Evolutionary Intelligence (2019) 12:73–82
https://doi.org/10.1007/s12065-018-0190-0

RESEARCH PAPER

Optimized regularized linear discriminant analysis for feature extraction in face recognition

Xiaoheng Tan[1] · Lu Deng[1] · Yang Yang[1] · Qian Qu[1] · Li Wen[1]

Received: 15 March 2018 / Revised: 21 September 2018 / Accepted: 21 November 2018 / Published online: 3 December 2018

Abstract
In a reduced dimensional space, linear discriminant analysis looks for a projective transformation that can maximizes separability among classes. Since linear discriminant analysis demands the within-class scatter matrix appear to non-singular, which cannot directly used in condition of small sample size (SSS) issues in which the dimension of image is much higher, while the number of samples isn't unlimited. Both the between-class and within-class scatter matrices are always exceedingly ill-posed in SSS problems. And many algorithms are suffered from small sample size issues still. To solve SSS problems, many methods including regularized linear discriminant analysis were proposed. In this article, a way was presented by optimized regularized linear discriminant analysis for feature extraction in FR which can not only fix the singularity problem existing in scatter matrix but also the problem of parameter estimation. The experiment is conducted on several databases and promising results are obtained compared to some state-of-the-art methods to demonstrate the effectiveness of the proposed approach.

Keywords Face recognition · Linear discriminant analysis · Regularized linear discriminant analysis · Recognition accuracy · Small sample size problems

1 Introduction

With the gradual growth of artificial intelligence application and technology, many industry perspective and research interest in face recognition (FR). Face recognition has some advantages, such as low cost, least intrusive and more data sources. However the processing and analysis of high-dimensional data in FR is still a challenge [1–3], robust FR remains a challenging task in small sample environments. There are mainly two stages for a common face recognition system: (1) robust and discriminant feature extraction, such as, principal component analysis (PCA) [4], linear discriminative analysis (LDA) [5], regularized linear discriminant analysis (RLDA) [6]. Null LDA (NLDA) [7], the orientation matrix is calculated in two steps. In the first stage, the data is projected on the null space of S_W and in the second stage it finds W that maximizes $\left|W^T S_B W\right|$, and spectral regression discriminative analysis (SRDA) [8]. Sparse graph-based discriminate analysis (SGDA) [9] was developed by preserving the sparse connection in a block-structured affinity matrix with class-specific samples. Using low-rank constraints, low-rank graph-based discriminate analysis (LGDA) [10] preserves the global structure in data. Sparse and low-rank graph-based discriminate analysis (SLGDA) method was developed in [10] to purse block-diagonal structured affinity matrix with sparsity and low-rank constraints. (2) Classifier construction, e.g. Nearest Neighbor (NN) [11]. But, many methods that include LDA-based statistical learning methods always affected by "small-sample-size" (SSS) problems [12].

Feature extraction has proved to be great in transforming high-dimensional space to lower one, and retain most of the intrinsic information in original data [13, 14]. PCA was originally used to remove zero value for *Sw*, and LDA was then executed in the reduced dimensional subspace. It has lighted that removed zero spaces include discriminatory information that cannot ignore. However, for supervised dimensionality reduction methods which are only suitable for single model data, classification performance is closely related to between-class separation, within-class compactness and equal emphasis on separation between classes [15]. In the technique of RLDA, the Sw matrix is regularized to deal with the *Sw*

✉ Lu Deng
526535207@qq.com

[1] College of Communication Engineering, Chongqing University, Chongqing 400044, China

singularity. The matrix Sw is approximated by $Sw+\eta I$. But it does not consider whether the definition of the scatter matrix is more reasonable. Li and Tang [16] presents the idea that traditional LDA algorithm is not optimal for the definition of between-class scatter matrices. It cannot help separate classes other than edge classes, and it may cause them to overlap with each other, resulting in discriminant performance degradation. Second, a fixed regularization parameter value was introduced in the RLDA, but it may not give the best classification. In [17], an approach, estimating η term by putting the modified Fisher's criterion maximize, presents better performance than other methods. In addition, close class pairs prone to overlap in the subspace, which is referred to as the class separation problem. A number of weighting methods was put forward to deal with this problem [18, 19], and the fundamental thought is to assign large weights to close class pairs. However, the problem cannot be solved thoroughly by those methods [20]. The proximity function proposed by [21] solves the shortcomings of the traditional distance function in high dimensional data. In view of that, this paper proposes an improved RLDA algorithm. It redefines the between-class and introduces a precise regularized parameter to control the deviation and variance of the eigenvalue. Finally, a better method of parameter estimation and the improved scatter matrices are combined.

This study is motivated by the fact that previous studies [6, 17, 21]. The frame of this article is as below. Part 2 gives a detailed introduction of mathematical derivation of regularized linear discriminant analysis algorithm. Section 3 introduces the method with the structure of improved scatter matrices and a precise regularized parameters. Section 4 is the simulation results and analysis. Lastly, the last part is the conclusions.

2 Regularized linear discriminant analysis

Among supervised dimensionality reduction methods, RLDA is the most popular discriminant analysis method to SSS problem, which is widely used in pattern recognition fields. Both the degree of deviation and variance are all decided by the extent of SSS problem. A related algorithm improved by *Friedman* under similar conditions, where estimating S_i for every sample type covariance matrix may not be very appropriate. The solving method, put forward by *Friedman*, is to add a regularization parameter, $\eta \cdot I$, so have $S_i = S_i + \eta I$. I is identity matrix and η is a regularization term. This regularization has influence on increasing the smaller ones and decreasing the larger eigenvalues, thus offsetting the biasing. To stabilize the smallest eigenvalues is another effect of the regularization.

