Learning Regularized LDA by Clustering

Yanwei Pang, *Senior Member, IEEE*, Shuang Wang, and Yuan Yuan, *Senior Member, IEEE*

***Abstract*— As a supervised dimensionality reduction technique, linear discriminant analysis has a serious overfitting problem when the number of training samples per class is small. The main reason is that the between- and within-class scatter matrices computed from the limited number of training samples deviate greatly from the underlying ones. To overcome the problem without increasing the number of training samples, we propose making use of the structure of the given training data to regularize the between- and within-class scatter matrices by between- and within-cluster scatter matrices, respectively, and simultaneously. The within- and between-cluster matrices are computed from unsupervised clustered data. The within-cluster scatter matrix contributes to encoding the possible variations in intraclasses and the between-cluster scatter matrix is useful for separating extra classes. The contributions are inversely proportional to the number of training samples per class. The advantages of the proposed method become more remarkable as the number of training samples per class decreases. Experimental results on the AR and Feret face databases demonstrate the effectiveness of the proposed method.**

***Index Terms*— Dimensionality reduction, face recognition, feature extraction, linear discriminant analysis (LDA).**

I. Introduction

THE ratio r of the number n of training samples per class to the number D of features is closely related to the generalization ability of a pattern classification system. Generally speaking, the generalization ability increases with the ratio. On one hand, when the number of training samples is fixed, it is helpful to reduce the number of features. Dimensionality reduction techniques such as linear discriminant analysis (LDA), principal component analysis (PCA), and locality preserving projection (LPP) [11] are powerful for reducing the number of features and increasing the ratio r. On the other hand, the performance of the dimensionality reduction techniques deteriorates when the number n of training samples per class is very small. This problem is serious in many cases. For example, in face recognition system, only several face images per person can be obtained in enrolling stage. High recognition rates have been reported when n is larger than or equal to 5. But, we find that the recognition rates drop severely when $n < 5$.

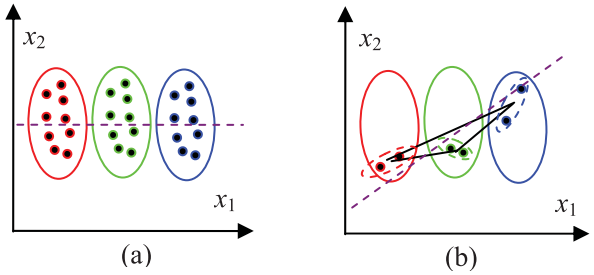

Fig. 1. LDA is sensitive to the number of samples per class. (a) Each class has $n = 9$ samples from which the correct discriminant direction (dashed-line) can be computed. (b) Samples are sampled from the same distribution as (a) but they are too sparse (i.e., $n = 2$). The dashed ellipses correspond to the wrong within-class scatter matrix computed from the sparse samples and the solid lines connect the means of each class. The resulting discriminant direction is denoted by the solid line.

In this paper, we develop a novel algorithm to enhance LDA when n is very small. Our algorithm is evaluated not only in face and ear recognition systems but can also be applied in other pattern recognition systems. LDA (i.e., Fisherface in face recognition community) makes use of between- and within-class scatter matrices for maximization of separability in a d-dimensional subspace with $d < D$. When the number n of training images per person is very small, there is a large bias in the computed between- and within-class scatter matrices against the underlying ones. This situation is shown in Fig. 1. In Fig. 1(a), $n = 9$ and $r = n/D = 9/2 = 4.5$ are large enough to reflect the underlying distribution (i.e., elliptical shape) and LDA is able to seek the optimal discriminant direction denoted by the dashed line. But, in Fig. 1(b) only $n = 2$ training samples are sampled from the big, solid ellipse. We can find from Fig. 1(b) that the samples are too sparse to describe the solid elliptical distributions and the underlying within-class scatter matrix. From the sparse samples, one can only merely estimate the wrong distributions shown by the dashed elliptical shapes and their corresponding within-class scatter matrix. The solid lines in Fig. 1(b) connect the means of the three classes. The triangles of the mean vectors together with the dashed ellipses determine a biased discriminant direction shown by the dashed line.

Fig. 1 shows that both the within-class and between-class scatter matrices are significantly biased from the correct ones when the number of training samples per class is very small. To deal with the small-sample-size problem in LDA, we, in this paper, propose to use unsupervised clustering to aug-

Manuscript received May 18, 2013; revised February 11, 2014; accepted February 12, 2014. Date of publication April 16, 2014; date of current version November 17, 2014. This work was supported in part by the State Key Program of National Natural Science of China under Grant 61232010, in part by the National Basic Research Program of China 973 Program under Grant 2014CB340404 and Grant 2014CB340403, in part by the National Natural Science Foundation of China under Grant 61172121, Grant 61172143, Grant 61271412, and Grant 61222109, in part by the Open Funding Project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant BUAA-VR-13KF, and in part by the Program for New Century Excellent Talents in University under Grant NCET-10-0620.

Y. Pang and S. Wang are with the School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: pyw@tju.edu.cn; shuangning07@hotmail.com).

Y. Yuan is with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: yuany@opt.ac.cn).

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/TNNLS.2014.2306844

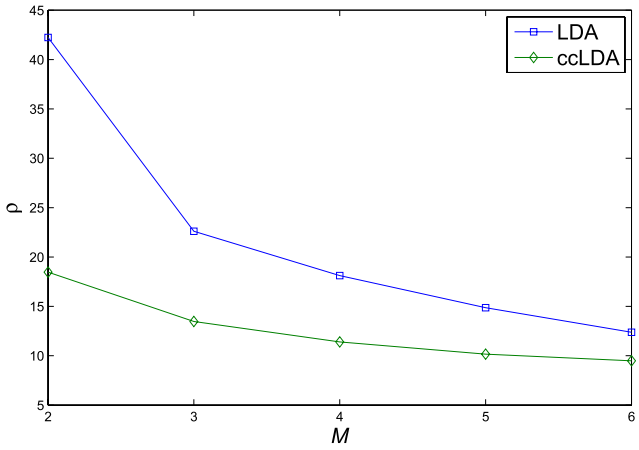

Fig. 2. Comparison of distance between $\mathbf{S}_w^{\text{ccLDA}}(M)$ and $\mathbf{S}_w^{\text{underlying}}$ and distance between $\mathbf{S}_w(M)$ and $\mathbf{S}_w^{\text{underlying}}$. Here, $\mathbf{S}_w^{\text{underlying}} \triangleq \mathbf{S}_w(7)$ and the AR face database is used.

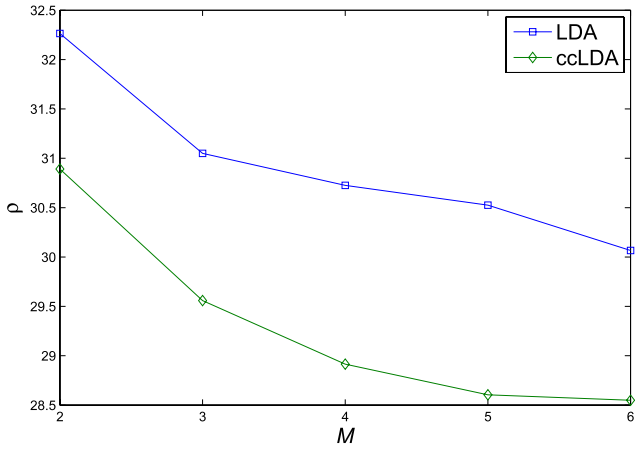

Fig. 3. Comparison of distance between $\mathbf{S}_b^{\text{ccLDA}}(M)$ and $\mathbf{S}_b^{\text{underlying}}$ and distance between $\mathbf{S}_b(M)$ and $\mathbf{S}_b^{\text{underlying}}$. Here, $\mathbf{S}_b^{\text{underlying}} \triangleq \mathbf{S}_b(7)$ and the AR face database is used.

ment both the within- and between-class scatter matrices for improving LDA. The motivation of the proposed method is as follows. A cluster consisting of similar samples of different classes contains some information of intraclass variations and the mean vectors of different clusters can distinguish different classes in some degree. Note that, the class labels are not used in the step of clustering because the clustering algorithm we employed is an unsupervised one. Fundamentally, our insight is that the between-cluster scatter matrix contributes to the underlying between- and the within-cluster scatter matrices contributes to the underlying within-class scatter matrix. Between- and within-cluster matrices are scatter matrices of clustered data. In computing between-cluster matrix, the total mean is the same as that in between-class scatter matrix while the component means are computed cluster by cluster instead of class by class. That is, the mean of each cluster is computed. The definitions of between- and within-cluster matrices are given in (13) and (14), respectively.

