IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 25, 2018, accepted November 19, 2018, date of publication December 5, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2885131

Robust Locally Discriminant Analysis via Capped Norm

ZHIHUI LAI[1,2], NING LIU[1], LINLIN SHEN[1], AND HENG KONG[3]

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[2]Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong
[3]The Shenzhen University General Hospital, Shenzhen University, Shenzhen 518060, China

Corresponding author: Zhihui Lai (lai_zhi_hui@163.com)

This work was supported in part by the Natural Science Foundation of China under Grant 61573248, Grant 61802267, Grant 61773328, Grant 61732011, and Grant 61703283, 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 and Grant JCYJ20160429182058044.

ABSTRACT Conventional linear discriminant analysis and its extended versions have some potential drawbacks. First, they are sensitive to outliers, noise, and variations in data, which degrades their performances in dimensionality reduction. Second, most of the linear discriminant analysis-based methods only focus on the global structures of data but ignore their local geometric structures, which play important roles in dimensionality reduction. More importantly, the total number of projections obtained by linear discriminant analysis (LDA) based methods are limited by the class number in the training data set. To solve the problems mentioned above, we propose a novel method called robust locally discriminant analysis via capped norm (RLDA), in this paper. By replacing L_2 -norm with $L_{2,1}$- norm to construct the robust between-class scatter matrix and using the capped norm to further reduce the negative impact of outliers in constructing the within-class scatter matrix, we can guarantee the robustness of the proposed methods. In addition, we also impose $L_{2,1}$-norm regularized term on projection matrix, so that its joint sparsity can be ensured. Since we redefine the scatter matrices in traditional LDA, the projection numbers we obtain are no longer restricted by the class numbers. The experimental results show the superior performance of RLDA to other compared dimensionality reduction methods.

INDEX TERMS Feature extraction, capped L_2-norm loss, $L_{2,1}$-regularization, manifold learning, discriminant analysis.

I. INTRODUCTION

As it is known to all, the curse-of-dimensionality is a difficult but essential problem in computer vision and pattern recognition. The reason for this phenomenon is that high-dimensional data is widespread and may lead to high computational complexity. Therefore, dimensionality reduction methods are frequently needed and widely utilized in practical application fields such as rough-fuzzy clustering [1], feature extraction [2], image recovery [3], feature selection [4]–[6], image preprocessing [7] and subspace clustering [8].

Generally, the dimensionality reduction methods can be roughly divided into three categories: (1) L_2-norm based methods subspace learning methods, (2) L_1-norm based methods, and (3) other robust jointly sparse feature selection methods. In what follows, we will review these three kinds of methods related to this paper.

As for the first category, the most popular methods include Principal Component Analysis (PCA) [9], Linear Discriminant Analysis (LDA) [10] and Locality Preserving Projections (LPP) [11]. Researchers have also tried to integrate the property of locality preserving and discriminant analysis theory together, and many of their studies achieved promising results. For example, Masashi combined the property of LDA and LPP and proposed local Fisher discriminant analysis (i.e. LFDA [12]). Other discriminant locality based methods include locality adaptive discriminant analysis (i.e. LADA [13]), local similarity and diversity preserving discriminant projection (i.e. LSDDP [14]), discriminant similarity and variance preserving projection (i.e. DSVPP [15]), local structure preserving discriminant analysis (i.e. LSPDA [16]), local maximal margin discriminant embedding (i.e. LMMDE [17]), fuzzy local discriminant embedding (i.e. FLDE [18]) and so on. However, each of these methods uses L_2 norm as the basic metric, and its square operation will magnify the negative effect of outliers. This leads to their lack of robustness, especially when training data is corrupted with noise.

In order to address this problem, researchers proposed L_1-norm based methods so that we can improve the robustness of the learned subspace. Inspired by this idea, L_1-norm based methods are widely used in feature extraction [19], coupled dictionary learning [20], adaptive sparse regression [21] and sparse subspace learning [22]. In recent years, L_1-norm based methods have been further developed. Zhang *et al.* [23] proposed L1-norm-based local optimal locality preserving LDA (i.e. LLDA_L1) and L1-norm-based global optimal locality preserving LDA (i.e. GLDA_L1). Furthermore, Yuan *et al.* [24] proposed a robust version of EPP based on L1-norm maximization (i.e. EPP-L1).

While L_1-norm based methods are superior to most of the L_2-norm based methods, they also have major problems as follows: (1) In most of the methods, L_1-norms are only employed as regularization terms, but L_2-norms are still dominant in loss function, so that these methods are still sensitive to outliers in a certain case. (2) Although we can obtain sparse projections through L_1-norm based methods, we cannot ensure their joint sparsity. To solve these problems, $L_{2,1}$-norm based methods were proposed and widely used in practical applications, among which jointly sparse feature extraction techniques are the most popular methods we would like to introduce in the next paragraph.

For the latter category of dimensionality reduction methods, (i.e. robust jointly sparse feature selection methods), researchers focus more on the $L_{2,1}$-norm based methods due to its simplicity and effectiveness. At first, Nie at el. proposed RFS [25], in which they introduce $L_{2,1}$-norm for feature selection. By employing $L_{2,1}$-norm as the basic metric and also imposing it on projection matrix as regularization term, we can enhance their robustness and ensure their joint sparsity. Following these idea, many feature selection methods using $L_{2,1}$-norm were also developed, including jointly sparse representation [26], multi-kernel learning [27], [28], sparse neighborhood preserving embedding [29] and so on.

In addition to the three categories mentioned above, capped norms have also been extended to robust principal component analysis problem since it is more robust than the former three categories. For example, Ma *et al.* [30] proposed a novel local linear regression for SISR based capped $L_{2,1}$-norm function recently. It was proposed to solve the problem in single image super resolution tasks, and studies showed that it achieves better reconstruction performance against other state-of-the-art methods. Moreover, Sun *et al.* [31] presented a novel nonconvex formulation for the RPCA problem using the capped trace norm and the capped L_1-norm. While their study was based on the assumption that no elements are missing in the sample data, this assumption was difficult to realize in real-word application. Therefore, matrix completion remains a major task in dimensionality reduction. Inspired by their study, Zhang at el. [32] proposed to employ the capped nuclear norm in matrix completion since it can reflect the rank more directly and accurately than the nuclear norm.

In addition, capped norm has also been utilized in clustering [33], semi-supervised learning [34] and feature selection [35] and anomaly detection [36]. These studies also show that methods integrating capped norms outperform many state-of-the-art methods.

However, as a new research field, capped norm based methods have potential drawbacks: (1) They ignore the relationship between within-class scatter matrix and between-class scatter matrix which is essential in preserving the global discriminant structure. As a result, we are not able to lessen the distance among the samples in the same class and enlarge the distances among the samples from different classes. (2) Some of the capped–norm based methods mentioned above do not consider local geometric structures, which will remarkably upgrade its performance in dimensionality reduction.

To address the first problem mentioned above, a feasible approach is to introduce fisher criterion into capped norm based methods, so that they can preserve the global discriminant structure to learn discriminative subspace. The second problem mentioned above can be alleviated by integrating the advantages of locality based methods since previous studies [37], [38], [23] have shown that preserving locality can effectively upgrade the performance in dimensionality reduction.

Robustness is a major measurement in the performance of dimensionality reduction. In this paper, we integrate the property of capped norm into linear discriminant methods.

The main contributions of this paper are as follows:

(1) We propose a novel learning method called Robust locally discriminant analysis (RLDA) via capped norm. In the proposed method, we introduce fisher criterion to preserve the global discriminant structure, and redefine the within-class and between-class scatter matrices in original LDA. That is, more robust $L_{2,1}$–norm based metric and capped norms are utilized to redefine the scatter matrices, so that we improve the robustness of our method. In addition, $L_{2,1}$–norm regularization was also introduced to make sure that the projections are jointly sparse.

