Neurocomputing 87 (2012) 41–50

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Reconstructive discriminant analysis: A feature extraction method induced from linear regression classification

Yi Chen*, Zhong Jin

School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, PR China

ARTICLE INFO

Article history:
Received 2 June 2011
Received in revised form
15 January 2012
Accepted 5 February 2012
Communicated by Xiaofei He
Available online 23 February 2012

Keywords:
Reconstructive discriminant analysis
Feature extraction
Dimensionality reduction
Linear discriminant analysis
Linear regression classification
Face recognition
Finger Knuckle print recognition

ABSTRACT

Based on linear regression, a novel method called reconstructive discriminant analysis (RDA) is developed for feature extraction and dimensionality reduction (DR). RDA is induced from linear Regression classification (LRC). LRC assumes each class lies on a linear subspace and finds the nearest subspace for a given sample. But the original space cannot guarantee that the given sample matches its nearest subspace. RDA is designed to make the samples match their nearest subspaces. Concretely, RDA characterizes the intra-class reconstruction scatter as well as the inter-class reconstruction scatter, seeking to find the projections that simultaneously maximize the inter-class reconstruction scatter and minimize the intra-class reconstruction scatter. Actually, RDA can also be seen as another form of classical linear discriminant analysis (LDA) from the reconstructive view. The proposed method is applied to face and finger knuckle print recognition on the ORL, extended YALE-B, FERET face databases and the PolyU finger knuckle print database. The experimental results demonstrate the superiority of the proposed method.

1. Introduction

With the last several decades, dimensionality reduction (DR) has drawn considerable attention in the areas of image processing and pattern recognition. Generally, in practical applications, the raw data may contain variations of illumination and noises, which probably lead to misclassifications. And, it is time-consuming to perform classification directly in the high-dimensional space. For robust recognition and fast computation, DR techniques are usually performed first before the classification step. Although the DR step may cause information loss, recent literatures [2,24] indicate that the lost information has no substantial impact on the classification results. Even more, the researchers achieve much higher recognition rates in the reduced subspace [50,51]. As a fundamental problem in many scientific fields, DR plays an important role in scientific research. The goal of DR is to find a meaningful low dimensional representation of high dimensional data. With respect to pattern recognition, DR is an effective way to overcome the "curse of dimensionality" [1]. And more importantly, it reveals the distinctive features from the original data for pattern matching [2].

In the task of pattern recognition, discriminant analysis has shown its significant discriminability and becomes the fundamental tool in many areas. By far, numerous discriminant analysis methods have been developed. Among the proposed methods, the most well-known technique is linear discriminant analysis (LDA) [3]. Based on Euclidean distance, LDA searches for the project axes on which the inter-class data points are far away from each other while the intra-class data points are close to each other. Unfortunately, it has been pointed out that there are still some drawbacks existed in LDA. For example, (1) it usually suffers from the small sample size (SSS) problem [4] when the within-class scatter matrix is singular; (2) it is only optimal for the case where the distribution of the data in each class is a Gaussian with an identical covariance matrix [47]; (3) LDA can only extract at most $c-1$ features (c is the number of total classes), which is suboptimal for many applications. Numerous LDA variants [4–17,41–43] have been developed to solve the limitations mentioned above. Recently, motivated by manifold learning algorithms [18–20], researchers proposed a family of locality characterization based discriminant analysis techniques [21–26,34]. Different from LDA, these techniques extract local discriminative information. Despite the different motivations of these algorithms, they can be nicely interpreted in a general graph embedding framework [19,22,26]. The graph embedding view of subspace learning provides us a powerful platform to develop various kinds of dimensionality reduction algorithms. However, the high computational cost restricts these algorithms to be applied to large scale high dimensional data sets. To address this issue, a strong tool named spectral regression (SR) [52–56] was proposed for efficient subspace learning.

* Corresponding author. Tel.: +86 25 8431 7297x403 (lab).
E-mail address: cystory@qq.com (Y. Chen).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2012.02.001

Although the existing discriminant analysis techniques achieve remarkable performances, we notice that these methods were designed independently of classifiers. At the classification stage of the pattern recognition progress, the classifier is usually selected by experience. Obviously, the subspaces learned by different discriminant analysis methods have different characteristics that are invisible to the classifiers. However, one specific classifier just explores the subspace following the classification rule rather than the characteristic of the subspace. Therefore, the DR method may not match the random selected classifier perfectly, which potentially degrades the performance of the pattern recognition system. To connect DR methods with classifiers, one feasible way is to design the DR methods according to the classification rule of a specific classifier. In literatures, we find Yang et al. have designed discriminant analysis methods [28,29] according to the minimal local reconstruction error (MLRE) measure based classifier and the local mean based nearest neighbor classifier (LM-NNC) respectively. By combining the discriminant analysis methods with their corresponding optimal classifiers, the researchers demonstrated remarkable improvements against conventional discriminant analysis methods.

Very recently, an important work called linear regression classification (LRC) [27] is reported by Naseem et al., where linear regression is applied to estimate the reconstruction error. Then the label of the probe image will be assigned as the class with a minimum reconstruction error. In Naseem et al.'s pioneer work, the downsampled images are directly used for classification. However, neither the original space nor the downsampled image space can guarantee that the intra-class reconstruction error is minimal. To obtain the best performance, the original space should have smaller intra-class reconstruction errors and larger inter-class reconstruction errors. Due to the variations of illumination and noises, the inter-class reconstruction error is probably smaller than the intra-class reconstruction error in the original space. Under this circumstance, the performance of LRC will degrade. In order to strengthen the performance of LRC, we first inherit the assumption and the classification rule of LRC. Based on the inherited assumption and classification rule, we aim to find a subspace that has smaller intra-class reconstruction errors and larger inter-class reconstruction errors. Then we present a new method called reconstructive discriminant analysis (RDA) for feature extraction and DR.

To have an intuitive impression, we show the characteristics of RDA, LDA and the MLRE-based feature extractor (MLREF). Based on Euclidean distance, LDA searches for the directions that are most discriminative to separate the samples belonging to different classes. Different from LDA, MLREF and RDA are representation-based methods. MLREF finds the projections on which samples can be best represented by their local intra-class neighbors. Motivated by the classification rule of LRC, RDA finds the projections on which samples can be best expressed by all of their intra-class samples.

The rest of the paper is organized as follows. Related works are reviewed in Section 2. In Section 3, RDA is described in detail. Connections with some related works are analyzed in Section 4. In Section 5, the experiments are presented on the well-known databases to demonstrate the effectiveness of the proposed method. Finally, conclusions are drawn in Section 6.

2. Related works

2.1. Linear discriminant analysis (LDA)

Assume we have n samples from c classes. Let n_i represents the training number of the ith class and $\mathbf{x}_i^j \in R^d$ denotes the jth sample of the ith class, $i=1,2,\ldots,c$, $j=1,2,\ldots,n_i$. The objective function of LDA is as follows:

$$\mathbf{w}_{opt} = \arg\max_{\mathbf{w}} \frac{|\mathbf{w}^T\mathbf{S}_b\mathbf{w}|}{|\mathbf{w}^T\mathbf{S}_w\mathbf{w}|} \tag{1}$$

where

$$\mathbf{S}_b = \sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T \tag{2}$$

$$\mathbf{S}_w = \sum_{i=1}^{c}\left(\sum_{j=1}^{n_i}(\mathbf{x}_i^j-\mathbf{m}_i)(\mathbf{x}_i^j-\mathbf{m}_i)^T\right) \tag{3}$$

$\mathbf{m}_i$ is the average vector of the ith class, and $\mathbf{m}$ is the average vector of all samples. The optimal projections are the generalized eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ corresponding to the largest generalized eigenvalues.

2.2. Linear regression classification (LRC)

LRC is based on the assumption that samples from a specific object class lie on a linear subspace. Using this concept, a linear model is developed. In this model, a probe image is represented as a linear combination of class-specific samples. Thereby the task of recognition is defined as a problem of linear regression. Least-squares estimation (LSE) [31–33] is used to estimate the reconstruction coefficients for a given probe image against all class models. Finally, the label is signed as the class with the most precise estimation.