The goal of the proposed method is to use clustered data and corresponding between- and within-cluster scatter matrices to over the small-sample-size problem such as the one in Fig. 1(b). Using the proposed method, the dashed line in Fig. 1(b) is expected to turn toward the direction of the dashed line in Fig. 1(a). Specifically, the contributions and merits of this paper are summarized as follows.

1) We discover that the within-cluster scatter matrix can compensate the biased within-class scatter matrix computed from the limited number of samples per class even though a cluster consists of samples mainly from different classes (Fig. 2).
2) We discover that the between-cluster scatter matrix can compensate the biased between-class scatter matrix computed from the limited number of samples per class even though a cluster may contain samples of the same class (Fig. 3).
3) By augmenting the between-class scatter matrix with the between-cluster scatter matrix and augmenting the within-class scatter matrix with the within-cluster scatter matrix, we propose an enhanced version of LDA. Because both class-specific and cluster-specific information are used, we call the proposed LDA ccLDA where the first letter c means class and the next letter c stands for cluster.
4) The proposed method is suitable for the situation where there are a large number of classes but each class has a very small number of training samples. The advantages of the proposed method become more remarkable as the number of training samples per class decreases.
5) LDA is a special case of the proposed ccLDA method when the number of training samples per class is large. The proposed ccLDA can be regarded as a regularization version of LDA. Existing regularized LDA (RLDA) employs a regularization constant for regularizing while the proposed ccLDA uses the clusters for regularization, which contain much more useful information.

This paper is organized as follows. Section II reviews the related work. Section III describes two traditional LDA methods. Section IV presents the proposed ccLDA method. The experimental results are given in Section V. Finally, Section VI concludes this paper.

II. Related Work

Many variants of LDA have been proposed since it was successfully applied in face recognition [1], which is a typical sparse-sample pattern recognition problem [29], [28]. In this section, we briefly review the LDA techniques related to the proposed method.

We divide the related LDA techniques into the following four categories: generic LDA, RLDA, cluster-based LDA, and nonparametric LDA (NLDA).

A. Generic LDA

Generic LDA methods include classical PCA plus LDA (i.e., Fisherface in the community of face recognition) [2], marginal Fisher analysis (MFA) [33], Laplacianfaces [11], local Fisher discriminant analysis (LFDA) [27] and others [10], [19], [24], [32], [33]. Standard LDA (see Section III-A) is involved in inversion of within-class scatter matrix. However, it is common in pattern recognition that the

number of training samples is so small that the within-class scatter matrix is noninvertible and singular. PCA plus LDA conducts PCA prior to LDA to solve the singularity problem. MFA, an instance of graph embedding, simultaneously minimizes the intraclass compactness and maximizes the interclass separability. LFDA effectively combines the ideas of LPP and LDA, which attains between-class separation and within-class local structure preservation by defining a local within- and a local between-class scatter matrix.

B. Regularized LDA

RLDA modifies LDA by introducing a regularization term to scatter matrices [3], [6], [8], [14], [22]. If the within-class scatter matrix is singular or ill-conditioned, then the original LDA cannot work. RLDA can solve this problem by adding a diagonal matrix to the within-class scatter matrix. Lu *et al.* [18] proposed to replace the between-class scatter matrix with the combination of between- and within-class scatter matrices with a proper regularization parameter. Regularization technique can also be used in other subspace learning algorithms such as nonnegative matrix factorization [3].

C. Cluster-Based LDA

Cluster-based LDA is involved in data clustering. In LDA mixture model [15], complex data of many classes are partitioned into an appropriate number of clusters and LDA is applied to each cluster, independently. Li *et al.* [16] proposed an LDA-based clustering algorithm for unsupervised feature extraction. Its objective function is composed of within- and between-cluster scatter matrices. To deal with the single-sample problem in LDA, Pang *et al.* [21] proposed to substitute the within-class scatter matrix with the within-cluster scatter matrix with the between-class scatter matrix unchanged. In [34], spectral clustering method is used to identify the underlying group structure of the training data.

D. Nonparametric LDA

Unlike parametric LDA, which uses the parametric form of the between-class scatter matrix based on the Gaussian distribution assumption, NLDA is suitable for the non-Gaussian case due to its local property and nonparametric nature. Li *et al.* [17] extended two-class NDA [9] to multiclass cases where the discriminant information in both the principal space and the null space of the intraclass scatter matrix is used. Principal space is the subspace spanned by the eigenvectors corresponding to nonzero eigenvalues of the inversion of within-class scatter matrix multiplied by the between-class scatter matrix. Null space is the complement of the principal space and is defined by the subspace of eigenvectors corresponding to zero eigenvalues [31].

III. Traditional LDA

In this section, we describe standard LDA [2]. RLDA [9] is also discussed because it is closely related to the proposed method.

A. Standard LDA

LDA is a supervised linear dimensionality reduction method stemming from Fisher-Rao LDA [9].

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$ be the D-dimensional training data where N is the number of total training samples. The class label of the training sample $\mathbf{x}_i \in \mathbb{R}^D$ is denoted by $l(\mathbf{x}_i)$ with $l(\mathbf{x}_i) \in (1, \ldots, C)$ where C is the number of classes. Denote N_i the number of training samples of class i. The goal of linear dimensionality reduction is to find a proper transformation matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^{D\times d}$ so that $\mathbf{x} \in \mathbb{R}^D$ is mapped from the D-dimensional space to a low-dimensional space by

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \in \mathbb{R}^d \tag{1}$$

with $d < D$.

As a supervised dimensionality reduction method, LDA aims at finding the optimal transformation vector $\mathbf{w}$ that maximizes the Rayleigh coefficient

$$J_{\mathrm{LDA}}(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \tag{2}$$

where the between-class scatter matrix $\mathbf{S}_b$ and within-class scatter matrix $\mathbf{S}_w$ are defined, respectively, as

$$\mathbf{S}_b = \frac{1}{C}\sum_{i=1}^{C}(\mathbf{u}_i - \mathbf{u})(\mathbf{u}_i - \mathbf{u})^T \tag{3}$$

$$\mathbf{S}_w = \sum_{j=1}^{C}\sum_{l(\mathbf{x}_i)=j}(\mathbf{x}_i - \mathbf{u}_j)(\mathbf{x}_i - \mathbf{u}_j)^T. \tag{4}$$

In (3), $\mathbf{u}_i$ is the mean vector of samples of class i

$$\mathbf{u}_i = \frac{1}{N_i}\sum_{l(\mathbf{x}_j)=i}\mathbf{x}_j \tag{5}$$

and $\mathbf{u}$ is the mean vector of the whole data

$$\mathbf{u} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i. \tag{6}$$

It is proved that (2) can be reduced to the following eigenvalue decomposition problem:

$$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}. \tag{7}$$

The relationship between (2) and (7) is $\lambda = J(\mathbf{w})$.