(2) Different from LDA based methods, the proposed methods can obtain more projections than $C - 1$, where C is the total class number in training samples. That is, the dimensions of projections are not limited by the class number any more, and this is beneficial to feature extraction.

(3) We propose an iterative algorithm to solve the optimization problem. In addition, the convergence analysis and computational complexity analysis are also presented. Besides, extensive experiments are also conducted to evaluate the performance of the proposed method, and experimental results show that the proposed method is superior to the compared methods.

The rest of the paper is organized as follows. In section II, we define some notations and review some related works. In Section III, we present the formulation of RLDA and propose an iterative algorithm to solve the optimization problem. In Section IV, the convergence analysis and computational complexity of the proposed methods are presented. In Section V, a set of experiments are carried out to evaluate the performance in dimensionality reduction of the proposed RLDA algorithm and some other compared methods. Finally, we draw a conclusion in Section VI.

II. RELATED WORKS

In this section, we first present some notations used in this paper, and then review LDA, LPP and capped–norm.

A. NOTATIONS AND DEFINITIONS

In this paper, matrices and vectors are represented as uppercase letters and lowercase letters, respectively. Besides, let a^i and a_j be the i-th row and the j-th column of matrix A, separately.

We use $X \in R^{d\times n}$ to denote data matrix, where d and n represent the dimension of features and the total number of training samples, respectively.

The $L_{2,1}$-norm of a matrix $A \in R^{m\times n}$ is defined as:

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

B. LDA

In traditional LDA [10], the within-class scatter matrix and between-class scatter are defined as (2) and (3) respectively:

$$S_w = \frac{1}{N}\sum_{i=1}^{C}\sum_{j\in C_i}(x_j - \bar{x}_i)(x_j - \bar{x}_i)^T. \tag{2}$$

$$S_B = \frac{1}{N}\sum_{i=1}^{C} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T. \tag{3}$$

Fisher criterion in traditional linear discriminant analysis theory aims to minimize the within-class scatter matrix and meanwhile maximize the between-class scatter matrix, which can be written as (4) or (5) separately:

$$\min_B \frac{tr(B^T S_W B)}{tr(B^T S_B B)}, \tag{4}$$

$$\max_B \frac{tr(B^T S_B B)}{tr(B^T S_W B)}, \tag{5}$$

where B is the projection matrix.

The optimization problem in LDA can be converted into an eigen-decomposition problem. The optimal projections are made up of the eigenvectors of $S_W^{-1}S_B$ corresponding to the largest eigenvalues. In reality, S_W is sometimes singular due to small sample size problem, so we often employ PCA as a preprocessing step to reduce the dimensionality of training samples beforehand.

C. LPP

LPP [11] was proposed to preserve the local geometric information of the training data. Assuming that we use $P \in R^{d\times k}$ to represent the projection matrix we aim to learn, and then the objective function of LPP can be presented as (6):

$$\min_P \sum_{ij} \left\|P^T x_i - P^T x_j\right\|_2^2 W_{ij}, \tag{6}$$

where W is the local neighborhood graph and its definition are presented in (7):

$$W_{ij} = \begin{cases} 1, & if\ x_i \in C_k(x_j) or x_j \in C_k(x_i) \\ 0, & otherwise, \end{cases} \tag{7}$$

where $C_k(x_i)$ is a set consisting of the k nearest neighbors of x_i.

Based on the definitions above, (6) is equivalent to (8)

$$\min_P tr[P^T X(Q - W)X^T P], \tag{8}$$

where Q is a diagonal matrix and its elements are row (or column sum) of matrix W, namely, $Q_{ii} = \sum_{i=1}^{n} W_{ij}$.

The objective function of LPP can be formulated as follows:

$$\begin{aligned} &\min_P tr[(P^T XQX^T P)^{-1} P^T XRX^T P], \\ &s.t.\ P^T XQX^T P = I_d \end{aligned} \tag{9}$$

where $I_d \epsilon R^{d\times d}$ is an identity matrix, and $R = Q - W$. The minimization problem can be further converted to eigenvalue problem as follows:

$$X^T RX^T P = XQX^T P\Lambda, \tag{10}$$

where Λ is the eigenvalue matrix of matrix P. Eventually, the optimal matrix consists of k minimum eigenvalue solutions of the above generalized eigenvalue problem.

D. CAPPED NORM MINIMIZATION

As for the capped-L_1 norm based minimization problem in [39], [41], researchers hope to learn a weight matrix $K \in R^{d\times n}$ consisting of the weight vectors under n predictive models: $y_i \approx f_i(\mathrm{X}_i) = X_i \times k_i$. L_1 penalty was employed on each row of K to obtain a column vector. Later on, they impose the capped-L_1 penalty afterwards on the column vector. The capped norm minimization problem can be formulated as follows:

$$\min \left\{ l(K) + \lambda \sum_{j=1}^{d} \min(\left\|k^j\right\|_1, \varepsilon) \right\}, \tag{11}$$

where $l(K)$ denotes an empirical loss function of K, and $\lambda\ (> 0)$ represents a parameter balancing the empirical loss and the regularization term. $\varepsilon(> 0)$ denotes a thresholding parameter and k^j is the j-th row of the matrix K.

III. ROBUST LOCALLY SPARSE REGRESSION

In this section, we present the objective function of the proposed method, and then propose an iterative algorithm to solve the optimization problem.

A. MOTIVATION AND FORMULATION OF RLDA

Let $B \in X^{d\times k}$ be the projection matrix, where k is the dimension of the low-dimensional subspace we aim to learn. It is clear that we will map the original data points x_i into low-dimensional subspace y_i according to (12).

$$y_i = x^T B. \tag{12}$$

Within-class scatter value in conventional LDA is sensitive to outliers since it is defined using L_2-norm as the basic metric. To address this problem, in this paper, we redefine it by using the capped norm as the basic metric to further alleviate the negative effect of outliers, which can be presented as (13):

$$\sum_{i=1}^{c}\sum_{j=1}^{N_i} \min(\left\|(x_i^j - \bar{x}_i)^T B\right\|_2, \varepsilon), \tag{13}$$

where x_i^j represents the j-th sample in the i-th class.

However, the above definition only characterizes the within-class information and neglect the local geometric structure. As it is shown in the existing research, the locality is very important to improve the algorithm's performance and it is necessary to take it into consideration. Therefore, the locality in LPP is integrated in (13). Besides, we also expect the algorithm can select the key features to further improve the performance. As a result, the joint sparsity of the learned projections should be guaranteed and the $L_{2,1}$-norm is imposed on our projection matrix in the proposed model as a regularized term. Then we obtain:

$$\begin{aligned}\min_{B^TB=I} \sum_{i=1}^{c}\sum_{j=1}^{N_i} \min(\left\|(x_i^j - \bar{x}_i)^T B\right\|_2, \varepsilon) + \alpha \left\|B\right\|_{2,1} \\ +\beta \sum\nolimits_{ij} \left\|B^T x_i - B^T x_j\right\|_2^2 W_{ij},\end{aligned} \tag{14}$$

where $\alpha > 0$ is the regularization parameter of the $L_{2,1}$-norm and W is the k-nearest neighbor graph defined in (7).

Minimizing the first part of (14) indicates that, if x_i^j is a sample point in the i-th class, and $\bar{x}_i$ is the mean of i-th class, the low-dimensional representation of x_i^j and $\bar{x}_i$ should be close. Namely, $x_i^j B$ and $\bar{x}_i B$ should be close. However, if there are outliers in training samples, they will magnify the penalty in the loss function. Therefore, the capped norm is introduced to reduce the negative impact of outliers.

Similar to within-class scatter value, we also utilize $L_{2,1}$-norm based metric to redefine between-class scatter value as follows:

$$\left\|\begin{array}{c} N_1(\bar{x}_1^T - \bar{x}^T)B \\ \cdots \\ N_c(\bar{x}_c^T - \bar{x}^T)B \end{array}\right\|_{2,1} = \sum_{i=1}^{c} N_i \left\|(\bar{x}_i^T - \bar{x}^T)B\right\|_2, \tag{15}$$

where $\bar{x}_i$ is the mean of i-th class.