A training set, $z = \{z_i\}_{i=1}^{C}$ including C classes with every class $z_i = \{z_{ij}\}_{j=1}^{C_i}$ consisting of multiple partial images z_{ij}, a total of $N = \sum_{i=1}^{C} C_i$ images can be obtained in the set. For easing computation, each face image is standed by a lexicographic order of pixel elements (i.e. $z_{ij} \in R^J$). The length of it is $J (= Iw \times Ih)$. R^J represents the J-dimensional data space. The method, obtains a discriminant vector by lagging the proportion of the between-class scatter measure to the within-class scatter measure, can be formulated as:

$$W = \arg\max_{W} \frac{\left|W^T S_b W\right|}{\left|\eta(W^T S_b W) + W^T S_w W\right|}. \tag{1}$$

Among them, S_b is between-class scatter matrices, S_w is within-class scatter matrices, $W \in R^m$, $0 \le \eta \le 1$ is regularization parameter. And

$$S_b = 1/N \sum_{i=1}^{C} C_i(\bar{z}_i - \bar{z})(\bar{z}_i - \bar{z})^T \tag{2}$$

$$S_w = 1/N \sum_{i=1}^{C} \sum_{j=1}^{C_i} C_i(z_{ij} - \bar{z}_i)(z_{ij} - \bar{z}_i)^T, \tag{3}$$

where z_i is the average value (or center) of class i, $\bar{z}$ is the total mean (or center) of all the classes. In general, by using the training samples, z_i, $\bar{z}$ can be estimated, i.e., $z_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$ and $\bar{z} = 1/N \sum_{i=1}^{C=1} \sum_{j=i+1}^{C} x_{ij}$.

A series of discriminant vectors would be available by eigenvalue decomposition of $S_w^{-1} S_b$ according to (1) when Sw is full rank. The matrix of projection can be constructed by the eigenvectors that associated to the d eigenvalues those are largest, which is the suboptimal solution to (2). However, as mentioned in Sect. 1, there are some disadvantages on RLDA for FR and it can be improved. To sum up, algorithm proceeds in the following way.

3 Improved regularized linear discriminant analysis

Equation (2) is defined so that all the mean values of the sample and the average values of the classes are separated as much as possible, but the mean values of various class may be close to each other, resulting in overlapping of many samples of adjacent classes, resulting in a decrease in recognition performance. The reason for this problem is that the variance is the largest in the most discriminating projection direction obtained by the previous algorithm. So the edge class and other class can be separated as much as possible. It should be noted, however, that this direction does not help separate other class except the edge class, and it may cause

them to overlap with each other, resulting in a decline in discriminant performance.

Therefore, the existing algorithm for the definition of the scatter of the between-class is not optimal, because the edge class dominates the feature decomposing, resulting in the dimension reduction of the conversion matrix is too much emphasis on those who already have been better separated from the class. Then the overlapping of adjacent class be caused.

3.1 The model of improved between-class scatter matrices

Improved scatter matrices model was expressed as:

$$S_b = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} P_i P_j Close(\bar{z}_i, \bar{z}_j)(\bar{z}_i - \bar{z}_j)(\bar{z}_i - \bar{z}_j)^T, \tag{4}$$

where the Close function:

$$\text{Close}(\bar{z}_\text{i}, \bar{z}_\text{j}) = \frac{1}{\text{m}} \sum_{k=1}^{m} e^{-|\overline{z_{\text{ik}}} - \overline{z_{\text{jk}}}|}. \tag{5}$$

The value of a sample is in m-dimensional space. The range of the Close function is (0, 1), which indicates the proximity of the sample $\bar{z}_\text{i}$ to the sample $\bar{z}_\text{j}$. The closer, the greater the value of the function. On the contrary, the smaller, the value. Where P_i and P_j are the prior probability of class i and j, respectively, and $\bar{z}_\text{i}$ and $\bar{z}_\text{j}$ are the average values of i-th and j-th.

From Eq. (3), the larger the $\left\|\bar{z}_i - \bar{z}_j\right\|$ value, the smaller the weight assigned to them; on the other hand, the greater the weight assigned to them.

3.2 The model of improved within-class scatter matrices

Many algorithms need to train scatter matrices on larger databases. But in practice, the number of training set isn't unlimited. Under the condition of small sample, the model cannot be correctly and effectively represent the logic and characteristics of the model, and it is easy to obtain the problem of over-fitting of the scatter matrices, which makes the performance of face recognition significantly lower. Because the difference of characteristics of the same person is more susceptible to other factors, even greater than the difference between different characteristics, that is, the degree of scatter within the class is greater than the between-class scatter changes, making the estimation error greater. Therefore, under the condition of small sample, the within-class scatter is obviously more sensitive, and this paper also pays attention to solve the sensitivity of within-class scatter matrices to small samples.

When the data is less effective in data set, the effective information of samples can be robustly estimated by making full use of local data structure of the sample. When there is an outlier in the data set, the local data structure adjacent to the sample can also be used to represent the characteristics of the outlier. For the small sample caused by the over-fitting problem, you can solve the problem by smoothing. In this paper, the KNN algorithm was used to select the within-class scatter matrices of adjacent classes and by taking advantage of local data structure. The within-class scatter matrices are smoothed, the over-fitting problems caused by small samples can be solved.