Assume $\mathbf{X}_i$ is a class-specific model generated by stacking the n-dimensional image vectors

$$\mathbf{X}_i = [\mathbf{x}_i^1,\mathbf{x}_i^2,\ldots,\mathbf{x}_i^{n_i}] \in R^{n\times n_i}, \quad i=1,2,\ldots,c \tag{4}$$

Suppose $\mathbf{y}$ is a probe sample from the ith class, it should be represented as a linear combination of the images from the same class (lying on the same subspace), i.e.,

$$\mathbf{y} = \mathbf{X}_i\boldsymbol{\beta}_i, \quad i=1,2,\ldots,c \tag{5}$$

where $\boldsymbol{\beta}_i \in R^{n_i\times 1}$ is the reconstruction coefficients. Given that $n \geq n_i$, the system of equations in Eq. (5) is well conditioned and can be estimated by LSE:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{y} \tag{6}$$

The probe sample can be reconstructed by Eq. (7):

$$\begin{aligned}\hat{\mathbf{y}}_i &= \mathbf{X}_i\hat{\boldsymbol{\beta}}_i, \quad i=1,2,\ldots,c\\ &= \mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{y}\end{aligned} \tag{7}$$

Then the distance measure between the probe sample $\mathbf{y}$ and reconstructed sample $\hat{\mathbf{y}}_i$, $i=1,2,\ldots,c$ can be computed, and the label is signed as the class with the minimum distance, i.e.,

$$\min_i \|\mathbf{y}-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i\|^2, \quad i=1,2,\ldots,c \tag{8}$$

2.3. Minimal local reconstruction error measure based discriminant feature extraction

MLREF [28] is induced from the MLRE measure based Classifier (MLREC).The MLRE-based feature extractor aims to find the projections $\mathbf{P}$ that maximize the following criterion:

$$J(\mathbf{P}) = \frac{\mathrm{tr}(\mathbf{P}^T\mathbf{S}_b^L\mathbf{P})}{\mathrm{tr}(\mathbf{P}^T\mathbf{S}_w^L\mathbf{P})} \tag{9}$$

where

$$\mathbf{S}_w^{\mathrm{L}} = \sum_{\substack{i,j\\ s=i}}\left(\mathbf{x}_i^j-\sum_t w_{st}^{ij}\mathbf{x}_s^t\right)\left(\mathbf{x}_i^j-\sum_t w_{st}^{ij}\mathbf{x}_s^t\right)^T \tag{10}$$

$$\mathbf{S}_b^{\mathrm{L}} = \sum_{ij}\sum_{s\neq i}\left(\mathbf{x}_i^j - \sum_t w_{st}^{ij}\mathbf{x}_s^t\right)\left(\mathbf{x}_i^j - \sum_t w_{st}^{ij}\mathbf{x}_s^t\right)^T \tag{11}$$

and w_{st}^{ij} is the reconstruction coefficient which can be obtained by solving the following optimization problem:

$$\min_i \left\|\mathbf{x}_i^j - \sum_t w_{st}^{ij}\mathbf{x}_s^t\right\|^2 \tag{12}$$

subject to $\sum_t w_{st}^{ij} = 1$ and $w_{st}^{ij} = 0$ if $\mathbf{x}_s^t$ does not belong to the set of k-nearest neighbors of $\mathbf{x}_i^j$ in Class s. The optimal projections are the generalized eigenvectors of $(\mathbf{S}_w^{\mathrm{L}})^{-1}\mathbf{S}_b$ corresponding to the largest generalized eigenvalues.

3. Reconstructive discriminant analysis

3.1. Basic idea

As we mentioned above, our method is induced from LRC. LRC finds the class with a minimum reconstruction error (or equivalently, the nearest subspace) for a given sample. To obtain the optimal performance, for each sample, the intra-class reconstruction error should be smaller than the inter-class reconstruction error. Unfortunately, in the original space, the intra-class reconstruction error is probably larger than the inter-class reconstruction error due to the illuminations and noises. To achieve a better performance, we need to inherit the linear assumption [4,30] first. Therefore RDA has the same reconstruction strategy as LRC. Then we aim to find a subspace that minimizes the reconstruction error of the intra-class samples and maximizes the reconstruction error of inter-class samples simultaneously.

According to the linear subspace assumption, a probe image can be represented as a linear combination of the training images from the same class as shown in Eq. (5). The intra-class reconstruction error can be computed as follows:

$$\varepsilon_{ij} = \|\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j\|^2 \tag{13}$$

where $\boldsymbol{\beta}_i^j$ is the optimal reconstruction weights obtained by Eq. (6). Geometrically, to minimize the reconstruction error in Eq. (13) is to find a point $\hat{\mathbf{x}}_i^j$ on the subspace spanned by the samples from the ith class that is closest to $\mathbf{x}_i^j$. The intra-class reconstruction error is actually the distance to its own class. Apparently, in a discriminative space, this distance should be as small as possible. Based on the intra-class reconstruction error, we can define the intra-class reconstruction scatter of samples in the original space

$$\sum_i\sum_j \varepsilon_{ij} = \sum_i\sum_j \|\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j\|^2 \tag{14}$$

Similarly, the inter-class reconstruction error by the pth ($p\neq i$) class is

$$\varepsilon_{ij}^p = \|\mathbf{x}_i^j - \mathbf{X}_p\boldsymbol{\beta}_p^j\|^2 \tag{15}$$

To strengthen the separability of the different classes, a sample should be far away from its nearest subspaces spanned by other classes. Based on Eq. (15), for a given sample, we can find its k heterogeneous nearest subspaces (the $k$classes with the least inter-class reconstruction errors). Assume the subspace spanned by $\mathbf{X}_m$ ($m\neq i$) is one of the nearest subspaces of $\mathbf{x}_i^j$. Then the inter-class reconstruction scatter of the k nearest subspaces in the original space is defined as follows:

$$\sum_i\sum_j\sum_m \varepsilon_{ij}^m = \sum_i\sum_j\sum_m \|\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j\|^2 \tag{16}$$

From the geometrical view, the intra-class reconstruction scatter characterizes the compactness of the intra-class samples and the inter-class reconstruction scatter characterizes the separability of different classes. Naturally, a sample should be close to the subspace spanned by its intra-class samples. Simultaneously, it should be far away from the subspaces that other classes lie on. Therefore, it is obvious that larger inter-class reconstruction scatters and smaller intra-class reconstruction scatters will lead to better classification results.

We begin with LRC and use it as a steerer to induce a discriminant analysis method RDA. Furthermore, RDA and LRC share the same assumption and the same reconstruction strategy. We can expect that the proposed method is optimal for LRC. It is worthwhile to point out that some other classifiers such as sparse representation-based classification (SRC) [35] have the same classification rule as LRC. But they have different assumptions and reconstruction strategies. So RDA and SRC may not match perfectly.

Based on the above analysis, the motivation of RDA can be explained intuitively. RDA minimizes the intra-class reconstruction error and maximizes the inter-class reconstruction error at the same time. Geometrically, it pulls the samples to their own subspace and pushes the samples away from other subspaces. In other words, RDA aims to match the samples and their nearest subspaces. We know that, for each sample, LRC finds the nearest subspace rather than the nearest neighbor. LRC works effectively when the samples match their nearest subspaces. We can image that LRC should be more effective in the RDA subspace.

