Feature Extraction Using Class-oriented Regression Embedding

Yi Chen
School of Computer Science and Technology
Nanjing University of Science and Technology
Nanjing 210094, PR China
cystory@qq.com

Zhong Jin
School of Computer Science and Technology
Nanjing University of Science and Technology
Nanjing 210094, PR China
zhongjin@mail.njust.edu.cn

***Abstract*—Based on linear regression techniques, we present a new supervised learning algorithm called Class-oriented Regression Embedding (CRE) for feature extraction. By minimizing the intra-class reconstruction error, CRE finds a low-dimensional subspace in which samples can be best represented as a combination of their intra-class samples. This characteristic can significantly strengthen the performance of the newly proposed classifier called linear regression-based classification (LRC). The experimental results on the extended-YALE Face Database B (YaleB) and CENPARMI handwritten numeral database show the effectiveness and robustness of CRE plus LRC.**

Keywords-component; *Feature extraction, dimensionality reduction, linear regression-based classification, face recognition, handwritten numeral recognition*

I. INTRODUCTION

In real world applications, the data are usually high-dimensional. For robust representation and recognition, feature extraction or dimensionality reduction (DR) techniques are employed to find a meaningful low-dimensional representation of the high-dimensional data as a preprocessing step.

Among the existing feature extraction methods, Principal Component Analysis (PCA) [1] is the most popular one. PCA finds the projections that best represent the whole training samples. Thus, PCA is a global method which preserves the global Euclidean structure. Another representation-based method is neighborhood preserving embedding (NPE) [3]. In NPE algorithm, a sample is expressed by its local neighbors. As contrary to PCA, NPE utilizes the locality concept. From the point of view, NPE is a local method which preserves the local neighborhood structure. Motivated by sparse representation [7, 8], a novel DR technique called sparsity preserving projections (SPP) [10] was proposed. SPP treats each sample as a sparse combination of the whole samples. Therefore, SPP is a global sparse method.

To the best of our knowledge, the above techniques and the most existing DR methods are designed independently of classifiers. Therefore, the characteristics of the learned subspace are invisible to the classifiers. In other words, the classifiers usually do not make the best use of the characteristics of the learned subspace. Thus, the performance of the pattern recognition system potentially degrades.

In the past decades, nearest subspace (NS) classifications [13-15] have been drawn considerable attention. These NS approaches seek the best representation by samples in each class. Recently, two latest NS methods called sparse representation-based classification (SRC) [5] and linear regression-based classification (LRC) [6] were proposed for robust face recognition. SRC and LRC take the minimum reconstruction error criterion as the classification rule. Concretely, SRC and LRC assume a test sample belongs to the class with a minimum reconstruction error. For efficiency, downsampled images are directly used for classification instead of original images. However, due to the variations in lighting, facial expression, and pose, neither the original image space nor the downsampled image space can guarantee the minimum intra-class reconstruction for discrimination. In other words, the assumption of SRC and LRC may not hold well in practical application.

As the important parts of the pattern recognition system, the feature extraction methods and the classifiers should be complementary to each other. In order to achieve this goal and strengthen the discrimination of the pattern recognition system, we take the reconstruction strategy and classification rule of LRC into account. Then we induce a new supervised learning method called Class-oriented Regression Embedding (CRE). CRE aims to preserve the intra-class reconstruction relationships. By minimizing the intra-class reconstruction error, CRE finds the projections by which a sample can be best expressed by its intra-class samples. In this way, CRE can be integrated with LRC closely. In the CRE algorithm, we first construct the global reconstruction coefficient matrix using linear regression on the training samples and then the low-dimensional subspace is calculated to best preserve the intra-class reconstruction relationships. At the classification stage, the nearest subspace classifier LRC is applied to determine the labels of the samples.

The rest of the paper is organized as follows: NPE and LRC are reviewed in Section II. In Section III, CRE is described in detail. In Section IV, the experiments are presented on the well-known databases to demonstrate the effectiveness. Finally, conclusions are drawn in Section V.

978-1-4577-0121-4/11/$26.00 ©2011 IEEE

II. RELATED WORKS

In this section, we will introduce some related works called neighborhood preserving embedding and linear regression-based classification.