Let the training sample data set be: $z_{ij} \in R$, $i = 1,..., C$ and $j = 1,..., C_i$. C_i is the number of classes i in the training sample, C is class number, N is the total sample number, and z_{ij} represents the j-th face image of the i-th class of the training sample. Within-class scatter matrices model can be express as:

$$S_i = \sum_{j=1}^{i} (\text{z}_{ij} - \bar{z}_i)(z_{ij} - \bar{z}_i)^T. \tag{6}$$

The general within-class scatter matrices was formulated as:

$$S_\text{w} = 1/N \sum_{i=1}^{C} S_i. \tag{7}$$

Using the adjacent class to smooth the class divergence matrix

$$\tilde{S}_i = \beta S_i + (1 - \beta) \sum\nolimits_{k \in KNN(i)} \omega_k S_k. \tag{8}$$

$k \in \text{KNN}(i)$ represents the K nearest neighbors of class i. $\beta \in [0, 1]$ is the trade-off parameter, k is the weight parameter determined by the nearest neighbor system, the smaller the distance, the greater the weight.

Improved within-class scatter matrices model can be express as:

$$S_\text{w} = 1/N \sum_{i=1}^{C} \tilde{S}_i. \tag{9}$$

By the definition, $\tilde{S}_1$ is the result of the smoothing of S_i and S_k of S_i K nearest neighbor classes of S_i, and the problem of fitting can be solved by making full use of class i samples and adjacent class sample information. When a class has only one sample, the scatter matrix cannot be estimated effectively, but the scatter matrix can be approximated by using the neighboring class samples. The smoothing method takes full advantage of local data structure and reduces adverse effects of outliers in each class. The improved algorithm makes full use of the within-class scatter matrices of the class, and solves the problem of over-fitting of the general within-class scatter matrices, and obtains exact within-class scatter matrices.

3.3 A deterministic approach to RLDA

Let S_T, S_W, S_B denotes the total, within-class and between-class scatter matrix, respectively. The scatter matrices would be singular in condition of SSS. It is well known that the discriminant information does not exist in the zero space of S_T. Thus, the feature dimensionality from d-dimension can drop to rt-dimension (where rt is the rank of S_T) by advance processing of PCA. The range space of S_T matrix, $P_1 \in R^{d\times rt}$, will be applied as a transformation matrix. In reduced dimension, the scatter matrices are: $\mathrm{S_w} = \mathrm{P_1^T S_W P_1}$ and $\mathrm{S_b} = \mathrm{P_1^T S_B P_1}$. After this procedure $S_w \in R^{rt\times rt}$ and $S_b \in R^{rt\times rt}$ are decreased dimensional scatter matrix.

In RLDA, the regularization of within-class scatter matrix S_W was performed by adding η to diagonal elements of S_W; i.e., $S_w = S_w + \eta I$. The η make S_W presents non-singular and reversible which would benefit the revised Fisher's criterion maximized:

$$\mathrm{W} = \arg\max_{W} \frac{\left|\mathrm{W^T S_b W}\right|}{\left|\mathrm{W^T(S_w + \eta I)W}\right|}, \tag{10}$$

where $w \in R^{rt\times 1}$ is orientation vector. Avoid using any heuristic method in determining η, solving Eq. (10) in the belowing way. Denote

$$\mathrm{f} = \mathrm{W^T S_b W}. \tag{11}$$

Constraint condition:

$$\mathrm{g} = \mathrm{W^T(S_w + \eta I)W} - \mathrm{b} = 0. \tag{12}$$

$b > 0$ is constant. Under constrained curve g, the restricted relative maximum of f can be obtained. By putting its derivative to zero value, then

$$\frac{\partial(\mathrm{f} - \lambda g)}{\partial \mathrm{W}} = 2\mathrm{S_b W} - \lambda(2\mathrm{S_w} + 2\eta W) = 0.$$

Or

$$\left(\frac{1}{\lambda}\mathrm{S_b} - \mathrm{S_w}\right)\mathrm{W}\eta - \mathrm{W} = 0 \tag{13}$$

λ is Lagrange's multiplier ($\lambda \neq 0$). Shifting ηW from Eq. (13) into Eq. (12), we conclude

$$\mathrm{W^T S_b W} = \lambda \mathrm{b}. \tag{14}$$

And from Eq. (12) and Eq. (14), we can get

$$\lambda = \frac{\mathrm{W^T S_b W}}{\mathrm{W^T(S_w + \eta I)W}}. \tag{15}$$

We can observe that the left term of Eq. (15) is the Lagrange's multiplier, and to the right of Eq. (15) same as the Fisher's revised criterion. To large the modified Fisher's criterion, we need to maximize λ. So approximate value of λ can be got by maximizing $W^T S_b W / W^T S_w W$, W corresponding to the large eigenvalue of $\mathrm{S_w^{-1} S_b}$. But, $\mathrm{S_w^{-1}}$ can be replaced by its pseudoinverse for which it is singular and irreversible. We can get λ_{max} by decomposing the eigenvalue of matrix $\mathrm{S_w^+ S_b}$. $\mathrm{S_w^+}$ is the pseudoinverse of S_w. The value of λ_{max} can be substituted as follows:

$$\lambda \max = \max\left(\frac{\mathrm{W^T S_b W}}{\mathrm{W^T(S_w + \eta I)W}}\right) \approx \max\left(\frac{\mathrm{W^T S_b W}}{\mathrm{W^T S_w W}}\right) \tag{16}$$

$\approx$ the maximum eigenvalue of $\mathrm{S_w^+ S_b}$.