3.2. Fundamentals

The goal of RDA is to find the low-dimensional subspace into which the intra-class reconstruction scatter is minimized while at the same time the inter-class reconstruction scatter is maximized. Suppose we have obtained the optimal projections $\mathbf{P}=\{\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\ldots,\boldsymbol{\varphi}_d\}$ on which samples can be best represented by all of their intra-class samples. Project each data point $\mathbf{x}_i^j$ to the subspace:

$$\mathbf{y}_i^j = \mathbf{P}^T\mathbf{x}_i^j \tag{17}$$

The intra-class reconstruction scatter of samples in the subspace is

$$\begin{aligned}\sum_i\sum_j \|\mathbf{y}_i^j - \mathbf{Y}_i\boldsymbol{\beta}_i^j\|^2 &= \sum_i\sum_j (\mathbf{y}_i^j - \mathbf{Y}_i\boldsymbol{\beta}_i^j)^T(\mathbf{y}_i^j - \mathbf{Y}_i\boldsymbol{\beta}_i^j)\\ &= \sum_i\sum_j (\mathbf{P}^T\mathbf{x}_i^j - \mathbf{P}^T\mathbf{X}_i\boldsymbol{\beta}_i^j)^T(\mathbf{P}^T\mathbf{x}_i^j - \mathbf{P}^T\mathbf{X}_i\boldsymbol{\beta}_i^j)\\ &= \mathrm{tr}\left(\sum_i\sum_j [\mathbf{P}^T(\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)][\mathbf{P}^T(\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)]^T\right)\\ &= \mathrm{tr}\left(\mathbf{P}^T\left[\sum_i\sum_j (\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)(\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)^T\right]\mathbf{P}\right)\\ &= \mathrm{tr}(\mathbf{P}^T\mathbf{S}_w^R\mathbf{P})\end{aligned} \tag{18}$$

where $\mathrm{tr}(\cdot)$ is the notation of trace operator, and

$$\mathbf{S}_w^R = \sum_i\sum_j (\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)(\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)^T \tag{19}$$

is called the intra-class reconstruction scatter matrix. It is easy to prove that $\mathbf{S}_w^R$ is a nonnegative definite matrix.

The inter-class reconstruction scatter of samples in the subspace is

$$\begin{aligned}&\sum_i\sum_j\sum_m \|\mathbf{y}_i^j - \mathbf{Y}_m\boldsymbol{\beta}_m^j\|^2\\ &= \sum_i\sum_j\sum_m (\mathbf{y}_i^j - \mathbf{Y}_m\boldsymbol{\beta}_m^j)^T(\mathbf{y}_i^j - \mathbf{Y}_m\boldsymbol{\beta}_m^j)\\ &= \sum_i\sum_j\sum_m (\mathbf{P}^T\mathbf{x}_i^j - \mathbf{P}^T\mathbf{Y}_m\boldsymbol{\beta}_m^j)^T(\mathbf{P}^T\mathbf{x}_i^j - \mathbf{P}^T\mathbf{Y}_m\boldsymbol{\beta}_m^j)\end{aligned}$$

$$= \sum_i \sum_j \sum_m [\mathbf{P}^T(\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)][\mathbf{P}^T(\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)]$$

$$= \mathrm{tr}\left(\sum_i \sum_j \sum_m [\mathbf{P}^T(\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)][\mathbf{P}^T(\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)]^T\right)$$

$$= \mathrm{tr}\left(\mathbf{P}^T\left[\sum_i \sum_j \sum_m (\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)(\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)^T\right]\mathbf{P}\right)$$

$$= \mathrm{tr}(\mathbf{P}^T\mathbf{S}_b^R\mathbf{P}) \tag{20}$$

where

$$\mathbf{S}_b^R = \sum_i \sum_j \sum_m (\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)(\mathbf{x}_i^j - \mathbf{X}_m\boldsymbol{\beta}_m^j)^T \tag{21}$$

is called the inter-class reconstruction scatter matrix.

To maximize the inter-class reconstruction scatter and minimize the intra-class reconstruction scatter simultaneously, we tend to find the projections that maximize the following criterion:

$$J(\mathbf{P}) = \frac{\mathrm{tr}(\mathbf{P}^T\mathbf{S}_b^R\mathbf{P})}{\mathrm{tr}(\mathbf{P}^T\mathbf{S}_w^R\mathbf{P})} \tag{22}$$

In a special case, when $\mathbf{P}$ is one-dimensional vector, i.e., $\mathbf{P}=\boldsymbol{\varphi}$, then the criterion changes to

$$J(\boldsymbol{\varphi}) = \frac{\boldsymbol{\varphi}^T\mathbf{S}_b^R\boldsymbol{\varphi}}{\boldsymbol{\varphi}^T\mathbf{S}_w^R\boldsymbol{\varphi}} \tag{23}$$

We can find the optimal solutions $\mathbf{P}=\{\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\ldots,\boldsymbol{\varphi}_d\}$ by solving the d generalized eigenvectors of Eq. (24) corresponding to d largest eigenvalues

$$\mathbf{S}_b^R\boldsymbol{\varphi} = \lambda\mathbf{S}_w^R\boldsymbol{\varphi} \tag{24}$$

3.3. Implementation of RDA in small sample size cases

If the dimension of the original subspace is larger than the total number of training samples, $\mathbf{S}_w^R$ is always singular because the following proposition holds:

Proposition 1. *The rank of the intra-class reconstruction scatter matrix $\mathbf{S}_w^R$ is equal or less than n, i.e., $rank(\mathbf{S}_w^R) \le n$, where n is the total number of training samples.*

Proof. First of all, let us define the intra-class reconstruction coefficient matrix:

$$\boldsymbol{\beta}_i = [\boldsymbol{\beta}_i^1, \boldsymbol{\beta}_i^2, \ldots, \boldsymbol{\beta}_i^{n_i}] \tag{25}$$

Based on the intra-class reconstruction coefficient matrix in Eq. (25), we can further define the global reconstruction coefficient matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 & \cdots & \cdots & 0 \\ \vdots & \boldsymbol{\beta}_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \boldsymbol{\beta}_c \end{bmatrix} \tag{26}$$

Suppose $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_c]$ is the column sample matrix. According to Eq. (26), we can rewrite the intra-class reconstruction scatter matrix as follows:

$$\mathbf{S}_w^R = \sum_i \sum_j (\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)(\mathbf{x}_i^j - \mathbf{X}_i\boldsymbol{\beta}_i^j)^T = (\mathbf{X}-\mathbf{X}\boldsymbol{\beta})(\mathbf{X}-\mathbf{X}\boldsymbol{\beta})^T$$

$$= \mathbf{X}(\mathbf{I}-\boldsymbol{\beta})(\mathbf{I}-\boldsymbol{\beta})^T\mathbf{X}^T = \mathbf{X}\mathbf{M}\mathbf{X}^T \tag{27}$$

where $\mathbf{M}=(\mathbf{I}-\boldsymbol{\beta})(\mathbf{I}-\boldsymbol{\beta})^T$ and $\mathbf{I}$ is the identity matrix.

From the definitions, we know that $\mathbf{X}\in R^{d\times n}$ and $\mathbf{M}\in R^{n\times n}$, where d is the dimension of the image vector. It is easy to derive that $rank(\mathbf{X}) = rank(\mathbf{X}^T) \le \min(d,n)$ and $rank(\mathbf{M}) \le n$.

In practical applications, the image vector is very high-dimensional, i.e., $d \gg n$. Thus $rank(\mathbf{X}) \le n$ and $rank(\mathbf{M}) \le n$. Then we have $rank(\mathbf{S}_w^R) \le \min(rank(\mathbf{X}), rank(\mathbf{M})) \le n$. □

In a high-dimensional space, the sample images are generally linear independent, i.e., $rank(\mathbf{X})=n$. Meanwhile, $\boldsymbol{\beta}_i^j$ is composed of irregular decimals. In this case, the intra-class reconstruction coefficient vectors are generally linear independent too, i.e., $rank(\boldsymbol{\beta}_i)=n_i$. When $rank(\boldsymbol{\beta}_i)=n_i$, it is easy to prove $rank(\boldsymbol{\beta})=n$. We notice that $\mathbf{I}-\boldsymbol{\beta}$ only changes the diagonal entries of $\boldsymbol{\beta}$. According to the definition of $\boldsymbol{\beta}$, the diagonal block entries of $\boldsymbol{\beta}$ are the intra-class reconstruction coefficient matrix $\boldsymbol{\beta}_i$ which contains irregular decimals. Usually, the matrix $\mathbf{I}-\boldsymbol{\beta}$ is of full rank, i.e., $rank(\mathbf{I}-\boldsymbol{\beta})=n$. Then we have $rank(\mathbf{M})=rank((\mathbf{I}-\boldsymbol{\beta})(\mathbf{I}-\boldsymbol{\beta})^T)=n$. Finally, we derive $rank(\mathbf{S}_w^R) \le \min(rank(\mathbf{X}), rank(\mathbf{M})) = n$. From the above analysis, we find that $rank(\mathbf{S}_w^R) = n$ generally holds in a high-dimensional space.

