Neural Process Lett
DOI 10.1007/s11063-012-9252-y

Feature Extraction Based on Maximum Nearest Subspace Margin Criterion

Yi Chen · Zhenzhen Li · Zhong Jin

Abstract Based on the classification rule of sparse representation-based classification (SRC) and linear regression classification (LRC), we propose the maximum nearest subspace margin criterion for feature extraction. The proposed method can be seen as a preprocessing step of SRC and LRC. By maximizing the inter-class reconstruction error and minimizing the intra-class reconstruction error simultaneously, the proposed method significantly improves the performances of SRC and LRC. Compared with linear discriminant analysis, the proposed method avoids the small sample size problem and can extract more features. Moreover, we extend LRC to overcome the potential singular problem. The experimental results on the extended Yale B (YALE-B), AR, PolyU finger knuckle print and the CENPARMI handwritten numeral databases demonstrate the effectiveness of the proposed method.

Keywords Feature extraction · Dimensionality reduction · Face recognition · Finger knuckle print recognition · Linear regression classification

1 Introduction

Recently, the classifications based on reconstruction errors [1–6] have attracted a lot of researchers. Among the existing reconstruction errors based classifications, the most popular ones are SRC [5] and LRC [6]. The main difference between SRC and LRC is the reconstruction strategy. Let us take face recognition for example. SRC, which is based on sparse representation, represents a face image as a sparse combination of all the face images. Differently, based on linear regression model, LRC represents a face image as a linear combination of all the face images from one class. Although SRC and LRC have different

Y. Chen (✉) · Z. Jin
School of Computer Science and Technology, Nanjing University of Science and Technology,
Nanjing 210094, People's Republic of China
e-mail: cystory@qq.com

Z. Li
School of Information Engineering, Jiangxi Manufacturing Technology College, Nanchang 330095, China

Published online: 07 November 2012

reconstruction strategies, their classification rules are based on the same assumption: the probe image belongs to the class with the minimum reconstruction error. Suppose $\mathbf{x}$ is a probe image and $\hat{\mathbf{x}}_i (i = 1, 2, \ldots, c)$ is the reconstructed image by the ith class. The distance from $\mathbf{x}$ to the ith class is defined as $d_i = \left\| \mathbf{x} - \hat{\mathbf{x}}_i \right\|^2$. Then the label of $\mathbf{x}$ will be assigned as the class with the minimum reconstruction error $\min\limits_i d_i (\mathbf{x})$.

Unfortunately, although SRC and LRC have been successfully applied for face recognition, we notice the performances of SRC and LRC degrade severely under the illuminations and noises conditions. That means the intra-class reconstruction errors are probably larger than the inter-class reconstruction errors when the images contains variations of illuminations and noises. In other words, the classification rule of SRC and LRC may not hold well in the original space. To achieve a high performance, a good classification rule considers the characteristics of the feature space. Therefore, a classifier works more effectively only the feature space fits for the classification rule of the classifier. Then a question is: what is the optimal feature space for SRC and LRC?

Intuitively, the optimal feature space can fit for the classification rule of SRC and LRC as far as possible. To the best of our knowledge, most of the existing feature extraction methods such as [7–15] are designed based on the data structures rather than the classifications. Without considering the classification rule of SRC and LRC, the feature subspaces learned by the above methods may not hold the assumption of SRC and LRC well. Therefore, the performances of SRC and LRC potentially degrade. To enhance the performances of SRC and LRC, we propose the maximum nearest subspace margin criterion (MNSMC) according to the classification rules of SRC and LRC. Based on MNSMC, a new feature extractor is developed to find the optimal feature subspace for SRC and LRC.

The rest of the paper is organized as follows. Related works are reviewed in Sect. 2. In Sect. 3, MNSMC is described in detail. In Sect. 4, the experiments are presented on the well-known databases to demonstrate the effectiveness of the proposed method. Finally, conclusions are drawn in Sect. 5.