A. Neighborhood Preserving Embedding

NPE assumes each sample can be expressed by its k nearest neighbors. The goal of NPE is to preserve the local neighborhood structure. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n]$ be the input data points and the transformation vector $\mathbf{a}$ maps these n points to a set of points $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$.

$$\begin{aligned} \mathbf{a} &= \arg\min_{\mathbf{a}} \sum_i \left(\mathbf{y}_i - \sum_j \mathbf{W}_{ij} \mathbf{y}_j \right)^2 \\ &= \arg\min_{\mathbf{a}} \mathbf{a}^T \mathbf{X} \mathbf{M} \mathbf{X}^T \mathbf{a} \end{aligned} \quad (1)$$

where $\mathbf{W}$ denotes the weight matrix (Please see [3] for more details), $\mathbf{M} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W})$ and $\mathbf{I}$ is an identity matrix.

The following constraint is added to remove an arbitrary scaling factor in the projection:

$$\mathbf{a}^T \mathbf{X} \mathbf{X}^T \mathbf{a} = 1 \quad (2)$$

Using Lagrange multipliers, we have:

$$\mathbf{X} \mathbf{M} \mathbf{X}^T \mathbf{a} = \lambda \mathbf{X} \mathbf{X}^T \mathbf{a} \quad (3)$$

The optimal projections of NPE are the generalized eigenvectors of Eq.(3) corresponding to the smallest eigenvalues.

B. Linear Regression-based Classification

LRC assumes samples from a single class lie on a linear subspace [2]. Using this concept, a linear model is developed. In this model, a test image is represented as a linear combination of class-specific galleries. Thereby the task of recognition is defined as a problem of linear regression.

Assume we have n samples from c classes. Let n_i represents the training number of the i^{th} class and $\mathbf{x}_i^j \in \mathrm{R}^d$ denotes the j^{th} sample of the i^{th} class, $i = 1, 2, ..., c$, $j = 1, 2, ..., n_i$. $\mathbf{X}_i$ is a class-specific model generated by stacking the d -dimensional image vectors,

$$\mathbf{X}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2, ..., \mathbf{x}_i^{n_i}] \in \mathbf{R}^{d \times n_i} \quad (4)$$

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

$$\mathbf{y} = \mathbf{X}_i \boldsymbol{\beta}_i \quad (5)$$

where $\boldsymbol{\beta}_i \in \mathbf{R}^{n_i \times 1}$ is the reconstruction coefficients. Given that $d \geq n_i$, the system of equations in Eq.(5) is well conditioned and can be estimated by least-squares estimation (LSE) [11]:

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

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

$$\hat{\mathbf{y}}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}_i = \mathbf{X}_i \left(\mathbf{X}_i^T \mathbf{X}_i \right)^{-1} \mathbf{X}_i^T \mathbf{y} \quad (7)$$

Then we compute the distance measure between the test sample $\mathbf{y}$ and reconstructed sample $\hat{\mathbf{y}}_i$, $i = 1, 2, ..., c$ and rule in favor of the class with minimum distance, i.e.,

$$\min_i \left\| \mathbf{y} - \mathbf{X}_i \hat{\boldsymbol{\beta}}_i \right\|^2, \ i = 1, 2, ..., c \quad (8)$$

III. CLASS-ORIENTED REGRESSION EMBEDDING

A. Basic Idea of CRE

Usually, the DR methods are designed under some assumptions and more effective for the datasets that conform to the assumptions. Then the nearest neighbor classifier (NNC) is employed to classify the test samples. However, the learned subspace may not be optimal for NNC. Obviously, if the DR method and the classifier are under different or even opposite assumption, more samples will be classified incorrectly. Meanwhile, the characteristics of the learned subspace are invisible to the classifiers. Consequently, the classifiers do not fully use the characteristics of the learned subspace for discrimination.

For example, Yang et al. [9] point out that that the classical Linear Discriminant Analysis (LDA) is optimal for minimum distance classifier (i.e. nearest class-mean classifier) rather than the NNC. Intuitively, the NNC is more suitable for the manifold learning algorithms because both of the manifold learning algorithms and the NNC are based on the "locality" concept.

We argue that the designs of feature extraction methods should take consideration of the model of one specific classifier. Based on the model of one specific classifier, the designed feature extraction methods can be bonded with the specific classifier closely. Otherwise, the performance of the pattern recognition system potentially degrades.