$\mathbf{S}_w^R$ is a d by d matrix but $rank(\mathbf{S}_w^R) \le n$. When the total number of the train samples is smaller than the dimension of the image vectors, $\mathbf{S}_w^R$ is singular. The inverse of $\mathbf{S}_w^R$ can not be calculated directly. To overcome the singularity problems, we first apply principal component analysis (PCA) [36] to reduce the dimension of the original subspace.

At the same time, the dimension of the PCA subspace can not be too small. Let us review the step of solving reconstruction coefficients in Eq. (6). If the number of training sample in each class exceeds the dimension of the PCA subspace, $\mathbf{X}_i^T\mathbf{X}_i$ will be singular too. In this case, the reconstruction coefficients can not be computed directly neither.

In conclusion, to avoid the singularity problems, we perform RDA based on the PCA-transformed features. The dimension of the PCA-transformed features should be larger than the number of training sample in each class and smaller than the total number of the training samples.

3.4. The algorithm of RDA

The main steps of the algorithm are summarized in Table 1.

4. Further discussions

4.1. Connections with LDA

While LDA aims to find the projections that maximize the inter-class scatter and simultaneously minimize the intra-class scatter, the proposed method seeks to find projections that maximize the inter-class reconstruction scatter and simultaneously minimize the intra-class reconstruction scatter. Obviously, the differences between RDA and LDA are the definitions of the scatters.

LDA tends to keep samples from the same class as near as possible and separate the different classes as far as possible. To achieve this goal, each sample is mapped to its corresponding class mean vector as near as possible. And at the same time, the different class mean vectors are mapped as far as possible in the reduced subspace. From Eqs. (2) and (3), we can see clearly that the distance from a sample $\mathbf{x}$ to the ith class (point-to-class distance) is simply defined as the distance to the corresponding class mean vector, i.e.,

$$d_i = \|\mathbf{x}-\mathbf{m}_i\|^2 \tag{28}$$

where $\mathbf{m}_i$ is the mean vector of the ith class.

Table 1
Algorithm of RDA.

Input: column sample matrix $\mathbf{X}$
Output: transform matrix $\mathbf{P}_{RDA}$
Step 1: project the training samples into a PCA subspace spanned by its leading eigenvectors: $\hat{\mathbf{X}}=\mathbf{P}_{PCA}^T\mathbf{X}$
Step 2: construct the intra-class reconstruction scatter $\mathbf{S}_w^R$ and inter-class reconstruction scatter $\mathbf{S}_b^R$ using $\hat{\mathbf{X}}$
Step 3: solve the generalized eigenvectors of $\mathbf{S}_b^R\boldsymbol{\varphi}=\lambda\mathbf{S}_w^R\boldsymbol{\varphi}$ and construct $\mathbf{P}=\{\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\ldots,\boldsymbol{\varphi}_d\}$ corresponding to the d largest nonnegative eigenvalues
Step 4: output $\mathbf{P}_{RDA}=\mathbf{P}_{PCA}\mathbf{P}$

Thus, as Yang et al. [29] point out, the optimal classifier for LDA is a minimum distance classifier (i.e., nearest class-mean classifier).

In algorithm of RDA, we focus on the reconstruction error rather than the geometric Euclidean distance. Naturally, we aim to find the projections that minimize the intra-class reconstruction error and maximize the inter-class reconstruction error of each class simultaneously. From the geometric view, the reconstruction error can be seen as the point-to-class distance. In our algorithm, the distance from a sample $\mathbf{x}$ to the ith class is defined as the reconstruction error by the ith class, i.e.,

$$d_i=\|\mathbf{x}-\mathbf{X}_i\boldsymbol{\beta}_i\|^2 \tag{29}$$

Consequently, minimizing the intra-class reconstruction scatter is minimizing the distances from the samples to their own classes. Similarly, maximizing the inter-class reconstruction scatter is maximizing the distances to other classes. The ideas of RDA and LDA are exactly the same. From this point of view, RDA can be seen as another form of LDA with the different definitions of the point-to-class distance.

Although the idea of RDA is the same as LDA, RDA has its own characteristics. Compared with LDA, RDA can extract more features. In the algorithm of LDA, since the ranks of intra-class scatter and inter-class scatter are at most $n-c$ and $c-1$ respectively [4], where n is the total number of training samples and c is the number of the classes. Thus, LDA can extract at most $c-1$ features. In our algorithm, the feature number we can extract depends on the ranks of the intra-class reconstruction scatter and the inter-class reconstruction scatter. Generally, in a high-dimensional subspace, both the intra-class reconstruction scatter and the inter-class reconstruction scatter are of full rank. Thus, RDA can extract at most n features.

4.2. Connections with spectral regression discriminant analysis

Based on the graph embedding framework, SR performs regression after the spectral analysis of the graph. A different graph will lead to a different SR based approach. Particularly, using the graph generated by LDA, Cai et al. developed Spectral Regression Discriminant Analysis (SRDA) [55] and Spectral Regression Kernel Discriminant Analysis (SRKDA) [56] for linear cases and nonlinear cases respectively. In this subsection, we will discuss the differences and the relationships between RDA and the SR based discriminant analysis approaches.

The main purpose of SRDA and SRKDA is to enhance the performances of LDA and Kernel Discriminant Analysis (KDA) [57] respectively. The computational cost will be greatly reduced especially in the high dimensional space. Benefiting from SR, SRDA avoids the SSS problem. Unfortunately, unlike SRDA, RDA still suffers from the SSS problem. Moreover, since both SRDA and SRKDA employ the graph generated by LDA, similar to LDA, SRDA and SRKDA can extract at most $c-1$ features, where c is the number of the class. Differently, RDA can extract at most n features, where n is the total number of training samples.

SR casts the problem of learning an embedding function into a regression framework, which avoids eigen-decomposition of dense matrices. As a powerful dimensionality reduction framework, SR can be easily combined with RDA. After some simple algebraic formulations, we can interpret RDA as the graph embedding form. Therefore, SR can be integrated with RDA by applying SR to solve the graph generated by RDA. Meanwhile, we can extend RDA to its kernel version for nonlinear cases. Just like SRKDA, SR can also be combined with the kernel version of RDA. In the future works, we will do some further study on the characteristics of the SR based RDA and the kernel extension of RDA.

4.3. Connections with MLRE-based feature extractor

Although RDA is very similar to MLREF formally, there are three significant differences between RDA and MLREF.

Firstly, they are under different assumptions. MLREF is based on the locality concept. In the algorithm of MLREF, a sample is represented as a combination of its k nearest neighbors. Differently, RDA is based on the linear subspace assumption, i.e., a single class lies on a linear subspace and each samples can be represented as a combination of its intra-class samples. In other words, MLREF is a local method which preserves the local discriminant structure. RDA is a global method which preserves the global discriminant structure.

Secondly, MLRE-based feature extractor and RDA have different optimal classifiers respectively. RDA follows the assumption, reconstruction strategy and classification rule of LRC. The optimal classifier for RDA is LRC. And the optimal classifier for MLREF is the MLREC.

Thirdly, MLREF has two model parameters while RDA has only one. MLREF needs to set the different neighborhood sizes to characterize the intra-class and inter-class local neighbors. In the classification stage, the neighborhood size still needs to be specified for MLREC. However, the manually selected model parameter can not promise the optimal performance. And if the parameters in the feature extraction step unmatch with the parameter in the classification stage, the MLREC may not be optimal for the MLREF. In summary, it is hard to find the optimal neighborhood sizes for both MLREF and MLREC.