2 Related Works

In this section, we briefly review SRC and LRC. Suppose $\mathbf{x}_i^j \in \mathbb{R}^n (i = 1, 2, \ldots, c, j = 1, 2, \ldots, n_i)$ is the jth sample from the ith class and $\mathbf{X}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^{n_i}]$ is a set of all the samples from the ith class. Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_c]$ be the set of original training samples.

2.1 Sparse Representation-Based Classification

SRC considers sparse representation has natural discriminating power: taking face images into account, the most compact expression of a certain face image is generally given by the face images from the same class [5].

Denote by $\mathbf{z}$ a probe image. We represent $\mathbf{z}$ in a overcomplete dictionary whose basis vectors are training sample themselves, i.e.,

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

The sparsest solution to Eq. (1) can be sought by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min \|\boldsymbol{\beta}\|_0 , \text{ subject to } \mathbf{z} = \mathbf{X}\boldsymbol{\beta} \tag{2}$$

where $\|\cdot\|_0$ denotes the L_0-norm, which counts the number of nonzero entries in a vector.

Recent research efforts reveal that for certain dictionaries, if the solution $\hat{\boldsymbol{\beta}}$ is sparse enough, finding the solution of the L_0 optimization problem is equivalent to finding the solution to the following L_1 optimization problem [16–18]:

$$\hat{\boldsymbol{\beta}} = \arg\min \|\boldsymbol{\beta}\|_1 \text{ , subject to } \mathbf{z} = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

Then find the class with the minimum reconstruction error

$$\text{identity}\ (\mathbf{z}) = \arg\min_i \{\varepsilon_i\} \tag{4}$$

where $\varepsilon_i = \left\|\mathbf{z} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right\|^2$, $\hat{\boldsymbol{\beta}} = \left[\hat{\boldsymbol{\beta}}_1; \hat{\boldsymbol{\beta}}_2; \ldots; \hat{\boldsymbol{\beta}}_c\right]$ and $\hat{\boldsymbol{\beta}}_i$ is the coefficient vector associated with class i.

2.2 Linear Regression Classification

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) [19–21] 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.

Suppose $\mathbf{z}$ 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.,

$$\mathbf{z} = \mathbf{X}_i\boldsymbol{\beta}_i \tag{5}$$

where $\boldsymbol{\beta}_i \in \mathrm{R}^{n_i \times 1}$ 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:

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

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

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

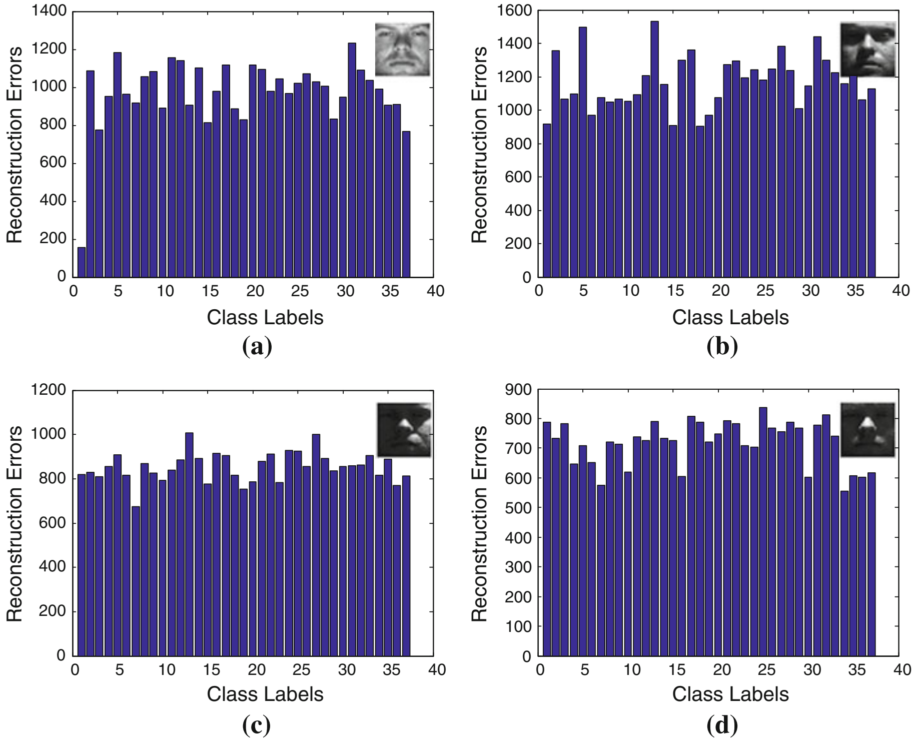
The label is signed as the class with the minimum reconstruction error, i.e.,