Inspired by the Fisher criterion, we finally obtain the objective function of RLDA as follows:

$$\begin{aligned}\min_{B^TB=I} \sum_{i=1}^{c}\sum_{j=1}^{N_i} \min(\left\|(x_i^j - \bar{x}_i)^T B\right\|_2, \varepsilon) \\ + \alpha \left\|B\right\|_{2,1} + \beta \sum\nolimits_{ij} \left\|B^T x_i - B^T x_j\right\|_2^2 W_{ij}. \\ s.t. \sum_{i=1}^{c} N_i \left\|(\bar{x}_i^T - \bar{x}^T)B\right\|_2 = cons\end{aligned} \tag{16}$$

The difference between the previously proposed SCM[41] and our RLDA is that we introduce the property of LPP to guarantee the joint sparsity of the projections, and further improve the performance of LDA based methods by redefining the scatter matrices defined in LDA.

B. THE OPTIMAL SOLUTION TO RLDA

As for the first part of the objective function in (16), we can compute it as it is shown in (17):

$$\begin{aligned}\min_{B^TB=I} \sum_{i=1}^{c}\sum_{j=1}^{N_i} \min(\left\|(x_i^j - \bar{x}_i)^T B\right\|_2, \varepsilon) \\ = \left\|X_{Rw}^T B\right\|_{2,1} \\ = tr(B^T X_{Rw} D_{Rw} X_{Rw}^T B) = tr(B^T S_{Rw} B),\end{aligned} \tag{17}$$

where $S_{Rw} = X_{Rw} D_{Rw} X_{Rw}^T$ is the redefinition of within-class scatter matrix in LDA. The data matrix X_{Rw} and the diagonal matrix D_{Rw} are redefined as (18) and (19), shown at the top of the next page, respectively, where Ind is an indicative function defined as (20):

$$Ind = \begin{cases} 1, & \text{if } \left\|(x_i^j - \bar{x}_i)^T B\right\|_2 \le \varepsilon \\ 0, & otherwise. \end{cases} \tag{20}$$

Moreover, the second part and the third part of (16) can be computed by (21) and (22) respectively:

$$\alpha \left\|B\right\|_{2,1} = \alpha tr(B^T D_b B). \tag{21}$$

$$\begin{aligned}\beta \sum\nolimits_{ij} \left\|B^T x_i - B^T x_j\right\|_2^2 W_{ij} = \beta tr\left[B^T X(D_1 - W)X^T B\right] \\ = \beta tr(B^T X L X^T B),\end{aligned} \tag{22}$$

where D_b is a diagonal matrix with its i-th diagonal element defined in (23), D_1 is a diagonal matrix whose elements are column (or row, since W is symmetric) sum of W, and L is the Laplacian matrix. The definition of D_1 and L are presented as (24) and (25), respectively.

$$D_{bii} = \frac{1}{2\left\|b^i\right\|_2}. \tag{23}$$

$$D_{1_{ii}} = \sum\nolimits_j W_{ij}. \tag{24}$$

$$L = D_1 - W. \tag{25}$$

$$X_{Rw} = ((x_1^1 - \bar{x}_1), \ldots, (x_1^{N_1} - \bar{x}_1), \ldots, (x_c^1 - \bar{x}_c), \ldots, (x_c^{N_c} - \bar{x}_c)). \tag{18}$$

$$D_{Rw} = \begin{bmatrix} \frac{Ind}{2\left\|(x_1^1 - \bar{x}_1)B\right\|_2} & & & & & & \\ & \ldots & & & & & \\ & & \frac{Ind}{2\left\|(x_1^{N_1} - \bar{x}_1)B\right\|_2} & & & & \\ & & & \ldots & & & \\ & & & & \frac{Ind}{2\left\|(x_c^1 - \bar{x}_c)B\right\|_2} & & \\ & & & & & \ldots & \\ & & & & & & \frac{Ind}{2\left\|(x_c^{N_c} - \bar{x}_c)B\right\|_2} \end{bmatrix}, \tag{19}$$

After these transformations, we can obtain:

$$\begin{aligned} \min_{B^T B=I} \; & tr(B^T X_{Rw} D_{Rw} X_{Rw}^T B) + \alpha tr(B^T D_b B) \\ & + \beta tr(B^T XLX^T B). \\ s.t. \; & \sum_{i=1}^{c} N_i \left\|(\bar{x}_i^T - \bar{x}^T)B\right\|_2 = cons \end{aligned} \tag{26}$$

In addition, the constraint term can be formulated as follows:

$$\begin{aligned} \sum_{i=1}^{c} N_i \left\|(\bar{x}_i^T - \bar{x}^T)B\right\|_2 &= tr(B^T X_{Rb} D_{Rb} X_{Rb}^T B) \\ &= tr(B^T S_{Rb} B), \end{aligned} \tag{27}$$

where $S_{Rb} = X_{Rb} D_{Rb} X_{Rb}^T$ is the redefinition of between-class scatter matrix in LDA. Besides, the data matrix X_{Rb} and the diagonal matrix D_{Rb} are redefined as (28) and (29), respectively.

$$X_{Rb} = (N_1(\bar{x}_1^T - \bar{x}^T), \ldots, N_c(\bar{x}_c^T - \bar{x}^T)). \tag{28}$$

$$D_{Rb} = \begin{bmatrix} \frac{1}{2\left\|N_1(\bar{x}_1^T - \bar{x}^T)B\right\|_2} & & \\ & \ldots & \\ & & \frac{1}{2\left\|N_c(\bar{x}_c^T - \bar{x}^T)B\right\|_2} \end{bmatrix}. \tag{29}$$

After that, we are able to obtain:

$$\begin{aligned} & \min_{B^T B=I} \; tr(B^T X_{Rw} D_{Rw} X_{Rw}^T B) + \alpha tr(B^T D_b B) + \beta tr(B^T XLX^T B) \\ & s.t. \; tr(B^T X_{Rb} D_{Rb} X_{Rb}^T B) = cons \end{aligned} \tag{30}$$

From (30), we can know that it is the trace optimization problem similar to classical LDA. Using Lagrangian multiplier methods we can obtain (31).

$$\min_{B^T B=I} \frac{tr[B^T(X_{Rw} D_{Rw} X_{Rw}^T + \alpha D_b + \beta XLX^T)B]}{tr(B^T X_{Rb} D_{Rb} X_{Rb}^T B)} \tag{31}$$

After some simple transformations, we can get the following generalized eigenfunction:

$$(X_{Rb} D_{Rb} X_{Rb}^T)^{-1}(X_{Rw} D_{Rw} X_{Rw}^T + aD_b + \beta XLX^T)B = B\Lambda. \tag{32}$$

where Λ is a diagonal matrix which is made up of eigenvalues of their corresponding eigenvectors, and their corresponding eigenvectors lies in each column of projection matrix B.

Thus, it is clear that the trace minimization problem can be solved by the eigen-decomposition of the matrix $(X_{Rb} D_{Rb} X_{Rb}^T)^{-1}(X_{Rw} D_{Rw} X_{Rw}^T + aD_b + \beta XLX^T)$, and the optimal matrix B is made up of eigenvectors corresponding to its smallest eigenvalue. Since the updating of D_b and D_{Rb} is related to the projection matrix B obtained from the last iteration, we need to develop an alternatively iterative algorithm to compute the optimal projection, which can be obtained by the algorithm procedures shown in TABLE 1.

IV. ALGORITHM ANALYSIS

In this section, we present the convergence analysis and computational complexity of our proposed RLDA algorithm.

A. CONVERGENCE ANALYSIS

To prove the convergence of the proposed algorithm, we need the following lemma.