Equation (16) will help us to seek the value of η by decomposing the eigenvalue of $1/\lambda S_b - S_w$ which will give $r_b = rank(S_b)$ finite eigenvalues. Since the dominant eigenvalue correspond to the largest discriminant eigenvector, η is considered to be the maximum eigenvalue. Then,

$$\eta = \Lambda_{\max}, \tag{17}$$

where $1/\lambda S_b - S_w = E\Lambda E^T$, $E \in R^{rt\times rt}$ is a matrix of eigenvectors, Λ is a diagonal matrix of corresponding eigenvalues. If η is determined, the projection vector W would be obtained by decomposing the eigenvalue of $(S_w + \eta I)^{-1} S_b$ which can be formulated as:

$$((\mathrm{S_w} + \eta \mathrm{I})^{-1}\mathrm{S_b})\mathrm{W} = \beta \mathrm{W}. \tag{18}$$

The m eigenvectors be obtained by Eq. (18) corresponding to the m highest eigenvalues to form W.

Algorithm

1:PCA is used to get the range space $P_1 \in R^{d\times rt}$ of matrix S_T and transform d-dimension to rt–dimension, and rt=rank(S_T). Find matrix $\mathrm{S_b} = \mathrm{P_1^T S_B P_1}$ and $\mathrm{S_w} = \mathrm{P_1^T S_W P_1}$ ($S_b \in R^{rt\times rt}$ and $S_w \in R^{rt\times rt}$)

2: Find the largest eigenvalue λ_{max} by carry outting EVD of $\mathrm{S_w^+ S_b}$

3: Calculate EVD of ($1/\lambda_{max} S_b - S_w$) to find its largest eigenvalue η

4: Calculate EVD of $(S_w + \eta I)^{-1} S_b$ to find h eigenvectors $w_j \in R^{rt\times 1}$ corresponding to the main eigenvalues, where $0 \leq m \leq r_b$ and r_b=rank(S_b). The eigenvectors $\mathrm{w_j}$ are column vectors of $W' \in R^{rt\times m}$

5: Find projection matrix $W \in R^{d\times m}$ in a d-dimension; i.e., $W = P_1 W'$

4 Simulation results and analysis

In this part, our approach is compared with a number of related state-of-the-art methods, including LGDA [10], NLDA [7], SRDA [8], SGDA [9] and SLGDA [10], etc. The range of normalized parameters in RLDA [6] is [0, 1]. RLDA in the following experiments has better results when the value obtained is 0.001. Our algorithm solves the difficulty of determining the normalized parameter values in RLDA. And the parameters introduced in our algorithm are β and k which can be seen in formula (8). When the values in the following experiments are 0.5 and 10 respectively, this algorithm achieves comparatively better results. Under different dimensions or training samples or classes, the parameters can be changed to other values so that the performance can be much better. NLDA verifies that the zero space of the between-class scatter matrices contains important discriminative information, but in some cases, the between-class scatter matrices may not contain null space. So the results were not so great sometimes. The Tikhonov regularizer was used in SRDA to control the model complexity, but the projection matrix obtained by it does not have orthogonality and it is not conducive to eliminating information redundancy between samples. All those algorithms of feature extraction are combined with NN classification algorithm for face recognition. The experiment is conducted on three face datasets, including the Extended Yale B [22], CMU PIE [23] and AR to evaluate performance. Details of datasets can be seen in Table 1 and Fig. 1.

The parameters in competing methods are adjusted to their best performance according to the suggestions in original papers.

Table 1 The three data sets used in our experiments

Datasets	#Sample	#Dimension	#Classes
Extended Yale B	2414	1024	38
AR	1400	1260	100
CMU PIE	1680	1024	68

4.1 2-D visualization experiment on CMU PIE dataset

In this part, the discriminate ability is showed by different methods using a partial CMU PIE [22] face database. In the experiment, each individual are randomly selected 7 images for training, and the remaining about 17 images were tested. Figure 2a–j visualize the testing data distribution along the first two dimensions obtained by different methods. From Fig. 2, we may draw several conclusions. First, considered small sample problem, NLDA [7], RLDA [6], SRDA [8] are superior to PCA [4] and LDA [5]. But overlaps are still serious. Second, SGDA [9] only uses the local neighborhood structure through sparse representation, which doesn't perform very well. Some parts of 5 classes mixed together can be seen in Fig. 2g. LGDA [10] shows better separation ability by introducing global low-rank regularization, But, still have significant overlaps among class 2, class 3 and class 5, and the distance between class 1 and class 3 is not far. With both sparse and low-rank constraints, SLGDA [10] performs better than the first two. However, class 2, class 3 and class 5 are still not separated as shown in Fig. 2i.