B. Regularized LDA

If the within-class scatter matrix is invertible, then (7) can be expressed as the standard eigenvalue decomposition problem

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{w} = \lambda\mathbf{w}. \tag{8}$$

However, (8) cannot be solved if within-class scatter matrix is singular (invertible), which is true when the number of training samples is small (i.e., $N < C + D$). To solve such a small-sample-size problem, RLDA adds a multiple of identity matrix $\gamma\mathbf{I}$ to the within-class scatter matrix $\mathbf{S}_w$. It is noted that the regularization parameter γ is larger than zero.

The corresponding objective function and eigenvalue decomposition problem become

$$J_{\mathrm{RLDA}}(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T (\mathbf{S}_w + \gamma \mathbf{I})\mathbf{w}} \tag{9}$$

and

$$(\mathbf{S}_w + \gamma \mathbf{I})^{-1} \mathbf{S}_b \mathbf{w} = \lambda \mathbf{w} \tag{10}$$

respectively.

IV. Proposed Method: ccLDA

In this section, the disadvantages of standard LDA and RLDA are discussed. Then the proposed ccLDA method is presented and why ccLDA can overcome the disadvantages is explained.

A. ccLDA

Standard LDA in Section III-A works well when the number of training samples is large enough to estimate the underlying Gaussian distribution of the data [Fig. 1(a)]. However, if the number of training samples is small, then the training samples are unable to reflect the underlying distribution [Fig. 1(b)]. Therefore, conducting standard LDA on the small number of training samples is apt to overfitting. Moreover, when the number N of training samples is smaller than the number C of classes plus the dimension D (i.e., $N < C+D$), the within-class scatter matrix $\mathbf{S}_w$ is singular (invertible) and (8) cannot be solved. This phenomenon is called small-sample-size problem.

RLDA in Section III-B attempts to solve the small-sample-size problem by regularizing the within-class scatter matrix $\mathbf{S}_w$ with a scalar γ. But this scalar-based method does not introduce much useful information for classification. In this section, we propose a more advanced regularized method, which we call ccLDA. The first letter c of ccLDA means class and the next letter c stands for cluster. As the number of training samples per class decreases, the superiority of ccLDA over traditional methods becomes more remarkable.

The proposed ccLDA is motivated by our discovery that a cluster consisting of similar samples of different classes contains some information of intra class variations and the mean vectors of different clusters contribute to distinguishing different classes. The scatter matrices of the clusters are used to regularize the between- and within-class scatter matrices.

The first stage of our method is to group the training data $\mathbf{X}$ into K nonoverlapping clusters

$$\mathbf{X}^i = \mathbf{X}_1^i \cup \mathbf{X}_2^i \cup, \ldots, \cup \mathbf{X}_K^i, \quad i = 1, \ldots, P \tag{11}$$

$$\mathbf{X}_i^p \cap \mathbf{X}_j^p = \Phi \quad \forall i \neq j. \tag{12}$$

The cluster label of a training sample $\mathbf{x}_i \in \mathbb{R}^D$ is denoted by $c(\mathbf{x}_i)$ with $c(\mathbf{x}_i) \in (1, \ldots, K)$. The cluster number K is usually smaller than the class number C.

The second stage is to compute the so-called between-cluster scatter matrix $\mathbf{S}_b^i$ and within-cluster scatter matrix $\mathbf{S}_w^i$

$$\mathbf{S}_b^i = \frac{1}{K} \sum_{j=1}^{K} \left(\mathbf{v}_j^i - \mathbf{v}^i\right)\left(\mathbf{v}_j^i - \mathbf{v}^i\right)^T, \quad i = 1, \ldots, P \tag{13}$$

$$\mathbf{S}_w^i = \sum_{j=1}^{K} \sum_{c(\mathbf{x}_k)=j} \left(\mathbf{x}_k - \mathbf{v}_j^i\right)\left(\mathbf{x}_k - \mathbf{v}_j^i\right)^T, \quad i = 1, \ldots, P. \tag{14}$$

In (13) and (14), $\mathbf{v}_j^i$ is the mean vector of samples in $\mathbf{X}_j^i$ and $\mathbf{v}^i$ stands for the mean vector of $\mathbf{X}^i$ with $\mathbf{v}^i = \mathbf{v}^k = \mathbf{u}$, $i \neq k$. $\mathbf{u}$ is defined in (6). Similar to $\mathbf{S}_b$ and $\mathbf{S}_w$, $\mathbf{S}_b^i$ and $\mathbf{S}_w^i$ are also scatter matrices. But $\mathbf{S}_b^i$ and $\mathbf{S}_w^i$ are the scatter matrix of the clustered data whereas (3) is the one of the original data.

Because most of the clustering algorithms (K-means in our experiments) are significantly sensitive to initial randomly selected cluster centers. To overcome the drawback, the clustering algorithm is conducted P times with different initialization. Thus, different final clustering results are obtained. Corresponding to different initializations, there are P between- and within-cluster scatter matrices. ccLDA regularizes the between- and within-class scatter matrices with averaged between- and within-cluster matrices, respectively. Formally, the objective function $J_{\mathrm{ccLDA}}(\mathbf{w})$ of the proposed ccLDA is as follows:

$$J_{\mathrm{ccLDA}}(\mathbf{w}) = \frac{\mathbf{w}^T (\alpha \mathbf{S}_b + (1-\alpha)\frac{1}{P}\sum_{i=1}^{P} \mathbf{S}_b^i)\mathbf{w}}{\mathbf{w}^T (\beta \mathbf{S}_w + (1-\beta)\frac{1}{P}\sum_{i=1}^{P} \mathbf{S}_w^i)\mathbf{w}}. \tag{15}$$

In (15), there are two regularization parameters: $1 \geq \alpha \geq 0$ and $1 \geq \beta \geq 0$. The objective function is identical to that of standard LDA when $\alpha = 1$ and $\beta = 1$. So the proposed ccLDA is identical to standard LDA when $\alpha = 1$ and $\beta = 1$. That is, standard LDA is a special case of the proposed method. But when $\alpha \neq 1$ and $\beta \neq 1$, J_{ccLDA} can be viewed as J_{LDA} augmented by $\mathbf{S}_b^i$ and $\mathbf{S}_w^i$. It is easy to be proved that maximizing (15) is equal to solving the following eigenvalue-decomposition problem:

$$\left[\beta \mathbf{S}_w + (1-\beta)\frac{1}{P}\sum_{i=1}^{P} \mathbf{S}_w^i\right]^{-1} \left[\alpha \mathbf{S}_b + (1-\alpha)\frac{1}{P}\sum_{i=1}^{P} \mathbf{S}_b^i\right] = \lambda \mathbf{w}. \tag{16}$$

By defining

$$\mathbf{S}_w^{\mathrm{ccLDA}} \triangleq \beta \mathbf{S}_w + (1-\beta)\frac{1}{P}\sum_{i=1}^{P} \mathbf{S}_w^i \tag{17}$$

$$\mathbf{S}_b^{\mathrm{ccLDA}} \triangleq \alpha \mathbf{S}_b + (1-\alpha)\frac{1}{P}\sum_{i=1}^{P} \mathbf{S}_b^i. \tag{18}$$

Equation (16) can be written as

$$\left[\mathbf{S}_w^{\mathrm{ccLDA}}\right]^{-1} \mathbf{S}_b^{\mathrm{ccLDA}} = \lambda \mathbf{w}. \tag{19}$$

d-dimensional features are extracted based on (1).

The parameters α, β, and K are usually set empirically. But they can also be considered as functions of M (i.e., the number

of training samples per class), C (i.e., the number of classes), and D (i.e., the dimension of training samples)

$$\alpha = f(M, C, D) \tag{20}$$
$$\beta = g(M, C, D) \tag{21}$$
$$K = z(M, C, D). \tag{22}$$

More simply, α, β, and K can be expressed as functions of M

$$\alpha = f(M) \tag{23}$$
$$\beta = g(M) \tag{24}$$

and

$$K = z(M) \tag{25}$$

respectively.