Lemma 1 [25]: For any nonzero vectors a and b, the following inequality holds:

$$\|a\|_2 - \frac{1}{2}\frac{\|a\|_2^2}{\|b\|_2} \leq \|b\|_2 - \frac{1}{2}\frac{\|b\|_2^2}{\|b\|_2} \tag{33}$$

From Lemma 1, we can easily obtain Corollary 1:

Corollary 1: For any nonzero vectors b_{k+1}^i and b_k^i, inequality in (34) holds:

$$\left\|b_{k+1}^i\right\|_2^1 - \frac{1}{2}\frac{\left\|b_{k+1}^i\right\|_2^2}{\left\|b_k^i\right\|_2^1} \leq \left\|b_k^i\right\|_2^1 - \frac{1}{2}\frac{\left\|b_k^i\right\|_2^2}{\left\|b_k^i\right\|_2^1} \tag{34}$$

Theorem 1: The algorithm shown in TABLE 1 will monotonically decrease the objective function in (16).

TABLE 1. Robust locally discriminant analysis via capped norm.

```
Input: Data matrix $X \in R^{d\times n}$, the symmetric matrix $W \in R^{n\times n}$, the numbers of iterations $ite$, dimensions $k(\le d)$, parameters $a$,
    $\beta$ and $\varepsilon$ .
Output: Low-dimensional features $y_i = B^T x_i$ ($i = 1,2,...,n$).
Step 1: Construct matrix $X_{Rw}$, $X_{Rb}$ using (18), (28) separately.
Step 2: Initialize matrix $B \in R^{d\times k}$ as an arbitrary orthogonal matrix.
Step 3: For $t = 1 : ite$ do
        - Solve the eigen-equation in (31) to update $B$
        - Update the matrix $D_b$ using (23)
        - Update the matrix $D_{Rw}$, $D_{Rb}$ using (19) , (29) respectively.
        - Set t to t+1
        - if objective function converges, end for
Step 4: Output the optimal projection $B$ for feature extraction
Step 5: Project the training samples to low-dimensional subspace $y_i = B^T x_i$ ($i = 1,2,...,n$)
```

Proof: Although D_{Rb} has also been updated during the iterative process, it was only used as a constraint. Therefore, the optimization problem in (16) can be denoted as $J(B, D_b, D_{Rw})$ for simplicity, and the value of $J(B, D_b, D_{Rw})$ is as follows:

$$J(B, D_b, D_{Rw}) = tr(B^T X_{Rw} D_{Rw} X_{Rw}^T B) + \alpha tr(B^T D_b B) + \beta tr(B^T XLX^T B). \tag{35}$$

Suppose we have obtained B_k in the k-th iteration, since B_k is updated by eigen-equation in (32), we can obtain inequality as follows:

$$J(B_{k+1}, D_b^k, D_{R_w}^k) \le J(B_k, D_b^k, D_{R_w}^k) \tag{36}$$

This is equivalent to:

$$\begin{aligned} &tr(B_{k+1}^T X_{Rw} D_{R_w}^k X_{Rw}^T B_{k+1}) + \alpha tr(B_{k+1}^T D_b^k B_{k+1}) \\ &\quad + \beta tr(B_{k+1}^T XLX^T B_{k+1}) \\ &\le tr(B_k^T X_{Rw} D_{R_w}^k X_{Rw}^T B_k) + \alpha tr(B_k^T D_b^k B_k) \\ &\quad + \beta tr(B_k^T XLX^T B_k) \end{aligned} \tag{37}$$

We can easily find that (37) can be rewritten as (38) utilizing the definition of $L_{2,1}$-norm and F-norm

$$\begin{aligned} &tr(B_{k+1}^T X_{Rw} D_{R_w}^k X_{Rw}^T B_{k+1}) \\ &\quad + \alpha tr(\sum_{i=1}^d \frac{1}{2} \frac{\|b_{k+1}^i\|_2^2}{\|b_k^i\|_2}) + \beta tr(B_{k+1}^T XLX^T B_{k+1}) \\ &\le tr(B_k^T X_{Rw} D_{R_w}^k X_{Rw}^T B_k) \\ &\quad + \alpha tr(\sum_{i=1}^d \frac{1}{2} \frac{\|b_k^i\|_2^2}{\|b_k^i\|_2}) + \beta tr(B_k^T XLX^T B_k), \end{aligned} \tag{38}$$

where b_k^i denotes the i-th row of matrix B_k.

From inequality in (34) and (38), we can easily get (39)

$$\begin{aligned} &tr(B_{k+1}^T X_{Rw} D_{R_w}^k X_{Rw}^T B_{k+1}) \\ &\quad + \alpha tr(\sum_{i=1}^d \frac{1}{2} \frac{\|b_{k+1}^i\|_2^2}{\|b_{k+1}^i\|_2}) + \beta tr(B_{k+1}^T XLX^T B_{k+1}) \\ &\le tr(B_k^T X_{Rw} D_{R_w}^k X_{Rw}^T B_k) \\ &\quad + \alpha tr(\sum_{i=1}^d \frac{1}{2} \frac{\|b_k^i\|_2^2}{\|b_k^i\|_2}) + \beta tr(B_k^T XLX^T B_k) \end{aligned} \tag{39}$$

On each side of the inequality, we can reformulate the second part and obtain:

$$\begin{aligned} &tr(B_{k+1}^T X_{Rw} D_{R_w}^k X_{Rw}^T B_{k+1}) \\ &\quad + \alpha tr(B_{k+1}^T D_b^{k+1} B_{k+1}) + \beta tr(B_{k+1}^T XLX^T B_{k+1}) \\ &\le tr(B_k^T X_{Rw} D_{R_w}^k X_{Rw}^T B_k) \\ &\quad + \alpha tr(B_k^T D_b^k B_k) + \beta tr(B_k^T XLX^T B_k) \end{aligned} \tag{40}$$

From (40) we can get (41), which means updating D_b decrease our objective function

$$J(B_{k+1}, D_b^{k+1}, D_{R_w}^k) \le J(B_k, D_b^k, D_{R_w}^k) \tag{41}$$

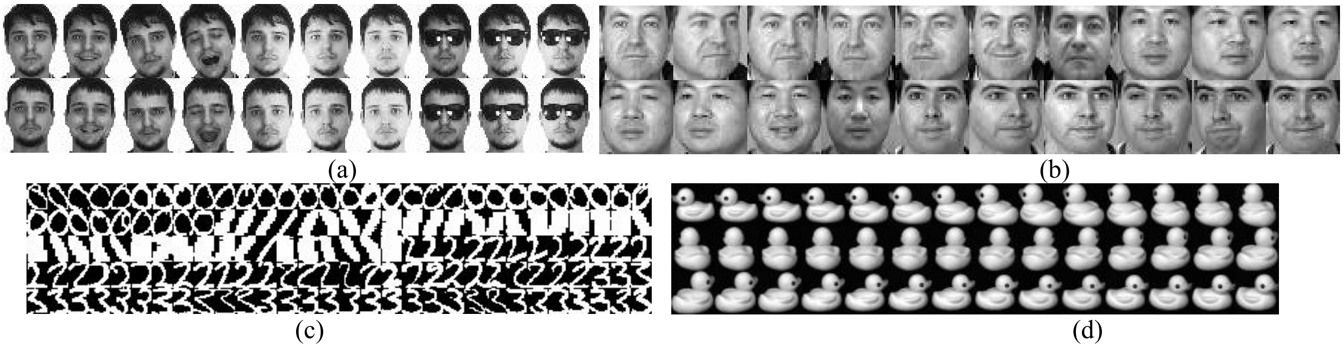
Therefore, the last step is to prove that the updating process of D_{Rw} also decrease the objective function. Namely, we aim to prove (42):

$$J(B_{k+1}, D_b^{k+1}, D_{R_w}^{k+1}) \le J(B_{k+1}, D_b^{k+1}, D_{R_w}^k) \tag{42}$$

Since the second and the third part of (30) are constants with respect to D_{Rw}, we can fix them and ignore their influence during the process of proving (42).