To consist with LRC, we assume that samples from a specific object class lie on a linear subspace and each sample can be represented as a combination of its intra-class samples. The intra-class reconstruction error can be computed as follows:

$$\boldsymbol{\varepsilon}_i^j = \left\| \mathbf{x}_i^j - \mathbf{X}_i \boldsymbol{\beta}_i^j \right\|^2 \quad (9)$$

According to the linear subspace assumption, the intra-class reconstruction error should be as small as possible. Naturally, we have the following objective function:

$$\min \sum_i \sum_j \boldsymbol{\varepsilon}_i^j = \sum_i \sum_j \left\| \mathbf{x}_i^j - \mathbf{X}_i \boldsymbol{\beta}_i^j \right\|^2 \quad (10)$$

By Eq.(10), a given image can be represented as a combination of its intra-class samples more precisely. In other words, Eq.(10) preserves the intra-class reconstruction relationships. It is worthwhile to point out that Eq.(10) is very important to LRC. As we introduced above, the classification rule of LRC is to find the class with minimum reconstruction error. Therefore, LRC works more effectively when the intra-class reconstruction error is minimal.

However, due to the variations of illumination, pose and etc., the original space may not obey the linear subspace assumption. We aim to find the low-dimensional subspace that minimizes the intra-class reconstruction error. Under the linear subspace assumption, the existing feature extraction methods are not optimal for LRC since they do not take intra-class reconstruction error into account. We believe that better results will be achieved if the intra-class reconstruction information is imposed in the objective function. To find a good subspace in line with LRC, we first inherit the assumption of LRC. Then we take the classification rule and reconstruction strategy of LRC into consideration. Finally, we present a novel supervised method called Class-oriented Regression Embedding (CRE) for feature extraction.

B. Formulation and algorithm

Our goal is to find the low-dimensional subspace in which the intra-class reconstruction error is minimal. Suppose we have obtained the optimal projections $\mathbf{P}=\{\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,...,\boldsymbol{\varphi}_d\}$. Project each data point $\mathbf{x}_i^j$ onto the subspace:

$$\mathbf{y}_i^j=\mathbf{P}\mathbf{x}_i^j \tag{11}$$

For a sample $\mathbf{x}_i^j$, we aim to represent it as a combination of the rest of the samples from the i^{th} class, i.e.,

$$\mathbf{x}_i^j=\mathbf{X}_i\boldsymbol{\beta}_i^j \tag{12}$$

where $\boldsymbol{\beta}_i^j$ be an n_i -dimensional vector in which the j^{th} element is equal to zero (implying the $\mathbf{x}_i^j$ is removed from $\mathbf{X}_i$).

Then let us define the intra-class reconstruction coefficient matrix:

$$\mathbf{W}_i=\left[\boldsymbol{\beta}_i^1,\boldsymbol{\beta}_i^2,...,\boldsymbol{\beta}_i^{n_i}\right] \tag{13}$$

Based on the intra-class reconstruction coefficient matrix in Eq.(13), we can further define the global reconstruction coefficient matrix:

$$\boldsymbol{\beta}=\begin{bmatrix}\mathbf{W}_1 & \cdots & \cdots & 0\\ \vdots & \mathbf{W}_2 & & \vdots\\ \vdots & & \ddots & \vdots\\ 0 & \cdots & \cdots & \mathbf{W}_c\end{bmatrix} \tag{14}$$

Suppose $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_c]$ is the column sample matrix. According to Eq.(14), we can rewrite the objective function as follows:

$$\begin{aligned}\min\sum_i\sum_j\boldsymbol{\varepsilon}_i^j&=\sum_i\sum_j\left\|\mathbf{y}_i^j-\mathbf{Y}_i\boldsymbol{\beta}_i^j\right\|^2\\&=tr\left(\sum_i\sum_j\left[\left(\mathbf{y}_i^j-\mathbf{Y}_i\boldsymbol{\beta}_i^j\right)\right]\left[\left(\mathbf{y}_i^j-\mathbf{Y}_i\boldsymbol{\beta}_i^j\right)\right]^T\right)\\&=tr\left(\left(\mathbf{Y}-\mathbf{Y}\boldsymbol{\beta}\right)\left(\mathbf{Y}-\mathbf{Y}\boldsymbol{\beta}\right)^T\right)\\&=tr\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{I}-\boldsymbol{\beta}\right)\left(\mathbf{I}-\boldsymbol{\beta}\right)^T\mathbf{X}^T\mathbf{P}\right)\\&=tr\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{I}-\boldsymbol{\beta}+\boldsymbol{\beta}^T+\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\mathbf{X}^T\mathbf{P}\right)\end{aligned} \tag{15}$$