Empirically, it is found that the following specific forms of (23) and (24) are proper for ccLDA:

$$\alpha = f(M) = 0.6 + 0.4\frac{M}{Q} \tag{26}$$
$$\beta = g(M) = 0.4 + 0.6\frac{M}{Q} \tag{27}$$

where Q is the number of training samples per class. M samples among the Q training samples are used for training. When M approaches 0, the values of α and β are equal to 0.6 and 0.4, respectively. When M approaches Q, both α and β become 1.

B. Why ccLDA Is Better

The usefulness of $\mathbf{S}_w^{\text{ccLDA}}$ and $\mathbf{S}_b^{\text{ccLDA}}$ can be clearly revealed if they are better than $\mathbf{S}_w$ and $\mathbf{S}_b$ in estimating the underlying within-class scatter matrix $\mathbf{S}_w^{\text{underlying}}$ and the underlying between-class scatter matrix $\mathbf{S}_b^{\text{underlying}}$. $\mathbf{S}_b$ and $\mathbf{S}_w$ are defined in (4) and (3), respectively.

Because $\mathbf{S}_w^{\text{ccLDA}}$, $\mathbf{S}_b^{\text{ccLDA}}$, $\mathbf{S}_w$, and $\mathbf{S}_b$ are dependent on the number M of training samples per class, we explicitly express these scatter matrices as $\mathbf{S}_w^{\text{ccLDA}}(M)$, $\mathbf{S}_b^{\text{ccLDA}}(M)$, $\mathbf{S}_w(M)$, and $\mathbf{S}_b(M)$. Without loss of generality, it is assumed that all the classes have the same number of training samples.

We have the following important observation which reveals why ccLDA is better than standard LDA and RLDA.

Observation 1: When the number M of the training samples becomes small, $\mathbf{S}_w^{\text{ccLDA}}(M)$ and $\mathbf{S}_b^{\text{ccLDA}}(M)$ are better than $\mathbf{S}_w(M)$ and $\mathbf{S}_b(M)$ in estimating the underlying within-class scatter matrix $\mathbf{S}_w^{\text{underlying}}$ and the underlying between-class scatter matrix $\mathbf{S}_b^{\text{underlying}}$.

To measure the distance between the scatter matrices, it is necessary to define a proper metric and the unknown underlying scatter matrices. The Observation 1 is to be verified by experimental results.

The underlying within-class and between-class scatter matrices are unknown and can be exactly computed only if M approaches infinity

$$\lim_{M\to\infty} \mathbf{S}_w(M) = \mathbf{S}_w^{\text{underlying}} \tag{28}$$
$$\lim_{M\to\infty} \mathbf{S}_b(M) = \mathbf{S}_b^{\text{underlying}}. \tag{29}$$

TABLE I
VALUES OF REGULARIZATION PARAMETERS IN OUR EXPERIMENTS

M	2	3	4	5	6	7
α	0.714	0.77	0.829	0.886	0.943	1
β	0.571	0.657	0.743	0.829	0.914	1

TABLE II
NUMBER OF CLUSTERS USED IN THE AR FACE DATABASE

M	2	3	4	5	6	7
K	5	8	12	8	5	1

It is impossible to have infinite number of training samples per class. But it is reasonable to estimate $\mathbf{S}_w^{\text{underlying}}$ and $\mathbf{S}_b^{\text{underlying}}$ by $\mathbf{S}_w(M)$ and $\mathbf{S}_b(M)$ with a relative large M, respectively. We conclude Observation 1 from the intermediate results on the AR [20] and Feret [25] face databases. The recognition rates of face recognition is given in Section V-A. Note that the intermediate results are used to verify Observation 1, which is the basis and motivation of the proposed ccLDA method. In the scenario of face recognition, only a small number of faces of one person are available [19]. In this case, $M = 2$ and $M = 3$ may be considered small, but $M \geqslant 5$ may be considered large. In the following experimental results, $\mathbf{S}_w^{\text{underlying}} \triangleq \mathbf{S}_w(7)$ and $\mathbf{S}_b^{\text{underlying}} \triangleq \mathbf{S}_b(7)$ are defined.

Because the covariance matrices lie on connected Riemannian manifold, common-used distance (e.g., Euclidean distance) is not suitable. In this paper, eigenvalue-based distance ρ is adopted [7], [23]

$$\rho(\mathbf{S}_1, \mathbf{S}_2) = \sqrt{\sum_{i=1}^{Q} \ln^2 \lambda_i(\mathbf{S}_1, \mathbf{S}_2)} \tag{30}$$

where $\lambda_1, \ldots, \lambda_Q$ are the Q largest generalized eigenvalues of the scatter matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$

$$\lambda_i \mathbf{S}_1 \mathbf{u}_i = \mathbf{S}_2 \mathbf{u}_i, \quad i = 1, \ldots, Q. \tag{31}$$

As stated above, to verify Observation 1, experiments on the AR face database are conducted. There are 117 persons. Each person has 7 face images among which M images are used for training and the remaining $7 - M$ images are used for testing. In the experiments, the regularization parameters α, β, and the number K of cluster are defined in (26), (27), and (34), respectively, and the K-means clustering algorithm runs $P = 25$ times. The parameters of the ccLDA are shown in Tables I and II. Fig. 2 shows the curves of $\rho(\mathbf{S}_w^{\text{ccLDA}}, \mathbf{S}_w^{\text{underlying}})$ and $\rho(\mathbf{S}_w, \mathbf{S}_w^{\text{underlying}})$ with $\mathbf{S}_w^{\text{underlying}} \triangleq \mathbf{S}_w(7)$. It is observed that

$$\rho(\mathbf{S}_w^{\text{ccLDA}}(M), \mathbf{S}_w^{\text{underlying}}) < \rho(\mathbf{S}_w(M), \mathbf{S}_w^{\text{underlying}}), \quad M = 2, \ldots, 6. \tag{32}$$

Fig. 3 shows the curves of $\rho(\mathbf{S}_b^{\text{ccLDA}}, \mathbf{S}_b^{\text{underlying}})$ and $\rho(\mathbf{S}_b, \mathbf{S}_b^{\text{underlying}})$ with $\mathbf{S}_b^{\text{underlying}} \triangleq \mathbf{S}_b(7)$. One can find from Fig. 3 that

$$\rho(\mathbf{S}_b^{\text{ccLDA}}(M), \mathbf{S}_b^{\text{underlying}}) < \rho(\mathbf{S}_b(M), \mathbf{S}_b^{\text{underlying}}), \quad M = 2, \ldots, 6. \tag{33}$$

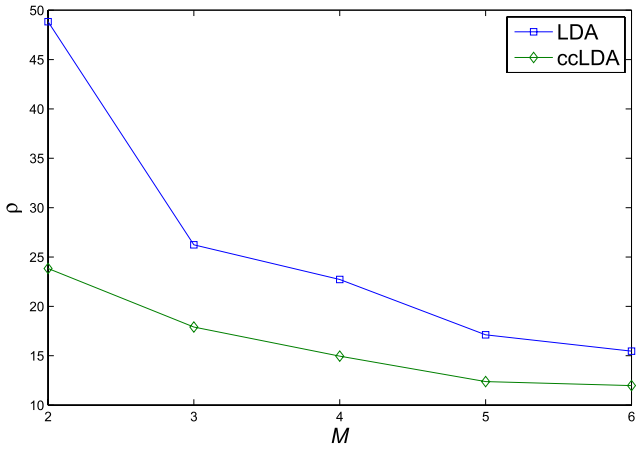

Fig. 4. Comparison of distance between $\mathbf{S}_w^{\text{ccLDA}}(M)$ and $\mathbf{S}_w^{\text{underlying}}$ and distance between $\mathbf{S}_w(M)$ and $\mathbf{S}_w^{\text{underlying}}$. Here, $\mathbf{S}_w^{\text{underlying}} \triangleq \mathbf{S}_w(7)$ and the Feret face database is used.