To prove the above inequality, we need following proposition as preparation.

Proposition 1: When B and D_b are fixed as constants, (35) becomes a non-convex capped-norm based optimization problem concerning only one variable D_{R_w}. This problem can be solved by utilizing the concave convex



(a) (b) (c) (d)

FIGURE 1. Image samples collected from (a) AR face database, (b) FERET face database, (c) Binary Alphadigits (BA) dataset, and (d) COIL-20 database.

procedure (CCCP) [40], the property of concave function and sub-gradient, and locally linear approximation simultaneously. Thus the updating process of D_{R_w} will monotonically decrease the objective function in (35).

Proof: The proposition can be derived from previous work of Lan *et al.* [41]. The detailed proof can be referred in [41]. (To be exact, the proof of this proposition can be referred to contents from (28) to (37) in [41]).

The main idea of his proof to the proposition is that, if non-convex objective function has a convex upper bound, we are able to minimize the objective function by minimizing its upper bound, so that (42) can be proved.

Combining (41) and (42), we arrive at the following inequality:

$$J(B_{k+1}, D_b^{k+1}, D_{R_w}^{k+1}) \leq J(B_k, D_b^k, D_{R_w}^k) \tag{43}$$

Therefore, we prove that the objective function of the proposed RLDA monotonically decrease and it will converge to a local optimal solution.

B. COMPUTATIONAL COMPLEXITY

In this subsection, we use d, n to denote the dimension of original training data, the number of training samples, respectively. Besides, let c and k represent the number of classes in the training data and the number of neighbors separately. Updating D_{Rb} and D_{Rw} costs $O(ckd)$ and $O(nkd)$ respectively. It takes us $O(nkd + nk^2c + nc)$ to update D_b. The main computational complexity comes from solving the eigen-equation in (32), so updating B will cost $O(d^3)$. In reality, the dimensions of original data are always bigger than other constants, so we only take computational complexity caused by eigen-equation into consideration. Therefore, if the iteration steps is T , the total computational complexity is $O(Td^3)$.

V. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of our proposed RLDA on four well-known datasets, namely, AR Face dataset, FERET dataset, Binary alpha digits (BA) dataset and COIL-20 database. Besides, a number of subspace learning methods are selected for comparison. These methods include conventional linear discriminant methods Linear Discriminant Analysis (i.e. LDA) [10], and its extended version low-rank linear regression method (i.e. LRLR) [42]. In addition, locality based methods such as conventional Locality Preserving Projections (i.e. LPP) [11] and its variant fast and orthogonal LPP (i.e. FOLPP) [43] are also used for comparison. Moreover, capped-norm based methods Robust Feature Selection via Simultaneous Capped L_2-Norm and $L_{2,1}$-Norm Minimization (i.e. SCM) [41], together with state-of-the-art subspace learning methods robust discriminant regression (i.e. RDR) [44], and low-rank linear embedding (i.e. LRLE) [45] are also utilized for comparison.

A. DETAILS OF THE DATABASES

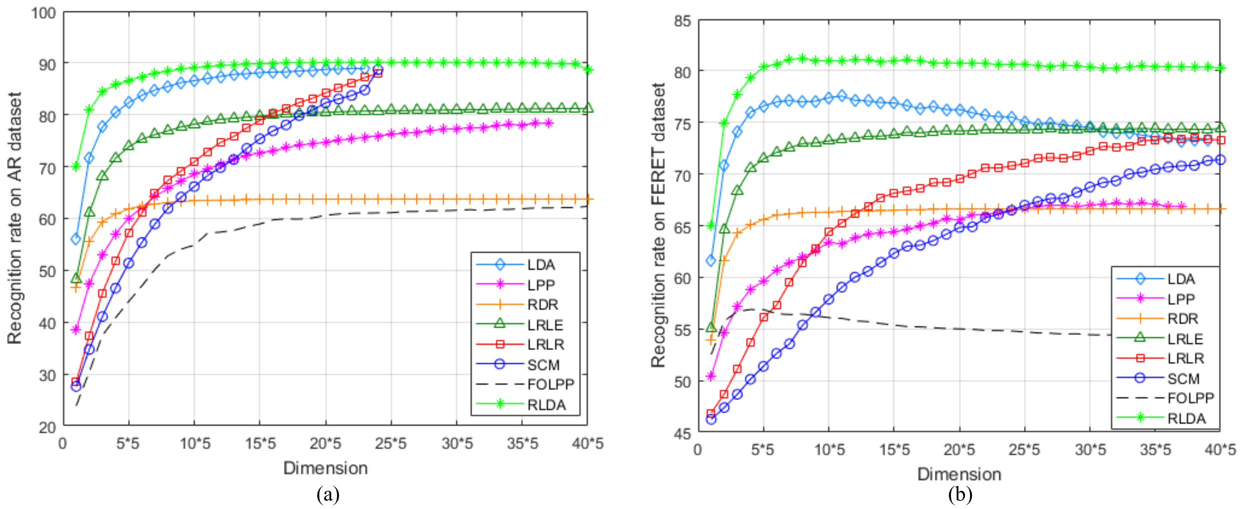
A subset of the AR face database includes 840 images of 120 classes are selected for experiments. A subset of FERET face dataset containing 800 images from 200 individuals (namely, each individual has 4 images) were used to conduct experiments. The Binary alpha digits database consists of 1404 binary handwritten images of 36 classes, and each class is corresponding to 39 images. The COIL-20 database is made up of 1440 images from 20 objects, and each object has 72 figures. Sample images from these datasets are shown in Fig.1.

B. EXPERIMENTAL SETTINGS

In this experiment, we randomly selected L images from each individual and used them as training samples. L was set as $L = 4, 5$ for AR dataset, $L = 3, 4$ for FERET dataset, $L = 10, 15$ for Binary alpha digits database , and $L = 6, 7$ for COIL-20 database. Experiments are performed 10 times repeatedly, and the average recognition of all the methods were calculated with the purpose of evaluating their performances in dimensionality reduction.

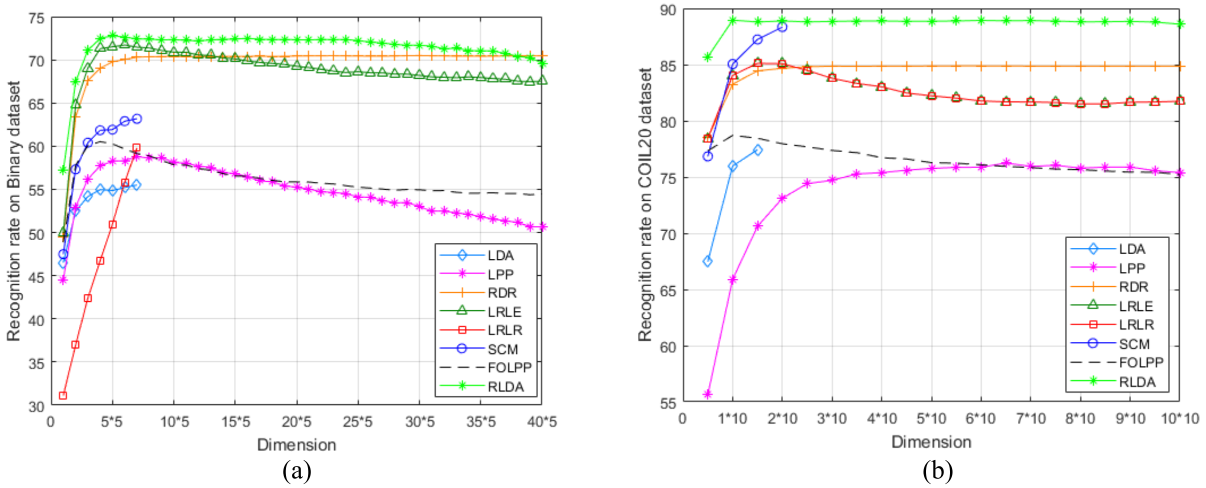
In each experiment, the range of subspace dimensions for AR, FERET and Binary alpha digits were from 5 to 200 with step 5, and for COIL-20 database, the dimension range was set as from 5 to 100 with step 5.

