution, we design an iterative algorithm to compute the optimal solution. It is shown that the optimal projection matrix can be obtained by solving a series of eigenfunctions in the iteration. Comprehensive analyses, including the convergence, computation complexity, and the similarity and difference between the proposed RDR and the classic methods, are presented. The robustness of the RDR is tested on some well-known image databases, in which there are different noises or occlusions. Compared with the $L_{2,1}$ norm-based FSSL methods, the potential disadvantage of the proposed RDR is that it cannot obtain the sparse projections for feature selection. One tractable method that might further increase the recognition rates of RDR is to introduce the sparsity in the projections, which will be explored in the future since directly solving the eigenfunction cannot obtain the sparse solutions (or sparse eigenvectors). Thus, we will continue to develop the robust sparse subspace learning methods based on the $L_{2,1}$ norm as the basic metric instead of only using it as the regularization term as in previous works.

References

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

[2] Y. Su, X. Gao, X. Li, and D. Tao, "Multivariate multilinear regression," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1560–1573, Dec. 2012.

[3] T. Strutz, *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*. Wiesbaden, Germany: Springer Vieweg and Teubner, 2010.

[4] Y. Li and A. Ngom, "Nonnegative least-squares methods for the classification of high-dimensional biological data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 2, pp. 447–456, Mar./Apr. 2013.

[5] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, "Object tracking via partial least squares analysis," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4454–4465, Oct. 2012.

[6] 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.

[7] Z. Zhang *et al.*, "Discriminative elastic-net regularized linear regression," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1466–1481, Mar. 2017.

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

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

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

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

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

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

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

[15] J. Ye, "Least squares linear discriminant analysis," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, Corvallis, OR, USA, 2007, pp. 1087–1093.

[16] 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, Jul. 1997.

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

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

[19] F. De la Torre, "A least-squares framework for component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1041–1055, Jun. 2012.

[20] F. Zheng, Y. Tang, and L. Shao, "Hetero-manifold regularisation for cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2016.2645565.

[21] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $L_{2,1}$-norms minimization," in *Proc. 23rd Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 1813–1821.

[22] Z. Ma *et al.*, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.

[23] Z. Ma, Y. Yang, N. Sebe, K. Zheng, and A. G. Hauptmann, "Multimedia event detection using a classifier-specific intermediate representation," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1628–1637, Nov. 2013.

[24] C.-X. Ren, D.-Q. Dai, and H. Yan, "Robust classification using $L_{2,1}$-norm based regression model," *Pattern Recognit.*, vol. 45, no. 7, pp. 2708–2718, Jul. 2012.

[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] Z. Lai, Y. Xu, J. Yang, L. Shen, and D. Zhang, "Rotational invariant dimensionality reduction algorithms," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2578642.

[27] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. 22nd Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 55. 2011, pp. 1294–1299.

[28] S. Ji, L. Tang, S. Yu, and J. Ye, "A shared-subspace learning framework for multi-label classification," *ACM Trans. Knowl. Disc. Data*, vol. 4, no. 2, pp. 1–29, May 2010.

[29] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.

[30] 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.
[31] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.
[32] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 12, pp. 2268–2269, 2000.
[33] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[34] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 585–591.
[35] S. Zhang, Z. Ma, and H. Tan, "On the equivalence of HLLE and LTSA," in *Advances in Neural Information Processing Systems*, vol. 14. Cambridge, MA, USA: MIT Press, 2002, pp. 585–591.
[36] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.
[37] 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.
[38] 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.
[39] W. K. Wong and H. T. Zhao, "Supervised optimal locality preserving projection," *Pattern Recognit.*, vol. 45, no. 1, pp. 186–197, Jan. 2012.
[40] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Beijing, China, 2005, pp. 1208–1213.
[41] M. Yu, L. Shao, X. Zhen, and X. He, "Local feature discriminant projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1908–1914, Sep. 2016.
[42] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA, USA: Academic Press, 1990.
[43] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[44] A. A. Martinez and R. Benavente, "The AR face database," CVC, Barcelona, Spain, Tech. Rep. 24, 1998.
[45] W. Di, L. Zhang, D. Zhang, and Q. Pan, "Studies on hyperspectral face recognition in visible spectrum with feature band selection," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 6, pp. 1354–1361, Nov. 2010.
[46] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

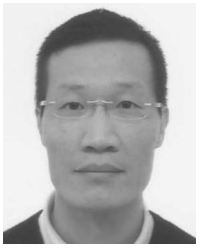
Zhihui Lai received the B.S. degree in mathematics from South China Normal University, Guangzhou, China, in 2002, the M.S. degree from Jinan University, Guangzhou, in 2007, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing,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

He has been a Research Associate, a Post-Doctoral Fellow, and a Research Fellow with Hong Kong Polytechnic University, Hong Kong. He has published over 60 scientific articles, including 30 papers published on top-tier IEEE TRANSACTIONS. His current research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization, and applications in the field of intelligent robot research.

Dr. Lai is an Associate Editor of the *International Journal of Machine Learning and Cybernetics*.

Dongmei Mo received the B.S. degree from Zhao Qing University, Zhaoqing, China. She is currently pursuing the M.S. degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His current research interests include artificial intelligence and pattern recognition.



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

He is currently a Professor with Hong Kong Polytechnic University. He has published over 50 scientific articles in refereed journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, the *International Journal of Production Economics*, the *European Journal of Operational Research*, the *International Journal of Production Research*, *Computers in Industry*, and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS. His current research interests include artificial intelligence, pattern recognition, and optimization of manufacturing scheduling, planning, and control.

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

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

Duoqian Miao received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1997.

He is a Professor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China, where he is also with the Department of Computer Science and Technology, the Director of the National Experimental Teaching Demonstration Center of Computer and Information Technology, and the Vice Director of the Key Laboratory of Embedded System and Service Computing Ministry of Education. His current research interests include rough sets, granular computing, principal curve, Web intelligence, and data mining.

Dr. Miao is a fellow of the International Rough Set Society.

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

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

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