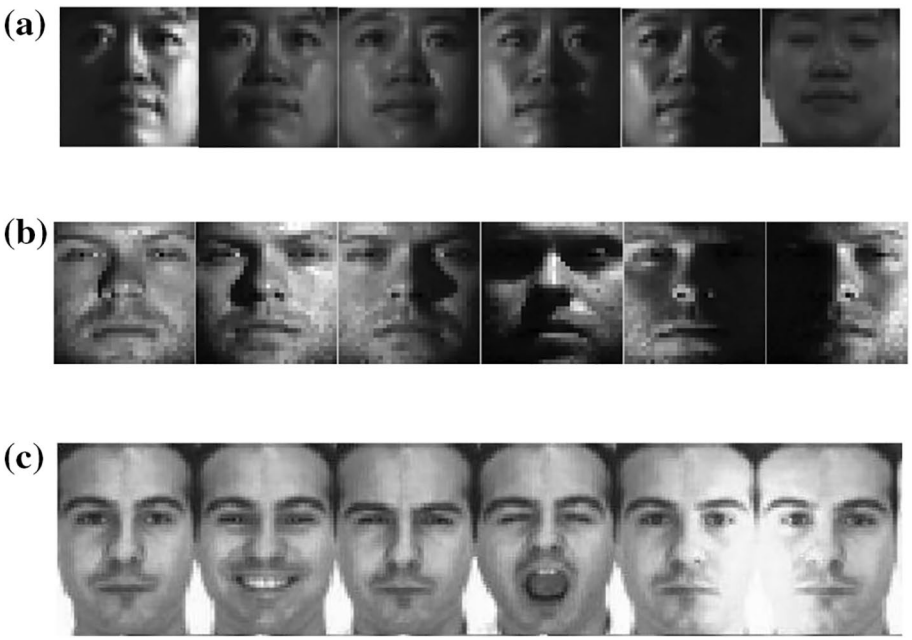

Fig. 1 Some facial images used in the experiments: **a** AR; **b** CMU PIE; **c** extended Yale B

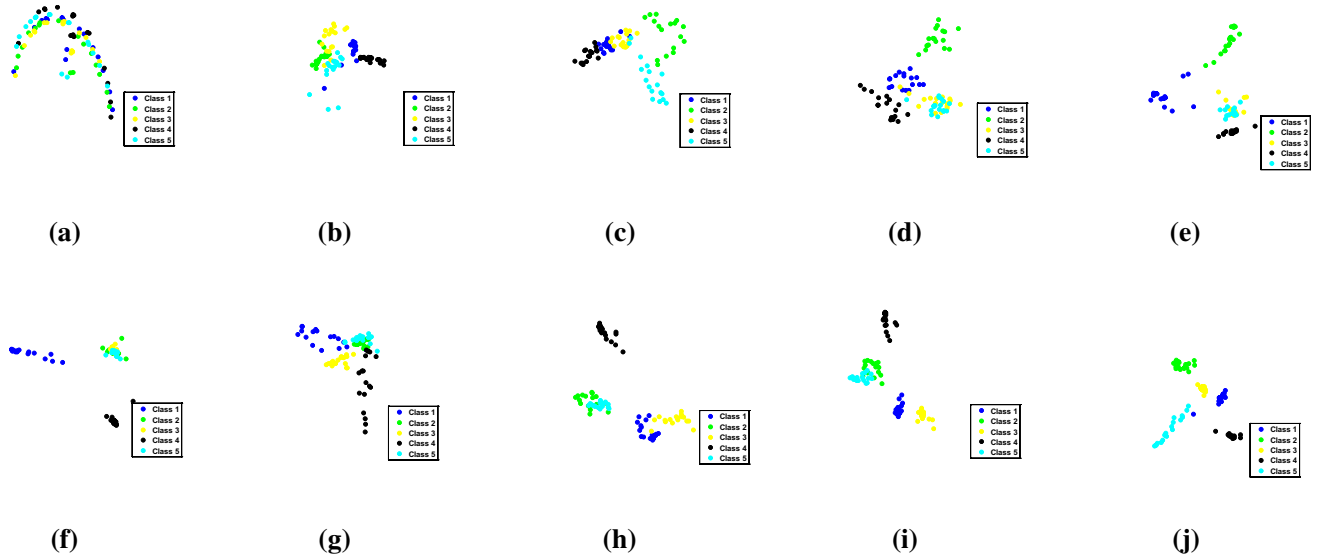

Fig. 2 Two-dimensional five-class CMU PIE data projected by different methods. **a** PCA; **b** LDA; **c** NLDA; **d** RLDA; **e** PCA + LDA; **f** SRDA; **g** SGDA; **h** LGDA; **i** SLGDA; **j** OURS

Contrastively, the proposed method show more clear boundaries among classes and shows stronger robustness in the following experiment.

4.2 Experiments on face recognition

4.2.1 CMU PIE database

The CMU PIE have surpassed four thousands images of sixty-eight individuals. Each person's image was obtained through 13 different postures. Here, we use a subset of poses close to the front, C07, for the experiment, which contains 1629 images of 68 people. Everyone has about 24 pictures. And all the facial images were pruned to 32×32 pixels. Each individual is selected to have a subset of p (= 2, 3 ...) samples for training, the rest for testing. For each p, we ran all of methods 10 times independently, and reported the average results in Table 2. The FR rates under different dimensions are shown in Table 3. Table 2 shows that our method almost exceeds the other methods in different experimental settings. The results in LGDA and SLGDA are similar with ours, while obviously lower than ours when training samples per subject are not much available. Also, LGDA and SLGDA can get better performance as ours under different dimensions when training samples per subject are fixed. The results in RLDA are better than in SRDA under different training samples per subject except p = 2. And the results in NLDA are better than both RLDA and SRDA. Our method has higher recognition rate under different dimensions. Figure 2 shows the recognition rate versus the number of training samples and feature dimensions on CMU PIE by using some methods. From Fig. 2b,

Table 2 Recognition rates under different number of training set (CMU PIE database)

Method	#Training samples per subject on CMU PIE dataset (50 feature dimension)				
	2	3	4	5	6
PCA	0.3182	0.4679	0.4958	0.5514	0.6782
LDA	0.7214	0.8344	0.8736	0.8945	0.9104
PCA + LDA	0.7087	0.8493	0.8849	0.9077	0.9201
SRDA	0.7488	0.8207	0.8617	0.8871	0.9081
NLDA	0.7733	0.8539	0.8924	0.9112	0.9224
RLDA	0.7547	0.8482	0.8861	0.9125	0.9165
SGDA	0.7401	0.8021	0.8747	0.9131	0.8943
LGDA	0.7388	0.8442	0.8954	0.9301	0.9320
SLGDA	0.7429	0.8428	0.8946	0.9302	0.9351
OURS	0.7937	0.8725	0.9098	0.9320	0.9341