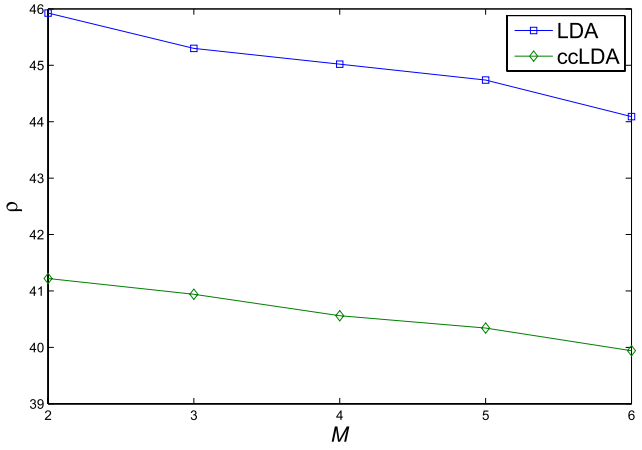

Fig. 5. Comparison of distance between $\mathbf{S}_b^{\text{ccLDA}}(M)$ and $\mathbf{S}_b^{\text{underlying}}$ and distance between $\mathbf{S}_b(M)$ and $\mathbf{S}_b^{\text{underlying}}$. Here, $\mathbf{S}_b^{\text{underlying}} \triangleq \mathbf{S}_b(7)$ and the Feret face database is used.

Figs. 2 and 3 experimentally support the Observation 1, which justifies the motivation of the proposed ccLDA method.

Figs. 4 and 5 show the experimental results on the Feret face database. The experimental setup is given in Section V. The figures have similar trends as those in Figs. 2 and 3. In the experimental results on both the AR and Feret face databases, the inequalities (32) and (33) hold. Therefore, it is considered that $\mathbf{S}_w^{\text{ccLDA}}(M)$ and $\mathbf{S}_b^{\text{ccLDA}}(M)$ are better than $\mathbf{S}_w(M)$ and $\mathbf{S}_b(M)$ in the sense of estimating the underlying within-class scatter matrix $\mathbf{S}_w^{\text{underlying}}$ and the underlying between-class scatter matrix $\mathbf{S}_b^{\text{underlying}}$.

V. Experimental Results

The experimental results in Section IV give the reason that our ccLDA method uses the average between-cluster scatter matrix $\frac{1}{P}\sum_{i=1}^{P}\mathbf{S}_b^i$ and within-cluster scatter matrix $\frac{1}{P}\sum_{i=1}^{P}\mathbf{S}_w^i$ for regularization in (15) and (16). In this section, we show the advantages of the proposed method over standard LDA and RLDA in terms of recognition rate. Experimental results are obtained from the AR face database [20], Feret face database [25], Indian face database [13], and Carreira–Perpinan ear data set [30].

Fig. 6. Seven normalized face images of one person in the AR face data set.

A. Experimental Results on the AR Face Data Set

In our experiments, 819 face images of 117 persons in the AR face dataset are used. Each person has seven nonoccluded face images among which M images are used for training and the rest $7-M$ images for testing. Each face image is normalized to 60×60 pixels. Seven normalized face images of one person are shown in Fig. 6. We preprocess the face images by applying PCA on the face images where 98% energy is reserved.

There are four parameters in (16) to be set: α, β, P, and K. The algorithm is not sensitive to the times P of clustering initialization as long as $P > 20$. In our experiments, P is set to 25. To show how the regularization parameters α and β influence the face recognition rate, we change α and β while the number d of extracted features (1) is fixed to 45. The recognition rates corresponding to the parameters are recorded and analyzed. We conclude that the functions in (26) and (27) are good for setting the values of α and β, respectively.

In (16), α and β reflect the weight of traditional between-class scatter matrix $\mathbf{S}_b$ and the weight of traditional within-class scatter matrix $\mathbf{S}_w$, respectively. Correspondingly, $(1-\alpha)$ and $(1-\beta)$ are the weights of average between-cluster scatter matrix $\mathbf{S}_b^{\text{ccLDA}}$ and within-cluster scatter matrix $\mathbf{S}_w^{\text{ccLDA}}$, respectively. From (26) and (27), we see that α and β linearly grow with M while $(1-\alpha)$ and $(1-\beta)$ are inversely proportional to M. Specifically, the values of α and β are shown in Table I.

The number K of cluster makes a great impact on classification accuracy. We establish a rule from the experimental results for determining K

$$K = z(M) = \lfloor 12 - 3.5 \times |M-4| \rfloor. \tag{34}$$

The values of K in our experimental results on the AR face database are shown in Table II. The recognition rates drop drastically if the values of K deviate much from those in Table II.

Fig. 7 shows the recognition rates of ccLDA, RLDA, and standard LDA when $M = 2$ face images per class are used for training while the remaining $7-M = 5$ face images are used for testing. It can be observed from Fig. 7 that the recognition rate of the proposed method is much higher than both standard LDA and RLDA. For example, when $d = 50$, ccLDA is 19% higher than LDA and 5% higher than RLDA. It is noted that the regularization parameter of RLDA is carefully selected so that it achieves its highest recognition rate.

Figs. 8–10 show the recognition rates of different methods when $M = 3$, 4, and 5, respectively. Investigating these

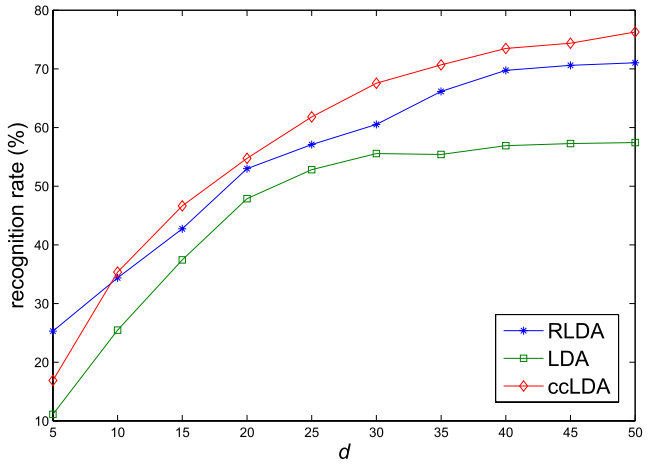

Fig. 7. Comparison of recognition rates on the AR face database when $M = 2$.

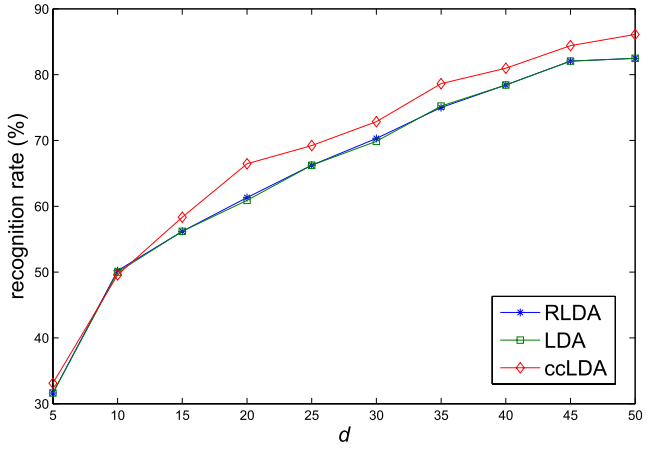

Fig. 8. Comparison of recognition rates on the AR face database when $M = 3$.