4.4. Connections with LRC

The design of the RDA method is intuitively based on the classification rule of LRC. Nevertheless, RDA and LRC are different in nature. RDA is a feature extraction method and LRC is a classifier. RDA inherits the assumption, reconstruction strategy and classification rule of LRC. Therefore, RDA has close connections with LRC. RDA is modeled by maximizing the point-to-intra-class reconstruction error and simultaneously minimizing the nearest point-to-inter-class reconstruction error. We can conclude that RDA finds the optimal subspace for LRC. It can be seen as a preprocessing step which can improve the performance of LRC significantly.

LRC finds the nearest subspace of a given sample. When the label of the given sample is the same as the label of the nearest subspace, the given sample can be classified correctly. In other words, LRC works more effectively on condition that the given sample matches its nearest subspace. Unfortunately, this condition does not hold well due to the illuminations and noises. RDA is designed to solve this problem. By maximizing the point-to-intra-class reconstruction error and minimizing the nearest point-to-inter-class reconstruction error simultaneously, RDA improves

the compatibility between samples and their nearest subspaces. Therefore, RDA and LRC can be seamlessly integrated into a pattern recognition system.

5. Experiments

5.1. Face recognition

To evaluate the performance of RDA plus LRC, we applied it for face recognition and compared with 4 DR methods (PCA, LDA, MLREF and SRDA) using 3 different classifiers (NN, MLREC and LRC). The code of SRDA is downloaded from http://www.zjucadcg.cn/dengcai/Data/SR.html. During the experiments, three well-known face image databases (ORL, extended YALE-B and FERET) were used to show the robustness and effectiveness of the proposed method.

The ORL database [37] contains images from 40 individuals, each providing 10 different images which were taken at different times and the facial expressions, details (glasses or no glasses) also vary. All images are grayscale and normalized to a resolution of 32×32 pixels for efficiency. Fig. 1 shows sample images of one person from ORL face database.

The YALE-B database [38,39] consists of 2414 frontal face images of 38 subjects under various lighting conditions. The database was divided in five subsets: subset 1 consisting of 266 images (seven images per subject) under nominal lighting conditions was used as the gallery. Subsets 2 and 3, each consisting of 12 images per subject, characterize slight-to-moderate luminance variations, while subset 4 (14 images per person) and subset 5 (19 images per person) depict severe light variations. The images are also grayscale and normalized to a resolution of 32×32 pixels. Fig. 2 shows some images of one person from the YALE-B face database.

The FERET database [40] includes 1400 images of 200 distinct subjects. Each subject has seven images. The subset involves variations in facial expression, illumination and pose. In our experiment, the facial portion of each original image is cropped automatically based on the location of eyes and resized to 40×40 pixels. Fig. 3 shows sample images of one person from FERET database.

On the ORL and FERET face database, i (i=3,4,5)images of one individual are randomly selected for training and the rest are used for test. On the YALE-B database, i (i=5,10,20)images of one individual are randomly selected for training and the rest are used for test. The details of the experimental databases are summarized in Table 2.

Fig. 1. Sample images of one person from the ORL face database.

Fig. 2. Sample images of one person from the YALE-B face database.

Fig. 3. Sample images of one person from FERET face database.

Table 2
Details of the ORL, YALE-B and FERET databases.

Database	Size	Number of subjects	Number of samples per subject	Number of training samples per subject
ORL	32×32	40	10	3/4/5
YALE-B	32×32	38	64	5/10/20
FERET	40×40	200	7	3/4/5

The experimental results of all the methods on three data sets are shown in Table 3, where the value in each entry represents the average recognition accuracy (in percentages) of 50 independent trials, and the number in the brackets is the dimension with the best recognition rate. The model parameter ($k$nearest subspaces) is also listed in Table 3. The recognition rates versus the dimensions are illustrated in Figs. 4–7.

5.2. Finger-knuckle-print recognition

We also have done some further study on the PolyU finger knuckle print (FKP) database [44–46] to evaluate the performance of the proposed method.

In this database, FKP images were collected from 165 volunteers, including 125 males and 40 females. Among them, 143 subjects were 20–30 years old and the others were 30–50 years old. The samples were collected in two separate sessions. In each session, the subject was asked to provide 6 images for each of the left index finger, the left middle finger, the right index finger and the right middle finger. Therefore, 48 images from 4 fingers were collected from each subject. In total, the database contains 7920 images from 660 different fingers. The average time interval between the first and the second sessions was about 25 day. The maximum and minimum time intervals were 96 day and 14 day respectively. All the samples in the database are histogram equalized and resized to 55×110.

For simplicity and efficiency, only the right index fingers of the FKP database are selected and used for recognition. Fig. 8 shows some sample images of the right index finger from one individual.

In the experiments, 5 sample images of the right index finger from one individual are randomly selected for training and the rest are for test. This procedure is repeated for 50 times and the maximal average recognition rates are shown in Table 4. We choose the nearest subspace parameter k=75. The recognition rates versus the variation of dimensions are illustrated in Figs. 9 and 10.

5.3. Evaluation of the experimental results

The above experiments show that the recognition rates of RDA plus LRC are higher than those of other combinations. But, is this difference statistically significant? In this section, we evaluate the experimental results using the null hypothesis statistical test based on Bernoulli model [48,49]. If the resulting p-value is below the desired significance level (i.e., 0.05), the null hypothesis is rejected and the performance difference between two algorithms is considered statistically significant. The evaluation results based on the statistical test are summarized as follows:

(1) On the ORL database, RDA plus LRC outperforms PCA plus NN significantly for all the tests (p=0.013, 0.016 and 0.017). And the results of the null hypothesis statistical tests also indicate that, compared with NN, LRC is more suitable for RDA in the trials with 4 and 5 training samples per class (p=0.023 and 0.030 respectively). In the other tests on the ORL database,

Table 3
Maximal recognition rates on the ORL, YALE-B and FERET databases.

Method	Database								
	ORL ($k=3$) Training number			YALE-B ($k=1$) Training number			FERET ($k=190$) Training number		
	3	4	5	5	10	20	3	4	5
PCA+NN	77.2(116)	81.4(152)	84.7(186)	36.1(176)	52.7(362)	68.9(727)	29.4(203)	33.0(242)	38.6(253)
PCA+LRC	81.4(95)	85.0(121)	88.7(142)	59.8(101)	82.7(148)	85.6(190)	40.7(298)	48.6(312)	52.0(335)
LDA+NN	83.1(39)	87.6(39)	91.7(39)	73.4(37)	78.1(27)	86.5(31)	61.9(33)	65.8(20)	69.9(199)
LDA+LRC	84.0(39)	90.8(39)	93.5(39)	65.3(37)	84.1(37)	87.4(37)	65.4(53)	73.4(30)	78.6(51)
MLREF+NN	81.6(65)	81.2(75)	84.0(81)	49.3(101)	56.7(151)	65.3(171)	57.0(159)	63.7(179)	71.0(191)
MLREF+LRC	85.3(65)	90.1(75)	92.3(81)	72.8(101)	87.6(151)	92.8(193)	71.3(159)	78.5(179)	84.3(191)
MLREF+MLREC	85.7(65)	90.8(75)	93.0(81)	70.3(101)	89.9(151)	93.4(193)	72.6(159)	81.7(179)	85.2(191)
LRC	82.1	88.6	92.3	58.0	81.7	90.9	42.0	50.6	55.4
RDA+NN	84.4(33)	89.3(29)	92.0(31)	68.2(71)	57.5(91)	50.1(119)	74.5(23)	77.2(29)	78.1(41)
RDA+LRC	**86.2(47)**	**91.8(65)**	**94.8(67)**	**80.2(119)**	**92.3(87)**	**97.4(91)**	**85.7(27)**	**91.4(29)**	**94.6(35)**
SRDA+NN	85.4(39)	91.4(39)	94.2(39)	72.6(37)	87.4(37)	95.5(37)	62.1(189)	74.7(197)	82.4(197)
SRDA+LRC	85.7(39)	91.6(39)	94.5(39)	71.9(37)	87.0(37)	95.1(37)	66.5(189)	77.9(197)	84.6(199)