$$\text{identity}\ (\mathbf{z}) = \arg\min_i \left\|\mathbf{z} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_i\right\|^2 \tag{8}$$

3 Feature Extraction by Maximum Nearest Subspace Margin Criterion

3.1 Motivation

In the application of face recognition, SRC and LRC demonstrate impressive results. When we investigate the misclassification, we find the variation of illuminations can significantly affect the classification results. An intuitive example is provided in Fig. 1 to state this problem. On the YALE-B face database, we select four face images of the first class under different illumination conditions. According to the classification rule of SRC and LRC, the face image can be classified correctly when the reconstruction error of the first class is minimal. However, taking LRC for example, as can be seen in Fig. 1, only the face image with the frontal illumination (Fig. 1a) has the minimum intra-class reconstruction error. In an extreme case,



(a)

(b)

(c)

(d)

Fig. 1 The reconstruction errors of four *images* of the first class under the different illumination conditions on the YALE-B database

i.e., the face image with the backlight illumination (Fig. 1d), the intra-class reconstruction error is even larger than most of the inter-class reconstruction errors.

The above example indicates the original image space is not suitable for classification due to the variations of illuminations. To solve this problem, we aim to find a feature subspace to reduce the impact of illuminations and enhance the performance of SRC and LRC. According to the classification rule of SRC and LRC, the feature subspace with larger inter-class reconstruction errors and smaller intra-class errors will lead to a good performance. To find the optimal feature subspace, we develop MNSMC for feature extraction.

3.2 Maximum Nearest Subspace Margin Criterion

In this section we will introduce our MNSMC in detail. First let us introduce some notations and preliminary definitions.

Let $\mathbf{x}_i^j \in \mathbb{R}^n (i = 1, 2, \ldots, c, j = 1, 2, \ldots, n_i)$ be the jth sample from the ith class and $\mathbf{X}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^{n_i}]$ is a set of all the samples from the ith class. Based on the reconstruction errors, we first define the nearest subspace.

Definition 1 (*Nearest subspace*). For a given sample $\mathbf{x}_i^j$, its nearest subspace N_i^j is the subspace spanned by the class with the minimum reconstruction error of $\mathbf{x}_i^j$.

To introduce the label information, we further define the two types of nearest subspaces.

Definition 2 (*Homogeneous nearest subspace*). For a given sample $\mathbf{x}_i^j$, its homogeneous nearest subspace is the subspace spanned by the ith class.

Definition 3 (*Heterogeneous nearest subspace*). For a given sample $\mathbf{x}_i^j$, its heterogeneous nearest subspace is the subspace spanned by the class with the minimum inter-class reconstruction error of $\mathbf{x}_i^j$.

Before describe the nearest subspace margin, we provide a way to compute the two types of distances: the point-to-intra-class distance and the point-to-inter-class distance. In this paper, we focus on the reconstruction errors. Thus, the point-to-intra-class distance is defined as the intra-class reconstruction error:

$$\varepsilon_{ij} = \left\| \mathbf{x}_i^j - \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j \right\|^2 \tag{9}$$

where $\tilde{\mathbf{X}}_{ij} = [\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^{j-1}, \mathbf{x}_i^{j+1}, \ldots, \mathbf{x}_i^{n_i}]$ indicates $\mathbf{x}_i^j$ is excluded from $\mathbf{X}_i$ and $\boldsymbol{\beta}_i^j$ is the reconstruction coefficient vector obtained from LSE, i.e.,

$$\boldsymbol{\beta}_i^j = \left(\tilde{\mathbf{X}}_{ij}^T \tilde{\mathbf{X}}_{ij} \right)^{-1} \tilde{\mathbf{X}}_{ij}^T \mathbf{x}_i^j \tag{10}$$