To release the singular value problem caused by the inverse calculation of the scatter matrix, we perform PCA as a pre-processing step before conducting our experiments.



(a) (b)

FIGURE 2. The recognition rates on (a) AR, (b) FERET database.



(a) (b)

FIGURE 3. The recognition rates vs. the dimension of each method on (a) Binary alpha digits database, (b) COIL-20 dataset.

In LPP, FOLPP and RLDA, the weight mode was set as Binary mode, and the neighbor graph W was defined in (7). The nearest neighbor parameter k was selected from the set $\{1, 2, 4, \ldots, C-1\}$, where C denotes the number of classes in training data. K Neighbors Classifier was used for classification.

C. EXPERIMENTAL RESULTS AND ANALYSIS

For all the methods, best average recognition rates and standard deviation, together with corresponding dimension and training time on AR, FERET, Binary alpha digits and COIL-20 face database are presented in Tables 2 to 6, respectively. The variation curves of average recognition rates in relation to the subspace dimension of each method on these databases are shown in Fig. 2 and Fig. 3.

From the tables and figures listed above, we can draw some conclusions as follows:

1. All the experimental results show that our proposed RLDA has superior performance to conventional LDA. The reason is that RLDA is more robust than LDA by redefining the scatter matrices. Moreover, we are able to address the problem in traditional LDA that it can only obtain $C-1$ projections at most, where C is the total number of classes in the training data. In addition, $L_{2,1}$-norm regularization ensures the joint sparsity of the projections, which can select more discriminative features. Furthermore, the locality discriminant information has been taken into consideration in RLDA, which can preserve the local geometric structure when the data is embedding in a latent manifold so as to further improve the performance.

2. We can find that the proposed RLDA always performs better than other compared methods in any low-dimensional subspace. These phenomena appear on all datasets, and show that the proposed method is much powerful than other methods in dimensionality reduction.

3. Since LRLR integrates the property of low-rank, it achieves comparatively better results than conventional L_2-norm based method. SCM utilizes $L_{2,1}$-norm based metric

TABLE 2. Performance on AR dataset.

L	LDA	LPP	RDR	LRLE	LRLR	SCM	FOLPP	**RLDA**
4	88.94±7.49 (115)(0.0468s)	78.38±9.28 (185)(0.0780s)	63.77±3.02 (200)(19.35s)	81.21±6.30 (200)(27.36s)	88.02±16.24 (120)(0.2964s)	88.69±16.73 (120)(0.3120s)	62.34±9.14 (200)(0.0624s)	**90.18±3.53** (200)(2.5116s)
5	91.78±6.22 (115)(0.0624s)	83.66±8.47 (185)(0.1092s)	70.76±3.05 (200)(26.55s)	86.17±6.02 (200)(29.65s)	90.90±15.25 (120)(0.3744s)	91.35±13.97 (120)(0.3588s)	70.73±9.67 (200)(0.0156s)	**92.76±2.93** (200)(2.6988s)

TABLE 3. Performance on FERET dataset.

L	LDA	LPP	RDR	LRLE	LRLR	SCM	FOLPP	**RLDA**
3	77.55±2.69 (195)(0.0624s)	67.17±3.84 (200)(0.1248s)	66.62±2.14 (200)(18.40s)	74.40±3.46 (200)(69.48s)	73.48±7.30 (200)(0.2808s)	71.38±7.36 (200)(0.7956s)	56.87±0.93 (200)(0.1404s)	**81.22±2.65** (200)(3.2448s)
4	78.00±1.57 (195)(0.1872s)	73.91±1.76 (200)(0.1560s)	77.63±1.45 (200)(20.35s)	82.75±2.34 (200)(79.71s)	82.96±5.04 (200)(0.4368s)	76.48±3.07 (200)(1.3728s)	72.86±1.81 (200)(0.1560s)	**86.78±1.76** (200)(3.8064s)

TABLE 4. Performance on binary alpha digits dataset.

L	LDA	LPP	RDR	LRLE	LRLR	SCM	FOLPP	**RLDA**
10	55.49±3.18 (34)(0s)	58.79±3.00 (200)(0.0624s)	70.48±3.47 (200)(9.5005s)	71.71±3.37 (200)(14.60s)	69.79±10.22 (34)(0.2964s)	63.17±5.56 (34)(0.2496s)	60.51±2.16 (200)(0.0624s)	**72.79±2.51** (200)(1.4196s)
15	68.91±3.22 (34)(0s)	72.91±2.82 (200)(0.1092s)	77.33±3.01 (200)11.7313s)	78.43±2.90 (200)(15.66s)	71.04±9.58 (34)(0.3588s)	75.89±6.89 (34)(0.312s)	70.23±1.89 (200)(0.1248s)	**79.18±2.24** (200)(1.5444s)

TABLE 5. Performance on Coil20 dataset.

L	LDA	LPP	RDR	LRLE	LRLR	SCM	FOLPP	**RLDA**
6	77.43±5.38 (15)(0s)	76.24±4.91 (100)(0.5772s)	84.81±1.62 (100)(3.5724s)	85.10±1.55 (100)(2.6988s)	71.86±11.09 (20)(0.1716s)	88.34±5.20 (20)(0s)	78.70±1.05 (100)(0.0624s)	**88.95±0.71** (100)(0.3120s)
7	80.38±3.40 (15)(0.0624s)	78.51±4.51 (100)(0.0624s)	84.87±1.44 (100)(5.8344s)	86.50±1.54 (100)(3.0420s)	75.91±19.26 (20)(0.3276s)	89.60±5.07 (20)(0.0624s)	80.74±1.12 (100)(0.0624s)	**90.35±0.82** (100)(0.5460s)

and capped norms to alleviate the negative of outliers, and the improvements in its robustness leads to its promising recognition rate especially on AR and COIL-20 database. However, both LRLR and SCM encounter the small-class problem. Namely, there exists an upper bound in the total number of the projections obtained by these two methods, and the upper bound is exactly the class number. The small-class problem limits their performances in some extent.

4. In addition to traditional subspace learning methods, we also compared the proposed RLDA with state-of-the art RDR and LRLE. These two methods also achieve prominent performance in dimensionality reduction mainly due to the integration of $L_{2,1}$-norm based metric and local geometric structure. Although both LPP and FOLPP preserve local discriminant information, their recognition rates are unsatisfying compared to the former two methods. The reason for this phenomenon may be that they are sensitive to outliers and variant in data owing to the utilization of L_2-norm based metric.

D. CONVERGENCE ANALYSIS

In section IV-A, we have proven that the proposed objective function is a monotonically decreasing function, and it will converge to local optimum. The value of objective function can be computed by equation in (16). Fig. 4 (a) and Fig. 4 (b) depict the variation of the value of objective function versus the number of iterations on AR and Binary alpha digits datasets, respectively. The figures show that the proposed method converges very fast. Similar phenomena can also be found on other databases, and we can draw a conclusion that the proposed algorithm will converge to the local optimal solutions within a few iterations.