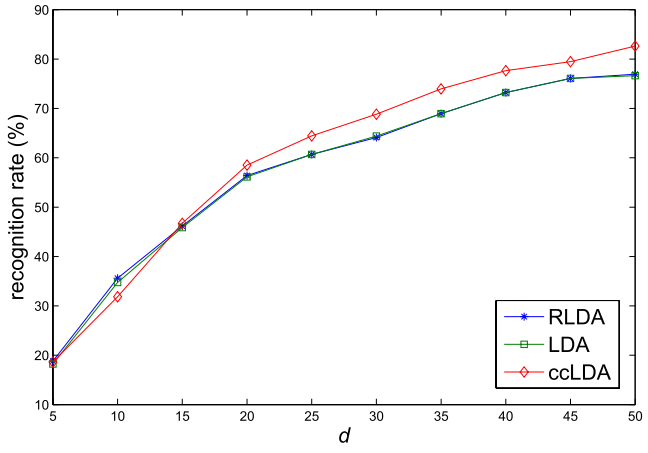

Fig. 9. Comparison of recognition rates on the AR face database when $M = 4$.

figures and Fig. 7, one can find that ccLDA becomes much better as M decreases and ccLDA approaches LDA as M increases. The experimental results demonstrate that the proposed method is suitable for the situation where only a small number of training samples are available.

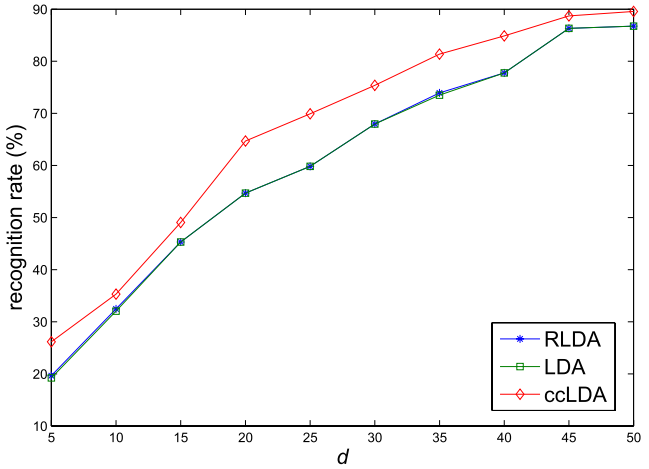

Fig. 10. Comparison of recognition rates on the AR face database when $M = 5$.

Fig. 11. Seven normalized face images of one person in the Feret face data set.

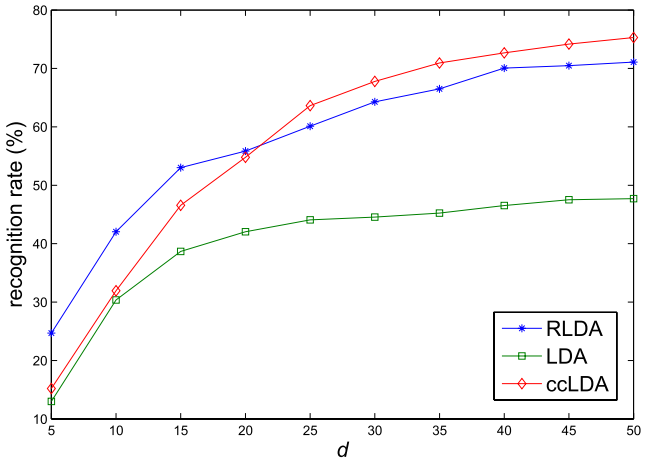

Fig. 12. Comparison of recognition rates on the Feret face database when $M = 2$.

B. Experimental Results on the Feret Face Data Set

Experiments are also conducted on the Feret face dataset. In our experiments, 1379 face images of 197 persons in the Feret dataset are used. In the experiments on the Feret face database, each person has seven nonoccluded face images among which M images are used for training and the rest $7-M$ images for testing. Each face image is normalized to 64×64 pixels. Seven normalized face images of one person are shown in Fig. 11. We also preprocess the face images by applying PCA on the face images where 98% energy is reserved.

The regularization parameters α, β, and K are the same as Tables I and II, which obey the functions in (26)–(34).

The recognition rates, varying with M, of ccLDA, RLDA, and LDA are shown in Figs. 12–15. The proposed method, ccLDA, outperforms significantly RLDA and LDA.

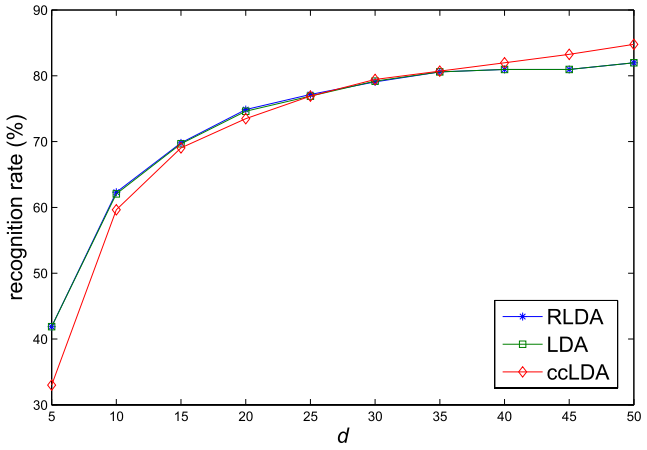

Fig. 13. Comparison of recognition rates on the Feret face database when $M = 3$.

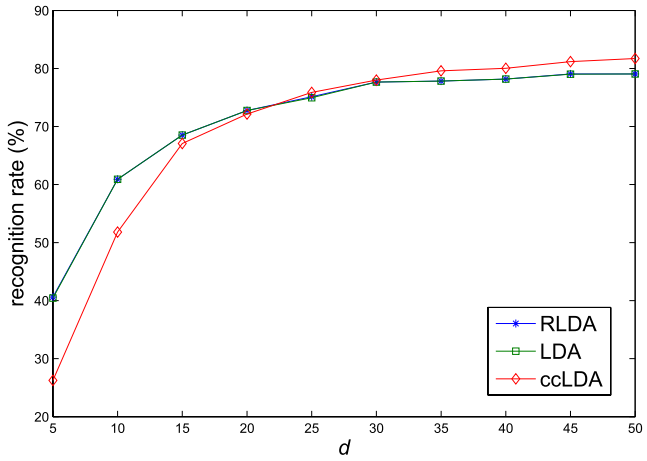

Fig. 14. Comparison of recognition rates on the Feret face database when $M = 4$.

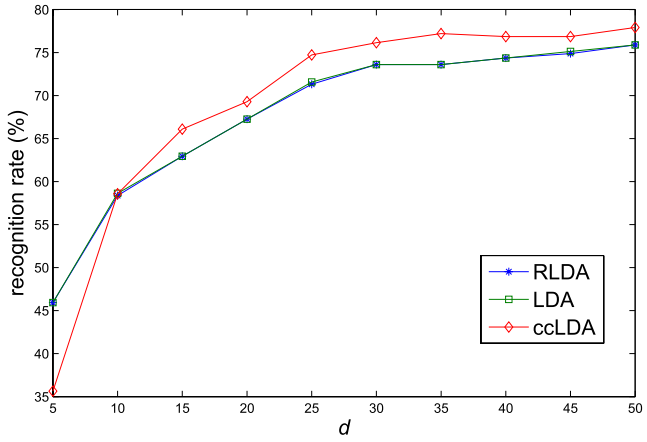

Fig. 15. Comparison of recognition rates on the Feret face database when $M = 5$.

For example, when $M = 2$ and $d = 50$, the recognition rate of ccLDA achieves 75.30% while the recognition rates of LDA and RLDA are merely 47.72% and 71.09%.

The experimental results in Figs. 12–15 clearly demonstrate the advantages of the proposed method. The advantages come from introducing between- and within-cluster scatter matrices.

Fig. 16. Eleven face images of one person in the Indian face data set.