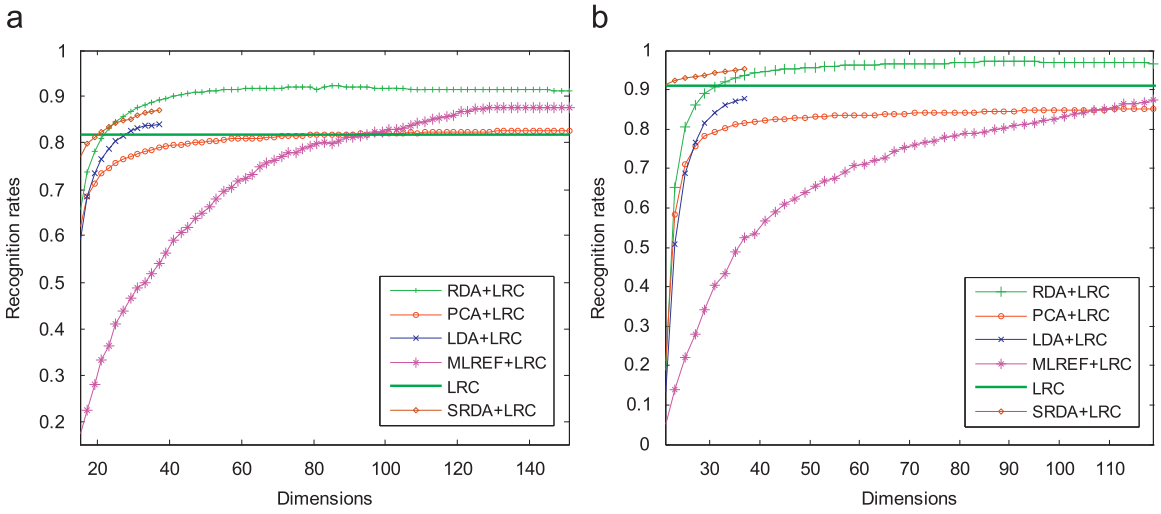

Fig. 4. The recognition rates curves using 6 methods plus LRC on the YALE-B database with (a) 10 and (b) 20 training samples each class respectively.

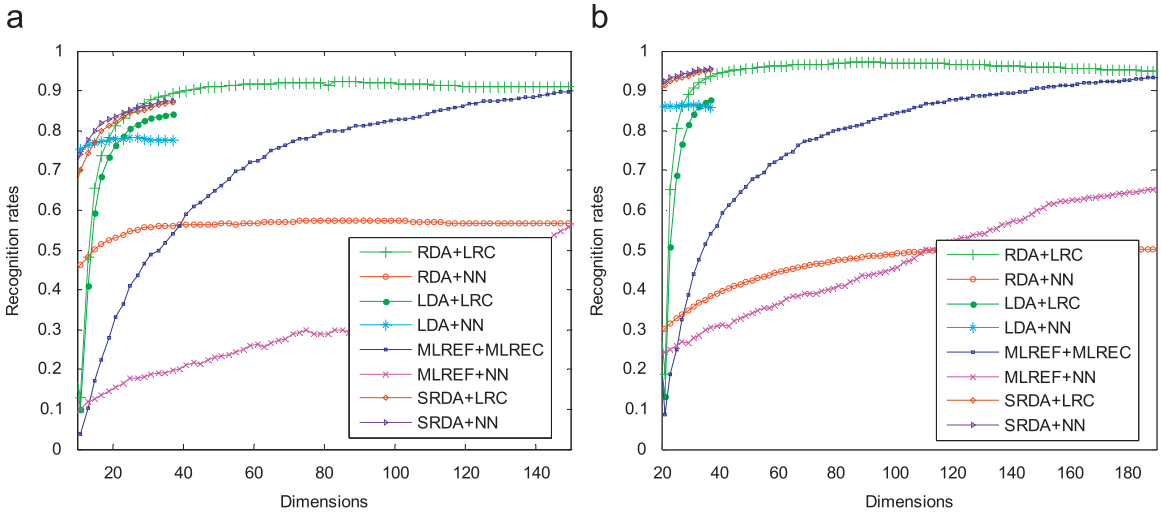

Fig. 5. The recognition rates curves using 4 methods plus LRC, NN and MLREC on the YALE-B database with (a) 10 and (b) 20 training samples each class respectively.

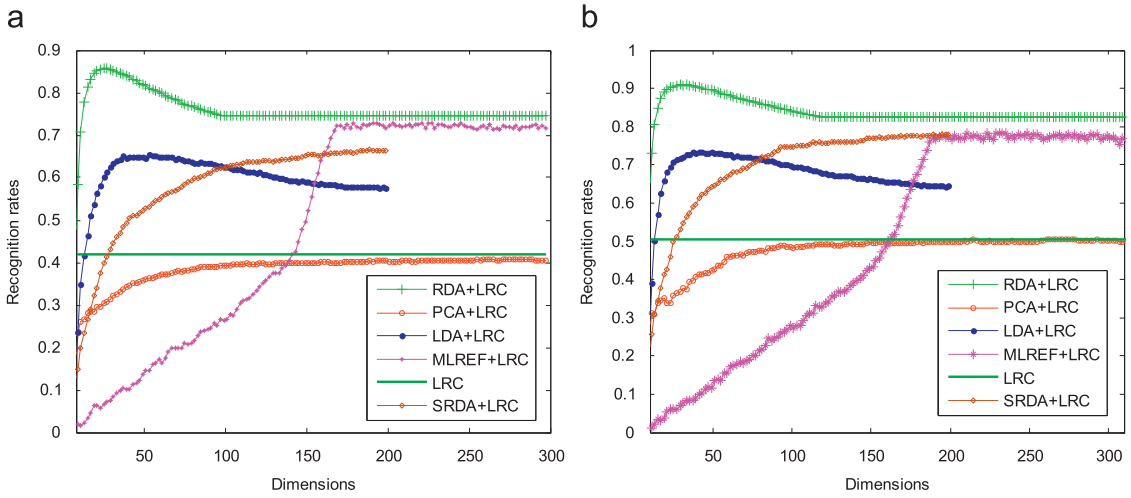

Fig. 6. The recognition rates curves using 6 methods plus LRC on the FERET database with (a) 3 and (b) 4 training samples each class respectively.

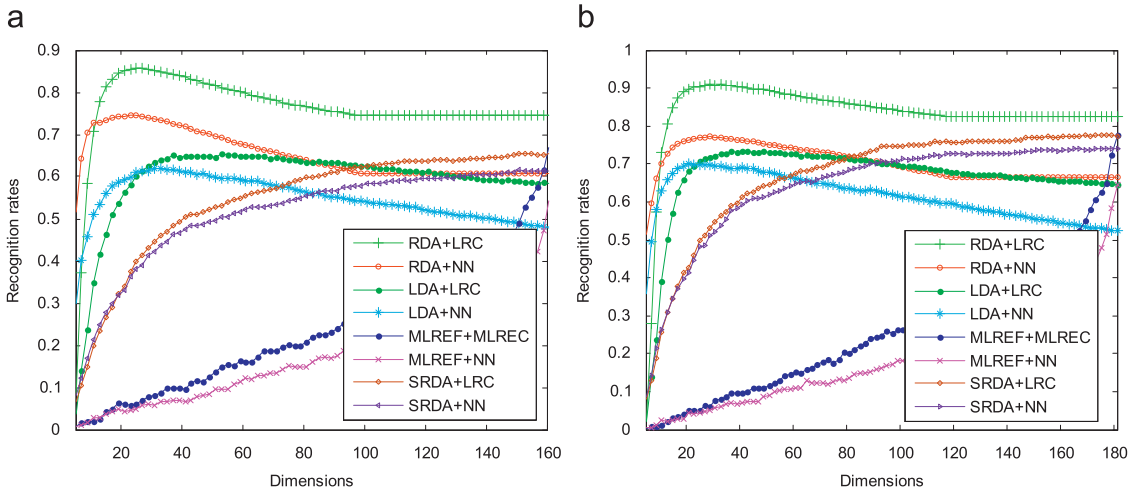

Fig. 7. The recognition rates curves using 4 methods plus LRC, NN and MLREC on the FERET database with (a) 3 and (b) 4 training samples each class respectively.