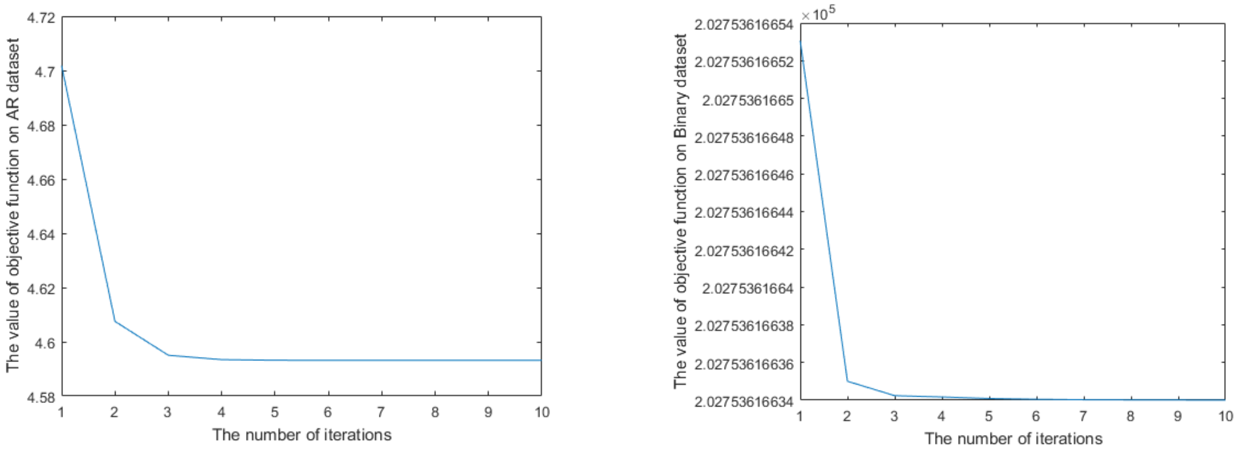
E. PARAMETERS DETERMINATION

Since the performance of our prosed RLDA was affected by the parameters α, β and ε, we can adjust these three parameters to find their optimal combination.

For the regularization prarmeter α, we first fix the other two parameters and vary α in the range of $10^{-5}, 10^{-4}, \ldots, 10^{5}$.

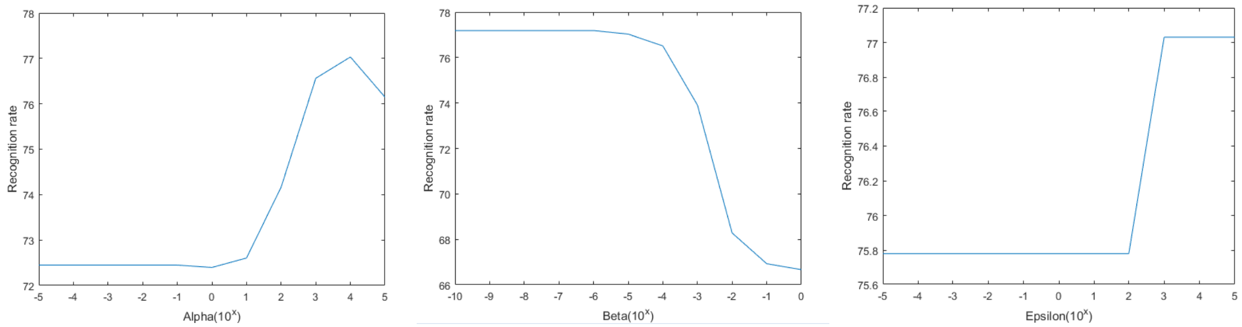
Fig. 5. (a) shows the variation curve of recognition rates versus the variations of parameter α on AR dataset when the other two parameters β and ε were fixed at 10^4, 10^{-5} respectively. It demonstrates that the recognition rate grows as we increase the value of parameter α, and climbs to the highest recognition rate when we set α as 10^4, but it falls when we continue increasing the value of parameter α. Therefore, the range of optimal α is $\alpha \in [10^2, 10^4]$.

Similarly, β and ε were selected from the range of $10^{-10}, 10^{-9}, \ldots, 10^{0}$, $10^{-5}, 10^{-4}, \ldots, 10^{5}$ separately. Fig. 5 (b) shows that the recognition rate continuously decline as parameter β varies from 10^{-5} to 10^{0}, so the range of optimal β is $[10^{-10}, 10^{-6}]$. Fig. 5 (c) indicates that the optimal range of parameter ε is $\varepsilon \in [10^2, 10^4]$. After estimating the



(a) (b)

FIGURE 4. Convergence on (a) AR dataset, (b) Binary alpha digits database.



(a) (b) (c)

FIGURE 5. Parameters determination of RLDA. (a) The variations of recognition rate vs. the parameter α. (b) The variations of recognition rate vs. parameter β. (c) The variations of recognition rate vs. the parameter ε.

optimal ranges of these three parameters using the strategy mentioned above, we can utilize the grid search with cross validation to obtain the optimal parameter combination versus the highest recognition rate on each dataset. The parameter selection processes on other datasets are similar to the process on AR dataset described above.

Eventually, the parameters we set on each database are as follows. For AR database, when $\alpha = 10^4$, $\beta = 10^{-5}$ and $\varepsilon = 10^5$, the proposed RLDA achieves its best performance. For FERET database, the optimal parameters for RLDA are set as $\alpha = 10^4$, $\beta = 10^{-5}$, $\varepsilon = 10^5$. For Binary alpha digits database, the optimal parameters for RLDA are set as $\alpha = 10^4$, $\beta = 10^{-5}$, $\varepsilon = 10^5$. For COIL-20 database, the optimal parameters for RLDA are set as $\alpha = 10^4$, $\beta = 10^{-4}$, $\varepsilon = 10^4$.

VI. CONCLUSIONS

In this paper, we propose a discriminant analysis method called RLDA for dimensionality reduction. In order to solve the problems in conventional LDA, we redefine the data matrix and the scatter matrices. In addition, by introducing capped-L_2 norm on loss function, we enhance the robustness of the proposed method. The $L_{2,1}$-norm regularization term is also utilized to select features with joint sparsity. RLDA takes local geometric structure into account and preserve locality of the data points as LPP. An iterative algorithm is designed to compute the optimal solutions of the proposed RLDA. Theoretical analysis, including convergence analysis and computational complexity, are also presented. To evaluate the performance of our method, we carried out experiments on four datasets, where we compared our proposed methods with seven methods, and the results illustrated that RLDA outperformed all the related methods and stat-of-the-art methods. In the future, we will further explore the property of capped norms to design more robust algorithms.

REFERENCES

[1] J. Zhou, Z. Lai, C. Gao, X. Yue, and W. Wong, "Rough-fuzzy clustering based on two-stage three-way approximations," *IEEE Access*, vol. 6, pp. 27541–27554, 2018.

[2] M. Wan and Z. Lai, "Multi-manifold locality graph embedding based on the maximum margin criterion (MLGE/MMC) for face recognition," *IEEE Access*, vol. 5, pp. 9823–9830, 2017.

[3] F. Zhang, G. Yang, Z. Yang, and M. Wan, "Robust recovery of corrupted image data based on L_{1-2} metric," *IEEE Access*, vol. 6, pp. 5848–5855, 2018.

[4] X. Chen, F. Nie, G. Yuan, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1525–1531.

[5] R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Supervised feature selection with a stratified feature weighting method," *IEEE Access*, vol. 6, p. 15087–15098, 2018.

[6] G. Yuan, X. Chen, C. Wang, F. Nie, and L. Jing, "Discriminative semi-supervised feature selection via rescaled least squares regression-supplement," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 8177–8178.

[7] 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.

[8] X. Chen, M. Yang, J. Z. Huang, and Z. Ming, "TWCC: Automated two-way subspace weighting partitional co-clustering," *Pattern Recognit.*, vol. 76, pp. 404–415, Apr. 2018.

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

[10] R. A. Fisher, "The statistical utilization of multiple measurements," *Ann. Hum. Genet.*, vol. 8, no. 4, pp. 376–386, 1938.

[11] X. He, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, no. 1, 2003, pp. 186–197.

[12] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[13] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2201–2207.

[14] Q. Hua, L. Bai, X. Wang, and Y. Liu, "Local similarity and diversity preserving discriminant projection for face and handwriting digits recognition," *Neurocomputing*, vol. 86, pp. 150–157, Jun. 2012.