Table 3 Recognition rates under the condition of different feature dimensions

Method	#Feature dimension (4 samples per subject) on CMU PIE				
	50	100	150	200	250
PCA	0.4958	0.5343	0.5534	0.5549	0.5565
LDA	0.8736	0.8626	0.8478	0.8400	0.8389
PCA + LDA	0.8853	0.8855	0.8871	0.8846	0.8864
NLDA	0.8849	0.8910	0.8911	0.8949	0.8870
RLDA	0.8861	0.8873	0.8889	0.8850	0.8874
SGDA	0.8747	0.8968	0.8983	0.9064	0.9019
LGDA	0.8954	0.8953	0.9000	0.9037	0.9093
SLGDA	0.8946	0.9005	0.9065	0.9104	0.9097
OURS	0.9013	0.9050	0.9055	0.9097	0.9076

Table 4 Recognition rates under different number of training set

Method	#Training samples per subject on AR dataset (50 feature dimensions)				
	2	3	4	5	6
PCA	0.3303	0.4401	0.4664	0.5688	0.6025
SGDA	0.5775	0.6418	0.6530	0.7167	0.8325
LGDA	0.5350	0.5655	0.5720	0.6033	0.7438
SLGDA	0.5671	0.6173	0.6330	0.7189	0.8300
LDA	0.7495	0.8848	0.9292	0.9502	0.9559
PCA + LDA	0.7667	0.8820	0.9184	0.9426	0.9555
SRDA	0.8154	0.8833	0.9097	0.9224	0.9235
NLDA	0.8478	0.9247	0.9534	0.9652	0.9692
RLDA	0.7821	0.8855	0.9213	0.9479	0.9554
OURS	0.8282	0.9204	0.9505	0.9651	0.9706

Table 5 Recognition rates under the condition of different feature dimensions

Method	#Feature dimension (4 samples per subject) on AR				
	50	100	150	200	250
PCA	0.4664	0.4950	0.5012	0.5050	0.5103
SGDA	0.6530	0.7170	0.6940	0.6900	0.6890
LGDA	0.5720	0.6220	0.6110	0.6150	0.6410
SLGDA	0.6330	0.7170	0.6970	0.7050	0.7010
LDA	0.9292	0.9381	0.9328	0.9227	0.9228
PCA + LDA	0.9184	0.9175	0.9226	0.9179	0.9197
NLDA	0.9534	0.9587	0.9529	0.9516	0.9505
RLDA	0.9174	0.9201	0.9203	0.9240	0.9261
OURS	0.9535	0.9579	0.9618	0.9596	0.9593

it is obvious that the whole training information does not provide significant advantage for classification, by which lead to computational costs instead. Thereby it's necessary to extra features.

4.2.2 Experiments on AR database

A subset of AR consisted of 50 men and 50 women in two sessions, with 6 lighting and 8 expression changes. From session 1, only seven images of light and expression changes, seven samples from another session. Each individual is selected to have a random subset of p (= 2, 3 ...) samples for training. For each p, we ran 10 times independently, and reported the average results in Tables. From Table 4, one can conclude that all the algorithms achieve better performance with the number of training samples per class increases, our method has higher recognition rate than other methods under numbers of every individual of training samples. The results in RLDA are better than in SRDA under different training samples per subject except p = 2. And the results in NLDA are better than both RLDA and SRDA. The FR rates under different dimensions were listed in Table 5. It can show that NLDA and our method exceed other methods in different experimental settings. The NLDA gain best outcomes on AR, which is slightly better than our method under 50 feature dimensions. But it doesn't better than our algorithm under higher dimensions. Particularly, because of the number of training samples per subject are just 14 in AR Database, the performance of SGDA, LGDA and SLGDA are drop sharply when p does not reach half. Figure 3 shows the recognition rate versus different number of training samples per class and feature dimensions by using some methods. From Fig. 3b, it is obvious that the whole training information does not provide significant advantage for classification, by which lead to computational costs instead.

4.2.3 Yale face database

About 2414 images of 38 people each and 64 frontal face of different lighting on Extended Yale B. In this experiment, the cropped and resized images were used which is 32 × 32 pixels. Figure 1 shows some example images of individual.

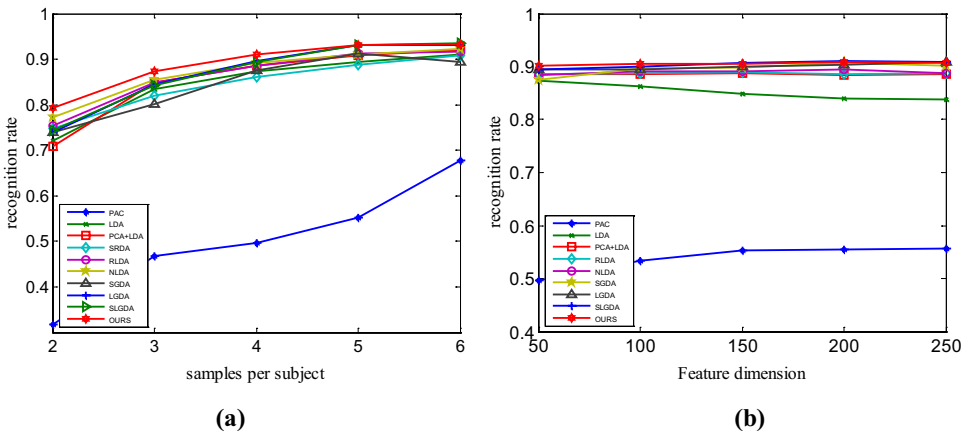

Fig. 3 Face recognition accuracy versus **a** number of training samples per subject, **b** feature dimension on CMU PIE