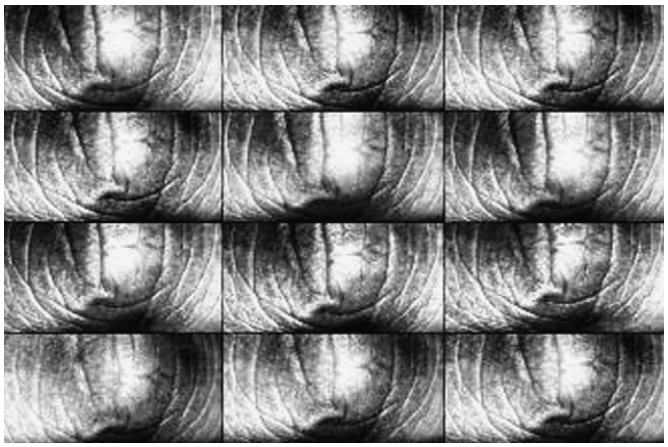

Fig. 8. Sample images of the right index finger from one individual.

although the recognition rates of RDA plus LRC are higher than those of other combinations, the performance differences between RDA plus LRC and other combinations are not statistically significant.

(2) On the YALE-B and FERET databases, RDA plus LRC is significantly better than other combinations under condition of variations in illuminations ($p < 0.001$).
(3) On the FKP database, RDA plus LRC also outperforms other combinations significantly ($p < 0.001$).

5.4. Discussions

From the above results, we can draw the following conclusions:

(1) RDA plus LRC has a good performance and surpasses other competing combinations significantly. The experimental results stated in Tables 3 and 4 show RDA plus LRC achieves higher recognition rates than other combinations.
(2) In most cases, the performance of LRC is higher than that of the NN classifier. In the sight of representation, the NN classifier classifies the test sample based on the best representation in terms of a single training sample, whereas LRC classifies the test sample based on the best linear representation in terms of all the training samples in a specific class. In other words, LRC regards a specific object class as a whole and extracts the subspace structure information, whereas the NN classifier treats each sample separately and ignores the subspace structure

Table 4
Maximal recognition rates on the finger knuckle print database.

Method	PCA+NN	PCA+LRC	LDA+NN	LDA+LRC	SRDA+LRC	SRDA+NN
Rates	89.4(37)	88.1(31)	86.7(31)	91.2(57)	93.3(99)	90.3(97)
Method	MLREF+MLREC	MLREF+NN	MLREF+LRC	LRC	RDA+NN	RDA+LRC
Rates	90.0(107)	83.5(107)	88.5(107)	86.1	89.9(37)	**93.9(69)**

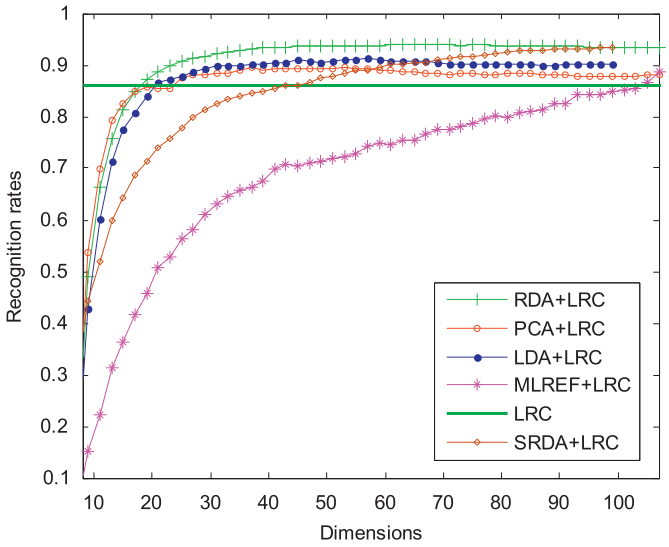

Fig. 9. Recognition rates curves using 6 methods plus LRC on the FKP database.

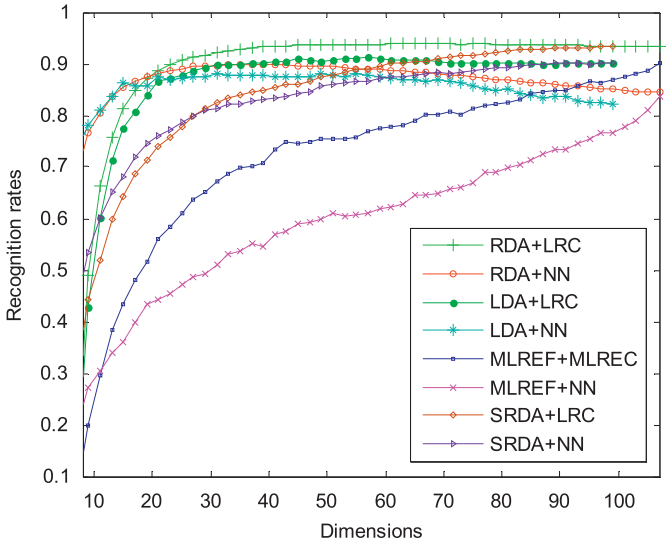

Fig. 10. Recognition rates curves using 4 methods plus LRC, NN and MLREC on the FKP database.

information. For a given sample, NN evaluates the similarity of each training sample by Euclidean distance and LRC evaluates the similarity of each object class. It is obvious that LRC utilizes the class structure information and classifies more accurately if the assumption is consistent with the subspace structure.

(3) Compared with other feature extraction methods, RDA is the most suitable methods for LRC. As can be seen from Tables 3 and 4, the framework of RDA plus LRC continuously outperforms other combinations. We observe that the features extracted by RDA are very discriminative. As demonstrated in Figs. 4–7, RDA achieves much higher recognition rates using very fewer features. It means LRC can utilize the RDA features effectively. Although MLREF is similar to RDA, it remains unclear how to match the parameters of MLREF and MLREC to achieve optimal performances. Without theoretical guidance, the manually selected parameters may probably lead to the performance loss.

(4) From the experimental results on the YALE-B database, we find that the NN classifier may not be robust for RDA under illumination conditions. The YALE-B database contains a wide range of illumination change from different directions. As the results show, with the growth of training number, the recognition rates get worse. Technically, minimizing the intra-class reconstruction error usually can not guarantee that a sample and its nearest neighbor are in the same class. Therefore, the NN classifier may not work effectively in the RDA subspace.

6. Conclusions

In this paper, a new method is proposed for feature extraction. If the reconstruction error is defined as the point-to-class distance, the ideas of RDA and LDA are exactly the same. Thus RDA can be viewed as another form of LDA from the reconstructive view. The projections of RDA uncover and separate the subspaces corresponding to different classes in the reduced subspace. RDA considers both the intra-class reconstruction scatter and the inter-class reconstruction scatter at the same time and seeks to find the projections maximizing the ratio of the inter-class reconstruction scatter and the intra-class reconstruction scatter. The experimental results on three popular face image databases and one FKP database demonstrate RDA plus LRC is more effective than other combinations of DR methods and classifications.

Compared with LDA, RDA has two significant advantages:

(1) RDA has natural connections to classifiers as RDA is induced from LRC. LRC can fully utilize the characteristic of the RDA subspace. RDA and LRC can be seamlessly integrated into a pattern recognition system.

(2) RDA overcomes the drawback of LDA: RDA can extract more features than LDA.

Just like LDA, RDA also suffers from the small sample size problems. To avoid singularity, in this paper, we use PCA to reduce the dimension of the original space first. However, this step may lose some null space information for discrimination. In the future work, we aim to take full advantage of the null space to further improve the performance of RDA.

Acknowledgments

This work is partially supported by the National Science Foundation of China under Grant nos. 90820306, 60873151, 60973098 and 61005008. We also thank Prof. Deng Cai for providing the code.

References

[1] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.
[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, Boston, MA, 1990.