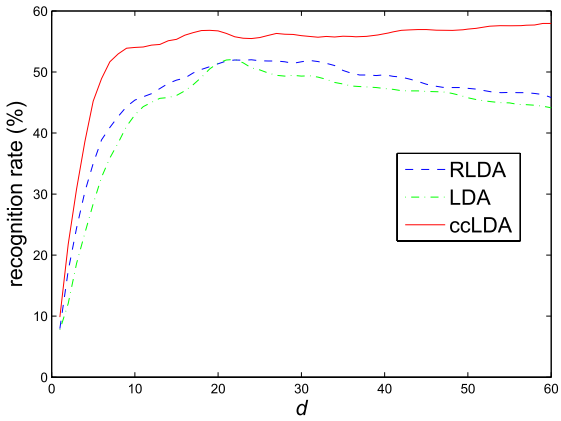

Fig. 17. Comparison of recognition rates on the Indian face database when $M = 2$.

The weights (i.e., $1 - \alpha$ and $1 - \beta$) of the cluster-based scatter matrices vary with the number of training face images. The weights are large when the number of available training samples is small. As M grows, the weights drop [see (23) and (24)].

The proposed ccLDA method can also be categorized to the regularized method. But RLDA regularizes the scatter matrix by a multiple of identity matrix $\gamma \mathbf{I}$, where γ is the regularization parameter. Instead, ccLDA uses cluster-based scatter matrices to regularize traditional class-based scatter matrices.

C. Experimental Results on the Indian Face Database

We have reported face recognition performance on faces of European and American. In this section, we conduct face recognition on faces of Asian. Specifically, the Indian face database is employed [13]. There are 22 females and 39 males in the database with 11 different poses for each individual. Among the 11 images of each individual (Fig. 16), M images are used for training and the rest $7 - M$ images for testing. The images are normalized according to the centers of the eyes. The face images are finally normalized to 32×32 pixels and then transformed to a low-dynamical subspace obtained by PCA where 99% energy is reserved.

The regularization parameters α, β, and K are the same as the ones in Section V-B. The experimental results of ccLDA, RLDA, and LDA are shown in Figs. 17–20. The recognition rates of the three methods grow with the number of features. Figs. 17 and 18 correspond to the results when two and three training samples per class are used, respectively. When two samples per class are used for training, ccLDA has much higher recognition rates than both RLDA and LDA. When three training samples per class are used, ccLDA is also

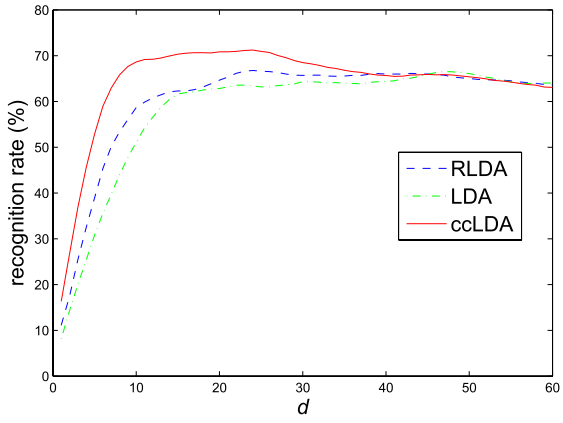

Fig. 18. Comparison of recognition rates on the Indian face database when $M = 3$.

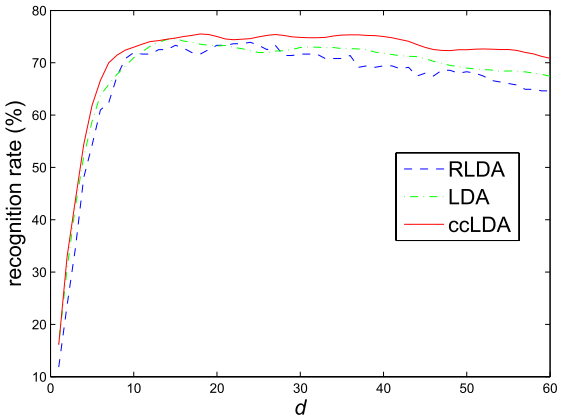

Fig. 19. Comparison of recognition rates on the Indian face database when $M = 4$.

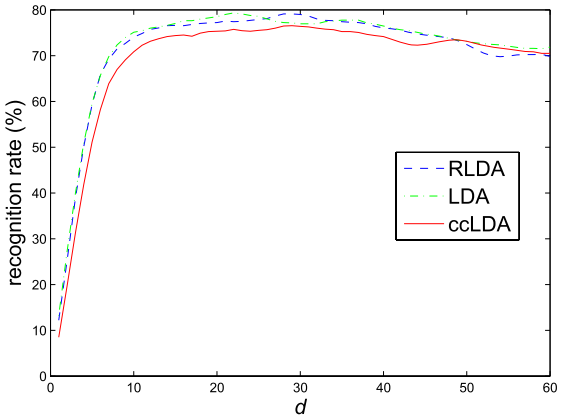

Fig. 20. Comparison of recognition rates on the Indian face database when $M = 5$.

the best; especially, a small number of features is employed. Fig. 19 shows that ccLDA is merely slightly better than LDA and RLDA when four training samples per class are used. When the number of training samples per class becomes five, the performance of ccLDA degrades significantly (Fig. 20).

The experimental results in Section V-A–C show that the proposed ccLDA algorithm is effective in face recognition where there are a large number of classes but each class has a very small number of training samples. In addition, the advantages of the proposed method become more remarkable as the number of training samples per class decreases.

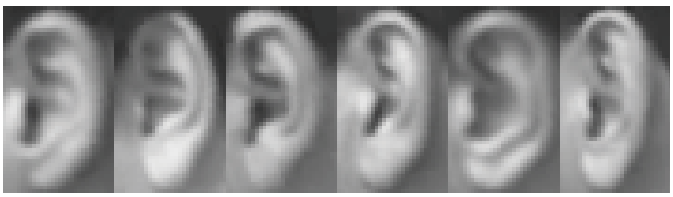

Fig. 21. Six ear images of the Carreira–Perpinan database.

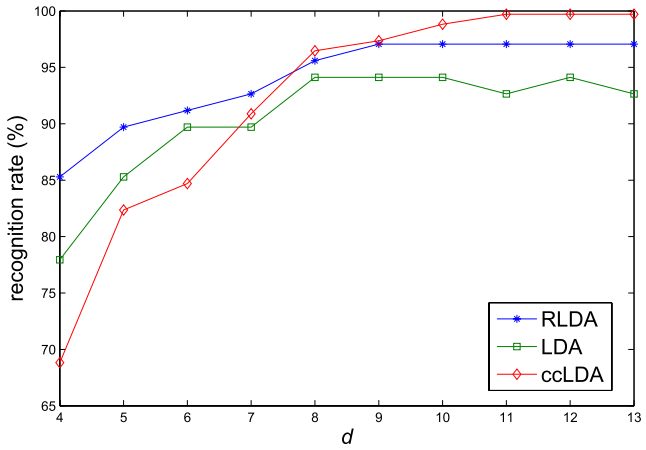

Fig. 22. Comparison of recognition rates on the Carreira–Perpinan ear database when $M = 2$.

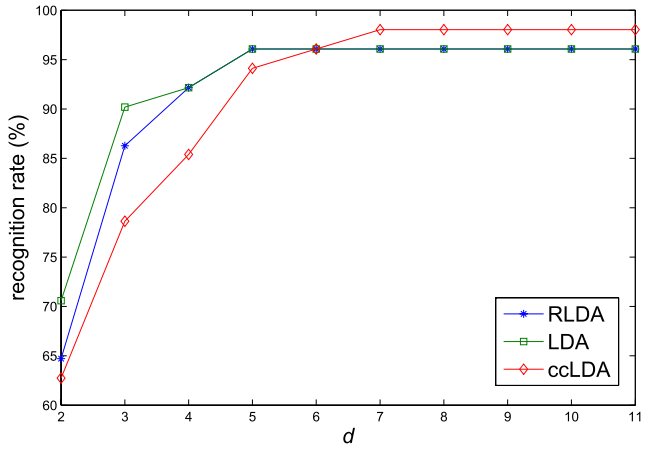

Fig. 23. Comparison of recognition rates on the Carreira–Perpinan ear database when $M = 3$.