To avoid degenerate solutions, we add a constraint:

$$\mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P}=\mathbf{I} \tag{16}$$

Then we have the objective function:

$$\begin{aligned}&\min_{\mathbf{P}} tr\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{I}-\boldsymbol{\beta}+\boldsymbol{\beta}^T+\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\mathbf{X}^T\mathbf{P}\right)\\&\text{s.t. } \mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P}=\mathbf{I}\end{aligned} \tag{17}$$

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

$$\begin{aligned}&\min_{\boldsymbol{\varphi}}\boldsymbol{\varphi}^T\mathbf{X}\left(\mathbf{I}-\boldsymbol{\beta}+\boldsymbol{\beta}^T+\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\mathbf{X}^T\boldsymbol{\varphi}\\&\text{s.t. } \boldsymbol{\varphi}^T\mathbf{X}\mathbf{X}^T\boldsymbol{\varphi}=1\end{aligned} \tag{18}$$

Using Lagrange multipliers we rewrite the objective function in Eq.(18) and taking derivatives then equaling them to zero, we have:

$$\mathbf{X}\left(\mathbf{I}-\boldsymbol{\beta}+\boldsymbol{\beta}^T+\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\mathbf{X}^T\boldsymbol{\varphi}=\lambda\mathbf{X}\mathbf{X}^T\boldsymbol{\varphi} \tag{19}$$

The optimal solutions can be obtained by solving the generalized eigenvectors in Eq.(19) corresponding to the smallest eigenvalues. It is should be noticed when the dimensionality of origin space is larger than the total number of the training samples, the matrix $\mathbf{X}\mathbf{X}^T$ in Eq.(19) is singular. To avoid the singularity, we first apply PCA on the original samples to reduce the dimensionality of the original input subspace.

The algorithm of CRE can be summarized as follows:

Input: Column sample matrix $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_c]$

Output: Transform matrix $\mathbf{P}_{CRE}$

Step 1: Project the training samples onto a PCA subspace: $\tilde{\mathbf{X}}=\mathbf{P}_{PCA}^T\mathbf{X}$

Step 2: Construct the global reconstruction coefficient matrix $\boldsymbol{\beta}$ using $\tilde{\mathbf{X}}$.

Step 3: Solve the generalized eigenvectors of $\tilde{\mathbf{X}}\left(\mathbf{I}-\boldsymbol{\beta}+\boldsymbol{\beta}^T+\boldsymbol{\beta}\boldsymbol{\beta}^T\right)\tilde{\mathbf{X}}^T\boldsymbol{\varphi}=\lambda\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\boldsymbol{\varphi}$ corresponding to the first d smallest eigenvalues $\mathbf{P}=\{\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,...,\boldsymbol{\varphi}_d\}$.

Step 4: Output $\mathbf{P}_{CRE}=\mathbf{P}_{PCA}\mathbf{P}$.

IV. EXPERIMENTS

To show the effectiveness of CRE, we apply our method for face recognition and handwritten numeral recognition. We compare CRE with PCA, LDA, NPE and SPP using NNC, SRC [5] and LRC respectively. As a baseline, we also give the classification results of the selected classifiers directly using the raw data without dimensionality reduction.

A. *Face recognition*

The YaleB database consists of 2414 frontal face images of 38 subjects under various lighting conditions. The images are grayscale and normalized to a resolution of 32×32 pixels. Fig.1 shows some images of one person from the YaleB face database.