[3] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second edition, John Wiley and Sons, Inc., 2000.
[4] P.N. Belhumeur, J. Hespanda, D. Kiregeman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[5] Daniel L. Swets, John Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.
[6] C.J. Liu, H. Wechsler, Robust coding schemes for indexing and retrieval from large face databases, IEEE Trans. Image Process. 9 (1) (2000) 132–137.
[7] J. Yang, J. Yang, Why can LDA be performed in PCA transformed space? Pattern Recognition 36 (2) (2003) 563–566.
[8] J. Ye, R. Janardan, C. Park, H. Park, An optimization criterion for generalized discriminant analysis on undersampled problems, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8) (2004) 982–994.
[9] W. Zhao, R. Chellappa, J. Phillips, Subspace Linear Discriminant Analysis for Face Recognition, Technical Report, CS-TR4009, University of Maryland, 1999.
[10] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminative common vectors for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 4–13.
[11] Z. Hong, J. Yang, Optimal discriminant plane for a small number of samples and design method of classifier on the plane, Pattern Recognition 24 (4) (1991) 317–324.
[12] J.H. Friedman, Regularized discriminant analysis, J. Am. Stat. Assoc. 84 (1989) 165–175.
[13] T. Hastie, R. Tibshirani, Penalized discriminant analysis, Ann. Stat. 23 (1995) 73–102.
[14] L.F. Chen, H.Y.M. Liao, M.T. Ko, G.J. Yu., A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (1) (2000) 1713–1726.
[15] H. Yu, J. Yang, A direct LDA algorithm for high dimensional data—with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.
[16] M. Loog, Approximate pairwise accuracy criterion for multiclass linear dimension reduction: generalization of the Fisher criterion, IEEE Trans. Pattern Anal. Mach. Intell. 26 (7) (2001) 762–766.
[17] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, in: Proceedings of Advances in Neural Information Processing Systems, 2003, pp. 97–104.
[18] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[19] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[20] J.B. Tenenbaum, V. de. Silva, J.C. Langford., A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[21] H.-T. Chen, H.-W. Chang, T.-L. Liu, Local discriminant embedding and its variants, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2005 (CVPR 2005), pp. 846–853.
[22] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extension: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[23] S. Yan, H. Wang, Semi-supervised learning by sparse representation, in: Proceedings of SDM, 2009.
[24] X. He, Deng Cai, S. Yan, H.J. Zhang, Neighborhood preserving embedding, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, 2005, pp. 1208–1213.
[25] J. Yang, D. Zhang, J. Yang, B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics, IEEE Trans. Pattern Anal. Mach. Intell. (2007) 650–664.
[26] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using Laplacian faces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[27] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2106–2112.
[28] Jian Yang, Zhen Lou, Zhong Jin, Jing-yu Yang, Minimal local reconstruction error measure based discriminant feature extraction and classification, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), Alaska, June 23–28, 2008, pp. 1–6.
[29] J. Yang, L. Zhang, J. Yang, D. Zhang, From classifiers to discriminators: a nearest neighbor rule induced discriminant analysis, Pattern Recognition (2011).
[30] R. Barsi, D. Jacobs, Lambertian reflection and linear subspaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2) (2003) 218–233.
[31] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning; Data Mining, Inference and Prediction, Springer, 2001.
[32] G.A.F. Seber, Linear Regression Analysis, Wiley-Interscience, 2003.
[33] T.P. Ryan, Modern Regression Methods, Wiley-Interscience, 1997.
[34] W. Yu, X. Teng, C. Liu, Face recognition using discriminant locality preserving projections, Image Vision Comput. 24 (2006) 239–248.
[35] J. Wright, A. Yang, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[36] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[37] F. Samaria, A. Harter, Parameterisation of a stochastic model for human face identification, in: Proceedings of Second IEEE Workshop Applications of Computer Vision, December 1994.
[38] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.
[39] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[40] P.J. Phillips, H. Wechsler, J.S. Huang, P.J. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, Image Vision Comput. 16 (5) (1998) 295–306.
[41] Wankou Yang, Changyin Sun, Lei Zhang, A multi-manifold discriminant analysis method for image feature extraction, Pattern Recognition 44 (8) (2011) 1649–1657.
[42] Wankou Yang, Jianguo Wang, Mingwu Ren, Jingyu Yang, Feature extraction based on laplacian bidirectional maximum margin criterion, Pattern Recognition 42 (11) (2009) 2327–2334.
[43] Z. Jin, J.Y. Yang, Z. Hu, Z. Lou, Face recognition based on the uncorrelated discrimination transformation, Pattern Recognition 34 (7) (2001) 1405–1416.
[44] L. Zhang, L. Zhang, D. Zhang, H. Zhu, Online finger-knuckle-print verification for personal authentication, Pattern Recognition 43 (7) (2010) 2560–2571.
[45] L. Zhang, L. Zhang, D. Zhang, Finger-knuckle-print: a new biometric identifier, in: Proceedings of the IEEE International Conference on Image Processing, 2009.
[46] The FKP Database: 〈http://www.comp.polyu.edu.hk/~biometrics/FKP.htm〉.
[47] S. Petridis, S.T. Perantonis, On the relation between discriminant analysis and mutual information for supervised linear feature extraction, Pattern Recognition 37 (5) (2004) 857–874.
[48] W. Yambor, B. Draper, R. Beveridge, Analyzing PCA-based face recognition algorithms: eigenvector selection and distance measures, in: H. Christensen, J. Phillips (Eds.), Empirical Evaluation Methods in Computer Vision, World Scientific Press, Singapore, 2002.
[49] J.R. Beveridge, K. She, B. Draper, G.H. Givens, Parametric and nonparametric methods for the statistical evaluation of human ID algorithms, in: Proceedings of the Third Workshop Empirical Evaluation of Computer Vision Systems, December 2001.
[50] Deng Cai, Xiaofei He, Jiawei Han, Hong-Jiang Zhang, Orthogonal Laplacianfaces for face recognition, IEEE Trans. Image Process. 15 (11) (2006) 3608–3614.
[51] Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, Hujun Bao, Locality sensitive discriminant analysis, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07), 2007.
[52] Deng Cai, Xiaofei He, Jiawei Han. Spectral Regression: A Unified Subspace Learning Framework for Content-Based Image Retrieval, ACM Multimedia, 2007.
[53] Deng Cai, Xiaofei He, Jiawei Han, Spectral regression: a unified approach for sparse subspace learning, in: Proceedings of the IEEE International Conference on Data Mining, 2007.
[54] Deng Cai, Spectral Regression: A Regression Framework for Efficient Regularized Subspace Learning, Ph.D. Thesis, UIUC, 2009.
[55] Deng Cai, Xiaofei He, Jiawei Han, SRDA: an efficient algorithm for large-scale discriminant analysis, IEEE Trans. Knowl. Data Eng. 20 (1) (2008) 1–12.
[56] Deng Cai, Xiaofei He, Jiawei Han, Speed up kernel discriminant analysis, VLDB J. (2011).
[57] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.-R. Müller, Fisher discriminant analysis with kernels, in: Proceedings of IEEE Neural Networks for Signal Processing Workshop (NNSP), 1999.

Yi Chen received the B.S. degree in the School of Computer Science and Technology, Nanjing University of Science and Technology (NUST), PR China, in 2003. Now he is pursuing the Ph.D. degree in Pattern Recognition and Intelligent Systems from Nanjing University of Science and Technology. His current research interests include pattern recognition, computer vision and image processing. E-mail: cystory@qq.com

Zhong Jin received the B.S. degree in mathematics, the M.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 1982, 1984 and 1999, respectively. He is a professor in the department of computer science, NUST. He visited the department of computer science and engineering, the Chinese University of Hong Kong from January 2000 to June 2000 and from November 2000 to August 2001 and visited the Laboratoire HEUDIASYC, UMR CNRS 6599, Universite de Technologie de Compiegne, France, from October 2001 to July 2002. He also visited the Centre de Visioper Computador, Universitat Autonoma de Barcelona, Spain, as the Ramon y Cajal program Research Fellow. His current interests are in the areas of pattern recognition, computer vision, face recognition, facial expression analysis and content-based image retrieval. E-mail: zhongjin@mail.njust.edu.cn