[15] P. Huang, C. Chen, Z. Tang, and Z. Yang, "Discriminant similarity and variance preserving projection for feature extraction," *Neurocomputing*, vol. 139, pp. 180–188, Sep. 2014.

[16] P. Huang, C. Chen, Z. Tang, and Z. Yang, "Feature extraction using local structure preserving discriminant analysis," *Neurocomputing*, vol. 140, pp. 104–113, Sep. 2014.

[17] P. Huang, Z. Tang, C. Chen, and Z. Yang, "Local maximal margin discriminant embedding for face recognition," *J. Vis. Commun. Image Represent.*, vol. 25, no. 2, pp. 296–305, 2014.

[18] P. Huang, Z. Yang, and C. Chen, "Fuzzy local discriminant embedding for image feature extraction," *Comput. Elect. Eng.*, vol. 46, pp. 231–240, Aug. 2015.

[19] G.-F. Lu, J. Zou, Y. Wang, and Z. Wang, "Sparse L_1-norm-based linear discriminant analysis," *Multimed. Tools Appl.*, vol. 77, no. 13, pp. 16155–16175, 2018.

[20] B. Yue, S. Wang, X. Liang, and L. Jiao, "Robust coupled dictionary learning with ℓ_1-norm coefficients transition constraint for noisy image super-resolution," *Signal Process.*, vol. 140, pp. 177–189, Nov. 2017.

[21] V. S. Aldea, M. O. Ahmad, and W. E. Lynch, "Hyperspectral classification with adaptively weighted L_1-norm regularization," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2016, pp. 1–4.

[22] G.-F. Lu, G. Tang, and J. Zou, "Spare L_1-norm-based maximum margin criterion," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 11–17, Jul. 2016.

[23] D. Zhang, X. Li, J. He, and M. Du, "A new linear discriminant analysis algorithm based on L_1-norm maximization and locality preserving projection," *Pattern Anal. Appl.*, vol. 21, no. 3, pp. 685–701, 2018.

[24] S. Yuan, X. Mao, and L. Chen, "Elastic preserving projections based on L_1-norm maximization," *Multimed. Tools Appl.*, vol. 77, no. 16, pp. 21671–21691, 2018.

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

[26] M. Qi, T. Wang, Y. Yi, N. Gao, J. Kong, and J. Wang, "Joint $L_{2,1}$ norm and fisher discrimination constrained feature selection for rational synthesis of microporous aluminophosphates," *Mol. Inform.*, vol. 36, no. 4, p. 1600076, Apr. 2016.

[27] P. Cao, X. Liu, J. Zhang, D. Zhao, M. Huang, and O. Zaiane, "$\ell_{2,1}$ norm regularized multi-kernel based joint nonlinear feature selection and over-sampling for imbalanced data classification," *Neurocomputing*, vol. 234, pp. 38–57, Apr. 2017.

[28] P. Cao *et al.*, "A $\ell_{2,1}$ norm regularized multi-kernel learning for false positive reduction in Lung nodule CAD," *Comput. Methods Programs Biomed.*, vol. 140, pp. 211–231, Mar. 2017.

[29] Y. Zhou, Y. Ding, Y. Luo, and H. Ren, "Sparse neighborhood preserving embedding via $L_{2,1}$-norm minimization," in *Proc. 9th Int. Symp. Comput. Intell. Design (ISCID)*, vol. 2, 2016, pp. 378–382.

[30] X. Ma, M. Zhao, Z. Zhang, J. Fan, and C. Zhan, "Anchored projection based capped $L_{2,1}$-norm regression for super-resolution," in *Proc. Trends Artif. Intell. (PRICAI)*, 2018, pp. 10–18.

[31] Q. Sun, S. Xiang, and J. Ye, "Robust principal component analysis via capped norms," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 311–319.

[32] F. Zhang, Z. Yang, Y. Chen, J. Yang, and G. Yang, "Matrix completion via capped nuclear norm," *IET Image Process.*, vol. 12, no. 6, pp. 959–966, 2018.

[33] Q. Lu, X. Li, Y. Dong, and D. Tao, "Subspace clustering by capped ℓ_1 norm," in *Proc. Chin. Conf. Pattern Recognit.*, 2016, pp. 663–674.

[34] M. Zhao, Z. Zhang, C. Zhan, and W. Wang, "Graph based semi-supervised classification via capped $\ell_{2,1}$-norm regularized dictionary learning," in *Proc. IEEE 15th Int. Conf. Ind. Inform. (INDIN)*, Jul. 2017, pp. 1019–1024.

[35] M.-J. Wu, J.-X. Liu, Y.-L. Gao, X.-Z. Kong, and C.-M. Feng, "Feature selection and clustering via robust graph-Laplacian PCA based on capped L_1-norm," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 1741–1745.

[36] D. Ma, Y. Yuan, and Q. Wang, "A sparse dictionary learning method for hyperspectral anomaly detection with capped norm," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 648–651.

[37] H. Jumahong and G. Alimjan, "Face recognition based on rearranged modular two-dimensional locality preserving projection," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 12, p. 1856016, May 2018.

[38] W. Hu, X. Cheng, Y. Jiang, K.-S. Choi, and J. Lou, "Locality preserving projections with adaptive neighborhood size," in *Proc. Intell. Comput. Theories Appl.*, 2017, pp. 223–234.

[39] P. Gong, J. Ye, and C. Zhang, "Multi-stage multi-task feature learning," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2979–3010, 2013.

[40] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, 2003.

[41] G. Lan, C. Hou, F. Nie, T. Luo, and D. Yi, "Robust feature selection via simultaneous sapped norm and sparse regularizer minimization," *Neurocomputing*, vol. 283, pp. 228–240, Mar. 2018.

[42] X. Cai, C. Ding, F. Nie, and H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1124–1132.

[43] R. Wang, F. Nie, R. Hong, X. Chang, X. Yang, and W. Yu, "Fast and orthogonal locality preserving projections for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 5019–5030, Oct. 2017.

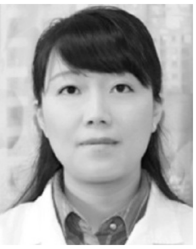
[44] Z. Lai, D. Mo, W. K. Wong, Y. Xu, D. Miao, and D. Zhang, "Robust discriminant regression for feature extraction," *IEEE Trans. Cybern.*, vol. 48, no. 8, pp. 2472–2484, Aug. 2018.

[45] Y. Chen, Z. Lai, W. K. Wong, L. Shen, and Q. Hu, "Low-rank linear embedding for image recognition," *IEEE Trans. Multimed.*, vol. 20, no. 12, pp. 3212–3222, Dec. 2018.

ZHIHUI LAI received the B.S. degree in mathematics from South China Normal University, the M.S. degree from Jinan University, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, China, in 2002, 2007, and 2011, respectively. He has published over 100 scientific articles. He has been a Research Associate, a Postdoctoral Fellow, and a Research Fellow with The Hong Kong Polytechnic University. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization, and applications in the fields of intelligent robot research. He is currently an Associate Editor of the *International Journal of Machine Learning and Cybernetics*.

NING LIU is currently pursuing the B.S. degree with Shenzhen University, Shenzhen. She is also with the College of Computer Science and Software Engineering, Shenzhen University.

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 the 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.



HENG KONG received the B.S. and M.D. degrees from Chongqing Medical University, the M.S. degree from Guangzhou Medical University, and the Ph.D. degree from Southern Medical University, China, in 2000, 2005, and 2008, respectively. She was a Visiting Scholar with the Cancer Center, Georgia Reagent University, Augusta, USA, from 2014 to 2016. She is currently a Professor The Shenzhen University General Hospital, Shenzhen University. She is also doing basic and clinical research associated with breast cancer. Her research interests include gene therapy, immunotherapy, early diagnosis and prognosis analysis of breast cancer, and tumor image processing and recognition using machine learning and artificial intelligent methods.