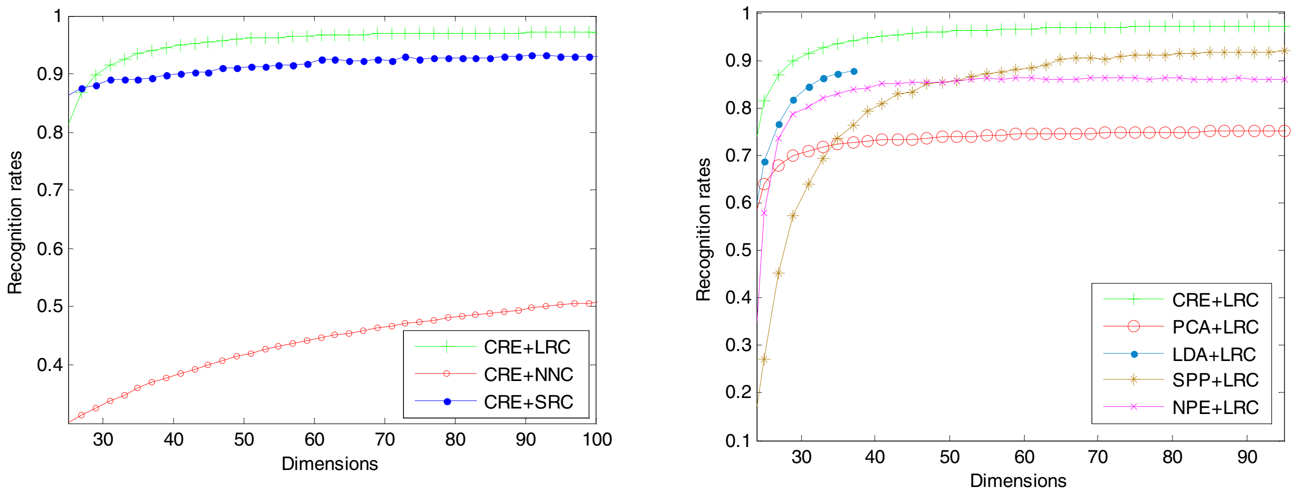
Figure 1. Some images of one person from the YaleB face database

On the YaleB database, 20 images of one individual are randomly selected for training and the rest are used for test. Since the experiments using SRC or SPP are computationally expensive, we repeated the procedures 10 times independently when SPP or SRC are employed. And the other experiments on the YALBE database are repeated 50 times. Table I is the maximal average recognition rates using NNC, SRC and LRC on the YaleB database. The numbers in the brackets are the dimensions with the best recognition rates.

The experimental results show CRE can enhance the performance of LRC. As can be seen from Table I, using CRE to reduce dimensionality first, LRC can achieve much higher recognition rates. Meanwhile, we find that, for LRC, CRE is more effective than other feature extraction methods. Different from other feature extraction methods, CRE preserves the intra-class reconstruction relationships. It means this characteristic is very important for LRC. We also observe that the recognition rate using CRE plus NN is very low. Theoretically, minimizing the intra-class reconstruction error can not guarantee that a given sample and its nearest neighbor belong to the same class. So NNC may not robust for CRE.

TABLE I. THE RECOGNITION RATES (%) ON THE YALEB DATABASE

Methods / Classifier	Baseline	PCA	LDA	NPE	SPP	CRE
NNC	42.1	68.9(727)	**91.3(37)**	82.7(152)	84.4(91)	54.3(161)
SRC	92.3	92.6(192)	91.7(37)	90.2(102)	92.0(82)	**93.4(90)**
LRC	90.9	85.6(190)	87.4(37)	85.3(240)	91.3(91)	**97.2(161)**



(a) (b)

Figure 2. (a) The recognition rates of CRE plus NNC/LRC/SRC on the YaleB database. (b) The recognition rates of 5 methods plus LRC on the YaleB database.

B. *Handwritten numeral recognition*

The Concordia University CENPARMI handwritten numeral database contains 10 numeral classes and each class has 600 samples. In our experiment, we randomly choose 50 samples of each class for training, the remaining 550 samples for testing. We repeat the procedures 10 times independently in the experiment. Table 4 is the maximal average recognition rates using NNC, SRC and LRC on the CENPARMI database. The numbers in the brackets are the dimensions with best recognition rates.

The experimental results on the CENPARMI handwritten numeral database also indicate that CRE plus LRC is more effective than other combinations.

CONCLUSION

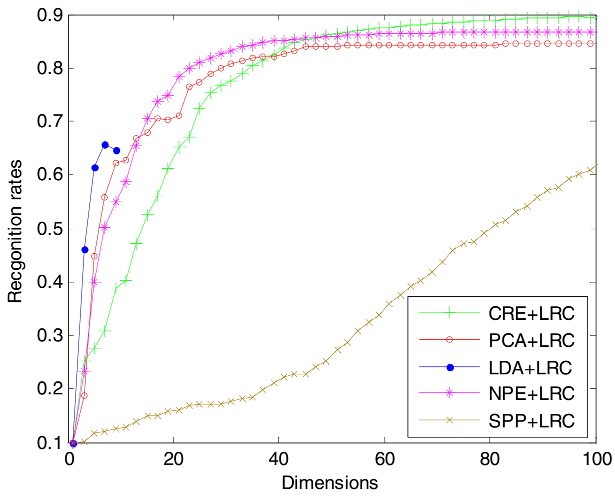
In this paper, we introduce minimal intra-class reconstruction error as a similarity measure and present a novel feature extraction method called class-oriented regression embedding. Different from existed feature extraction methods, CRE preserves the intra-class reconstruction relationships.