Table 6 Recognition rates under different number of training set

Method	#Training samples per subject on Yale dataset (30 feature dimension)						
	3	4	5	6	8	12	16
PCA	0.1973	0.2282	0.2515	0.2781	0.3156	0.3643	0.4083
LDA	0.5301	0.6119	0.6539	0.7088	0.7547	0.8309	0.8615
PCA + LDA	0.5668	0.6439	0.6763	0.7155	0.7631	0.8164	0.8471
SRDA	0.5752	0.6346	0.7027	0.7031	0.7289	0.6855	0.6614
NLDA	0.6113	0.6915	0.7421	0.7638	0.8156	0.8438	0.8507
RLDA	0.5654	0.6310	0.6830	0.7189	0.7577	0.8137	0.8481
SGDA	0.5469	0.6414	0.6919	0.7095	0.7928	0.8488	0.8887
LGDA	0.5126	0.6595	0.7176	0.7607	0.8317	0.8903	0.9114
SLGDA	0.5917	0.6626	0.7212	0.7758	0.8360	0.8963	0.9191
OURS	0.5975	0.6852	0.7474	0.7868	0.8388	0.8932	0.9137

Table 7 Recognition rates under different feature dimensions

Method	#Feature dimension (16 samples per subject) on Yale				
	15	20	25	30	35
PCA	0.2434	0.3182	0.3659	0.4083	0.4392
LDA	0.7580	0.8021	0.8424	0.8596	0.8730
PCA + LDA	0.7484	0.8069	0.8305	0.8477	0.8592
NLDA	0.8512	0.8568	0.8506	0.8446	0.8440
RLDA	0.7940	0.8059	0.8349	0.8481	0.8575
SGDA	0.8333	0.8654	0.8787	0.8887	0.8970
LGDA	0.8337	0.8609	0.9031	0.9014	0.9191
SLGDA	0.8386	0.8663	0.8980	0.9091	0.9258
OURS	0.8382	0.8652	0.8927	0.9088	0.9217

A subset of individuals with p (= 3, 4, 5...) samples were taken with labels as for training, and the remaining is used for testing. The experiment was ran for 10 times. From Table 6, one can conclude that all the methods achieve better performance along with the number of training images per subject grows. Then, we randomly select four images from every person for training, and use the remaining samples for testing. The results in RLDA are better than in SRDA under different training samples per subject except p = 3, 4, 5. And the results in NLDA are better than both RLDA and SRDA. SGDA doesn't present superior performance in this dataset as in CMU PIE. The results in LGDA and SLGDA are similar with ours in some cases, while obviously lower than ours when training samples per subject are not much available. Also, LGDA and SLGDA can get better performance as ours under different dimensions when training samples per subject are fixed. The rates, under the condition of different dimensions, were listed in Table 7. Our method exceeds the other methods in different experimental settings, and shows more robust in dealing with illumination problem in FR. Figure 4 vividly illustrate the recognition rate versus the number of training samples for each category and feature dimensions by using some methods. From Fig. 4b, it is obvious that the whole training information does not provide significant advantage for classification, by which lead to computational costs instead (Fig. 5).

5 Conclusions

The issue of small sample size in FR is studied in this paper. The algorithm of regularized linear discriminant analysis still has some disadvantages to fix the SSS problems.

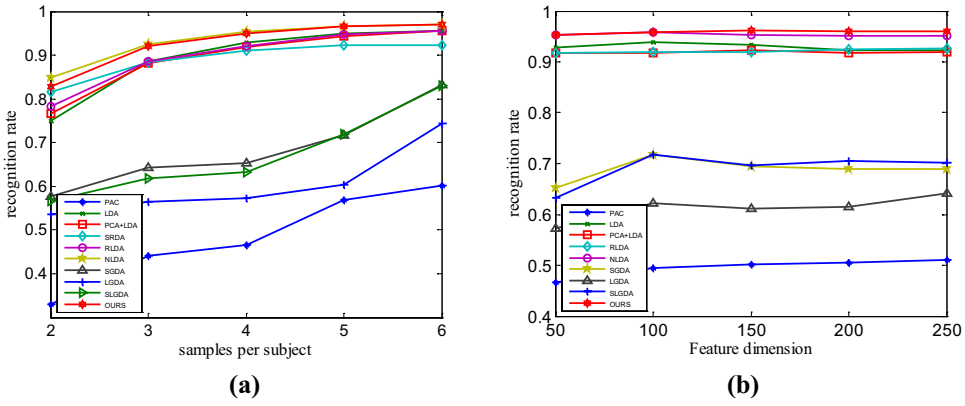

Fig. 4 Face recognition accuracy versus **a** number of training samples per class, **b** feature dimension on AR