D. Experimental Results on the Carreira–Perpinan Ear Data Set

Besides face, ear is also an effective biometric cue for person identification and recognition. In this section, experimental results on the Carreira–Perpinan ear database are given. The ear dataset of Carreira–Perpinan contains 17 subjects with six images for each subject [30]. Obviously, this is a small-sample-size problem in the task of ear recognition. The size of each image is 50×30. Fig. 21 shows six ear images.

The regularization parameters α and β are computed according to (26) and (27), respectively. The number K of clusters is selected according to (34). Figs. 22–24 show the recognition rates when $M = 2$, 3, and 4, respectively. Experimental

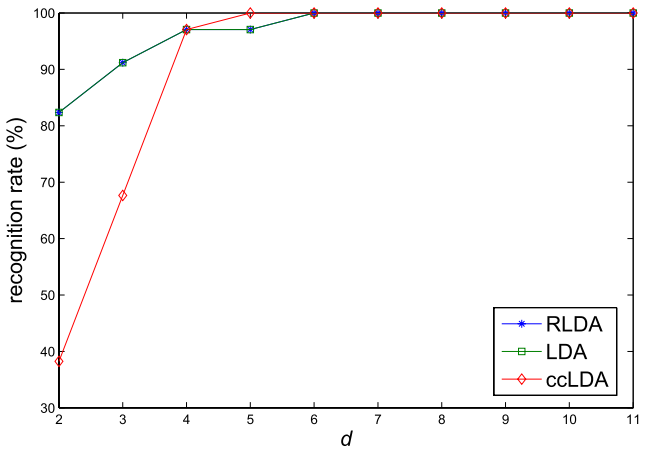

Fig. 24. Comparison of recognition rates on the Carreira–Perpinan ear database when $M = 4$.

results show that the proposed method exhibits remarkable performance in the situation where the number of training samples of each class is as small as 2 and the feature number $d > 9$ (Fig. 22). ccLDA also outperforms LDA and RLDA when $M = 3$ and $d > 6$ (Fig. 23). When the training number of each class is 4, ccLDA gets the same recognition rates as LDA and RLDA as long as the number d of features is larger than 6 (Fig. 24). But the recognition rates of ccLDA are lower than those of LDA and RLDA when the number of features is too small. Because the recognition rates of the three methods are all low and unacceptable when the feature number d is too small, high performance when d is reasonable large is important for practical pattern recognition system.

VI. Conclusion

In this paper, we have presented a novel method of RLDA called ccLDA. The proposed method regularizes both between- and within-class scatter matrices. The between-class scatter matrix is regularized by between- and the within-class scatter matrices is regularized by within-cluster scatter matrix. The regularized scatter matrices approach more closely to the underlying scatter matrices than the original scatter matrices do. We have also given empirical functions for setting the regularization parameters and cluster number. The experimental results on biometrics demonstrate that the proposed method is suitable for the situation where only a small number of training samples are available.

References

[1] S. An, W. Liu, S. Venkatesh, and H. Yan, "Unified formulation of linear discriminant analysis methods and optimal parameter selection" *Pattern Recognit.*, vol. 44, no. 2, pp. 307–319, 2011.

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

[3] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[4] D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *VLDB J.*, vol. 20, no. 1, pp. 21–33, 2011.

[5] L. Chen, I. W. Tsang, and D. Xu, "Laplacian embedded regression for scalable manifold regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 902–915, Jun. 2012.

[6] W. Ching, D. Chu, L. Liao, and X. Wang, "Regularized orthogonal linear discriminant analysis," *Pattern Recognit.*, vol. 45, no. 7, pp. 2719–2732, Jul. 2012.

[7] W. Forstner and B. Moonen, "A metric for covariance matrices," Dept. Geodesy Geoinform., Stuttgart Univ., Stuttgart, Germany, Tech. Rep., 1990.

[8] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[9] K. Fukunaga, *Statistical Pattern Recognition*. San Diego, CA, USA: Academic Press, 1990.

[10] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1087–1099, Jul. 2012.

[11] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[12] Y. Huang, D. Xu, and F. Nie, "Semi-supervised dimension reduction using trace ratio criterion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 519–526, Mar. 2012.

[13] V. Jain and A. Mukherjee. (2002). *The Indian Face Database* [Online]. Available: http://vis-www.cs.umass..edu/∼vidit/IndianFaceDatabase/

[14] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 383–394, Mar. 2008.

[15] H. Kim, D. Kim, and S. Y. Bang, "Face recognition using LDA mixture model," *Pattern Recognit. Lett.*, vol. 24, no. 15, pp. 2815–2821, Nov. 2003.

[16] C. Li, B. Kuo, and C. Lin, "LDA-based clustering algorithm and its application to an unsupervised feature extraction," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 1, pp. 152–163, Feb. 2011.

[17] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 755–761, Apr. 2009.

[18] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognit. Lett.*, vol. 26, pp. 181–191, Sep. 2005.

[19] A. M. Martinez and M. Zhu, "Where are linear feature extraction methods applicable," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1934–1944, Dec. 2005.

[20] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, May 2001.

[21] Y. Pang, J. Pan, and Z. Liu, "Cluster-based LDA for single sample problem in face recognition," in *Proc. Int. Conf. Mach. Learn. Cybern.*, 2005, pp. 4583–4587.

[22] Y. Pang, Z. Liu, and N. Yu, "A new nonlinear feature extraction method for face recognition," *Neurocomputing*, vol. 69, nos. 7–9, pp. 949–953, 2006.

[23] Y. Pang, Y. Yuan, and X. Li, "Gabor-based covariance matrices for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 989–993, Jul. 2008.

[24] C. H. Park and H. Park, "A comparison of generalized linear discriminant analysis algorithms," *Pattern Recognit.*, vol. 41, no. 3, pp. 1083–1097, 2008.

[25] P. J. Phillips, H. Moon, S. A. Rivzi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[26] R. Soares, H. Chen, and X. Yao, "Semisupervised classification with cluster regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 1779–1792, Nov. 2012.

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

[28] J. Tang, Z. Zha, D. Tao, and T. Chua, "Semantic-gap oriented active learning for multi-label image annotation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2354–2360, Apr. 2012.

[29] J. Tang, H. Li, G. Qi, and T. Chua, "Image annotation by graph-based inference with integrated multiple/single instance representations," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 131–141, Feb. 2010.

[30] C. Perpinan, "Compression neural networks for feature extraction: Application to human recognition from ear images," M.S. thesis, Dept. Phys., Tech. Univ. Madrid, Spain, 1995.

[31] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 1–6.

[32] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.
[33] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[34] J. Zhang, G. Sudre, X. Li, W. Wang, D. J. Weber, and A. Bagic, "Clustering linear discriminant analysis for MEG-Based brain computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 3, pp. 221–231, Jun. 2011.

Yanwei Pang (M'07–SM'09) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2004.

He is currently a Professor with the School of Electronic Information Engineering, Tianjin University, Tianjin, China. He has published more than 80 scientific papers, including 18 IEEE Transaction papers. His current research interests include object detection and recognition, and image processing.

Shuang Wang received the B.S. degree in electronic engineering from the Dalian University of Technology, Dalian, China, and the M.S. degree in signal processing from the Tianjin University, Tianjin, China, in 2011 and 2014, respectively.

Her current research interests include feature extraction and image analysis.

Yuan Yuan (M'05–SM'09) is a Researcher (Professor) with the Chinese Academy of Sciences, Beijing, China. Her current research interests include visual information processing and image/video content analysis.