According to the classification rule of LRC, LRC can work more effectively in the CRE subspace.

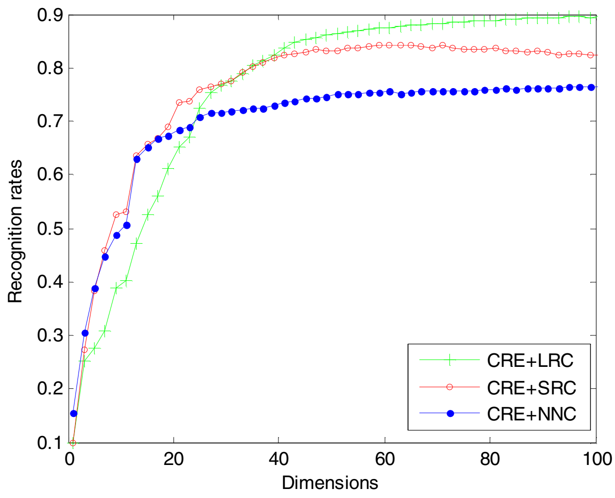
The proposed feature extraction and classification method is evaluated using the CENPARMI handwritten numeral database and the YaleB face image database. The experimental results indicate that the proposed method plus LRC is more effective than other combinations of DR methods and classifications.

TABLE II. THE RECOGNITION RATES (%) ON THE CENPARMI DATABASE

Methods / Classifier	Baseline	PCA	LDA	NPE	SPP	CRE
NNC	80.4	80.6(51)	**84.7(9)**	73.7(71)	77.5(31)	76.6(121)
SRC	80.0	**86.4(33)**	84.0(9)	86.0(80)	82.6(31)	84.5(59)
LRC	86.4	85.3(121)	65.7(7)	86.7(73)	71.1(119)	**90.4(113)**



(a)



(b)

Figure 3. (a) The recognition rates of 5 methods plus LRC on the CENPARMI database. (b) The recognition rates of CRE plus NNC/LRC/SRC on the CENPARMI database.

ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation of China under grant no.90820306, 60873151, 60973098，61005008.

REFERENCES

[1] I. Jolliffe, Principal component analysis. Springer (1986)

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell., 19 (2002) 711-720..

[3] X. He, D. Cai, S. Yan, and H.J. Zhang, Neighborhood preserving embedding. ICCV (2005) 1208–1213.

[4] J. Yang, D. Zhang, J. Yang, and B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics. IEEE Trans. Pattern Anal. Machine Intell.(2007) 650-664.

[5] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Machine Intell., 31 (2008) 210-227.

[6] I. Naseem, R. Togneri, and M. Bennamoun, Linear regression for face recognition. IEEE Trans. Pattern Anal. Machine Intell. (2010).

[7] P. Zhao, and B. Yu, On model selection consistency of Lasso. The Journal of Machine Learning Research 7 (2006) 2541-2563.

[8] R. Tibshirani, Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267-288

[9] J. Yang, Z. Lou, Z. Jin, J.Y. Yang, Minimal local reconstruction error measure based discriminant feature extraction and classification, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), vols. 1–12, 2008, pp. 173–178.

[10] L. Qiao, S. Chen, and X. Tan, Sparsity preserving projections with applications to face recognition. Pattern Recognition 43 (2010) 331-341.

[11] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer, 2001.

[12] J. Yang, L. Zhang, J-y Yang, Z David, From classifiers to discriminators: A nearest neighbor rule induced discriminant analysis, Pattern Recognition 44 (2011) 1387-1402.

[13] S.Z. Li, “Face Recognition Based on Nearest Linear Combinations,” Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 839-844, 1998.

[14] S.Z. Li and J. Lu, “Face Recognition Using Nearest Feature LineMethod,” IEEE Trans. Neural Network, vol. 10, no. 2, pp. 439-443,Mar. 1999.

[15] J.-T. Chien and C.-C. Wu, “Discriminant Wavelet Faces and Nearest Feature Classifiers for Face Recognition,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1644-1649, Dec. 2002.