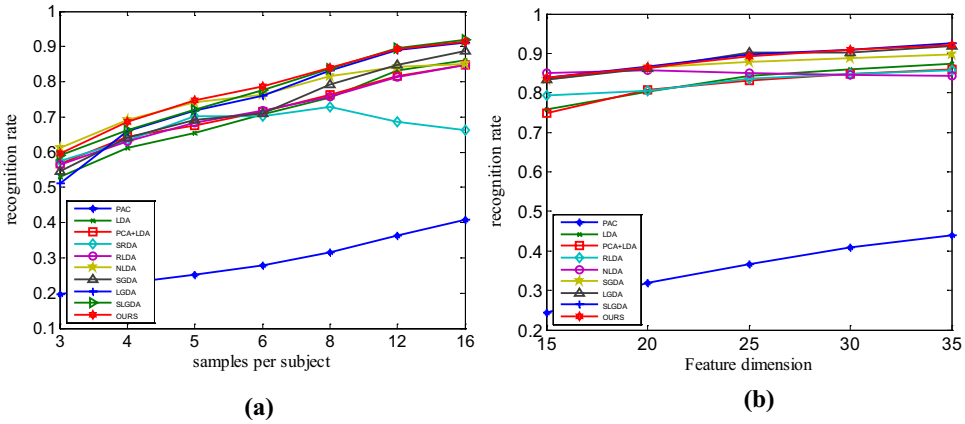

Fig. 5 Face recognition rates versus **a** number of training samples per subject, **b** feature dimension on Yale

Considering that the model of scatter matrices can be more reasonable and related parameter can be obtained by avoiding the process of heuristic. An improved algorithm is introduced, which cannot only fix the singularity problem of scatter matrix but also the problem of parameter estimation. PCA is simply to calculate, and performs well in some cases, but the performance is limited by its unsupervised nature. By introducing different discrimination standards to fix SSS problems, RLDA, NLDA, SRDA and etc perform well to some extent. SGDA, LGDA and SLGDA can adaptively select neighbors for graph construction, and use the labeled samples in the same class to find the representation of each sample for block-diagonal structure representations. However, due to the limited number of samples per class, this process may result in large representation error, which may not reveal the within-class adjacent relationship as well as ours do. So, SGDA, LGDA and SLGDA hardly perform better than the proposed method when training sample are not enough. The simulation results on the famous databases illustrate that the proposed method has much better performance than other methods and improves the face recognition.

Acknowledgements This work was supported by the National Natural Science Foundation of China (no. 61571069) and Project no. 106112017CDJQJ168817 supported by the Fundamental Research Funds for the Central Universities.

References

1. Zhang L, Zhang D (2017) Evolutionary cost-sensitive extreme learning machine. IEEE Trans Neural Netw Learn Syst 28(12):3045–3060
2. Zhang L, Zhang D (2016) Visual understanding via multi-feature shared learning with global consistency. IEEE Trans Multimed 18(2):247–259
3. Sha C, Zhao H (2017) Design and analysis of associative memories based on external inputs of continuous bidirectional associative networks. Neurocomputing 266:433–444
4. Jolliffe I (2005) Principal component analysis. Wiley Online Library, New York
5. Lee S, Park YT, d'Auriol BJ et al (2012) A novel feature extraction method based on normalized mutual information. Appl Intell 37(1):100–120
6. Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84(405):165–175
7. Chen L-F, Liao H-Y, Ko M-T, Lin J-C, Yu G-J (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognit 33:1713–1726
8. Cai D, He X, Han JS (2008) An efficient algorithm for large-scale discriminant analysis. IEEE Trans Knowl Data Eng 20(1):1–12
9. Ly NH, Du Q, Fowler JE (2014) Sparse graph-based discriminant analysis for hyperspectral imagery. IEEE Trans Geosci Remote Sens 52:3872–3884
10. Li W, Liu J, Du Q (2016) Sparse and low-rank graph for discriminant analysis of hyperspectral imagery. IEEE Trans Geosci Remote Sens 54:4094–4105
11. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
12. Lu GF, Wang Y, Zou J (2016) Graph maximum margin criterion for face recognition. Neural Process Lett 44(2):1–19
13. Kasun LLC, Yang Y, Huang GB et al (2016) Dimension reduction with extreme learning machine. IEEE Trans Image Process 25(8):1–1
14. Zhang Q, Deng K, Chu T (2016) Sparsity induced locality preserving projection approaches for dimensionality reduction. Neurocomputing 200:35–46
15. Shao G, Sang N (2014) Max–min distance analysis by making a uniform distribution of class centers for dimensionality reduction. Neurocomputing 143:208–221
16. Li K, Tang P (2014) An improved linear discriminant analysis method and its application to face recognition. Appl Mech Mater 556:4825–4829
17. Sharma A, Paliwal KK (2015) A deterministic approach to regularized linear discriminant analysis. Neurocomputing 151:207–214
18. Tao D, Li X, Wu X, Maybank SJ (2009) Geometric mean for subsPACe selection. IEEE Trans Pattern Anal Mach Intell 31(2):260–274
19. Loog M, Duin R, Haeb-Umbach R (2001) Multiclass linear dimension reduction by weighted pairwise Fisher criteria. IEEE Trans Pattern Anal Mach Intell 23(7):762–766
20. Bian W, Tao D (2011) Max–min distance analysis by using sequential SDP relaxation for dimension reduction. IEEE Trans Pattern Anal Mach Intell 33:1037–1050

21. Shao C, Lou W, Yan L-M (2011) Optimization of algorithm of similarity measurement in high-dimensional data. Comput Technol Dev 21(2):1–4
22. Georghiades AS, Belhumeur PN, Kriegman DJ (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660
23. Sim T, Baker S, Bsat M (2003) The CMU pose, illumination, and expression database. IEEE Trans Pattern Anal Mach Intell 25(12):1615–1618