For each $\mathbf{x}_i^j$, we can find its k heterogeneous nearest subspaces N_{ij}^e. The point-to-inter-class distance is defined as:

$$\eta_{ij} = \sum_{\mathbf{X}_m \in \mathrm{N}_{ij}^e} \frac{\left\| \mathbf{x}_i^j - \mathbf{X}_m \boldsymbol{\beta}_m^j \right\|^2}{\left| \mathrm{N}_{ij}^e \right|} \tag{11}$$

where $|\cdot|$ represents the cardinality of a set, $\mathbf{X}_m$ is one of the k heterogeneous nearest subspaces and $\boldsymbol{\beta}_m^j$ is the reconstruction coefficient vector.

Based on the point-to-intra-class and the point-to-inter-class distance, we can define the nearest subspace margin.

Definition 4 (*Nearest subspace margin*). The nearest subspace margin γ_i^j for $\mathbf{x}_i^j$ is defined as follows.

$$\gamma_i^j = \eta_{ij} - \varepsilon_{ij} \tag{12}$$

This margin measures the distances from $\mathbf{x}_i^j$ to its own class and similar heterogeneous classes. According to the classification rule of SRC and LRC, a large η_{ij} and a small ε_{ij}, i.e., a large γ_i^j, indicate $\mathbf{x}_i^j$ is easy to be classified correctly. Then the total nearest subspace margin for the whole data set is defined to be as follows

Definition 5 (*Total nearest subspace margin*). The total nearest subspace margin γ for all the samples is defined as:

$$\gamma = \sum_i \sum_j \gamma_i^j = \sum_i \sum_j \left(\eta_{ij} - \varepsilon_{ij} \right) = \sum_i \sum_j \eta_{ij} - \sum_i \sum_j \varepsilon_{ij} \tag{13}$$

Geometrically, $\sum_i \sum_j \eta_{ij}$ measures the class separability and $\sum_i \sum_j \varepsilon_{ij}$ measures the compactness of the intra-class samples. From the definitions, a large $\sum_i \sum_j \eta_{ij}$ and a small $\sum_i \sum_j \varepsilon_{ij}$, i.e., a large γ, will lead to a good separability.

3.3 Linear Feature Extraction

When performing dimensionality reduction, we aim to find a mapping $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d] \in \mathbb{R}^{n \times d}$ from the original space to some feature space such that γ is maximized after the transformation, i.e.,

$$\mathbf{W} = \arg\max_{\mathbf{W}} \gamma\left(\mathbf{W}\right) \tag{14}$$

Let $\mathbf{y}_i^j = \mathbf{W}^T \mathbf{x}_i^j$ be the image of $\mathbf{x}_i^j$ in the projected subspace. Then

$$\begin{aligned}
&\sum_i \sum_j \sum_{\mathbf{X}_m \in \mathrm{N}_{ij}^e} \frac{\left\| \mathbf{y}_i^j - \mathbf{Y}_m \boldsymbol{\beta}_m^j \right\|^2}{\left| \mathrm{N}_{ij}^e \right|} \\
&= \sum_i \sum_j \sum_{\mathbf{X}_m \in \mathrm{N}_{ij}^e} \frac{\left\| \mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \mathbf{X}_m \boldsymbol{\beta}_m^j \right\|^2}{\left| \mathrm{N}_{ij}^e \right|} \\
&= \sum_i \sum_j \sum_{\mathbf{X}_m \in \mathrm{N}_{ij}^e} \frac{\left(\mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \mathbf{X}_m \boldsymbol{\beta}_m^j \right)^T \left(\mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \mathbf{X}_m \boldsymbol{\beta}_m^j \right)}{\left| \mathrm{N}_{ij}^e \right|} \\
&= tr \left(\sum_i \sum_j \sum_{\mathbf{X}_m \in \mathrm{N}_{ij}^e} \frac{\left(\mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \mathbf{X}_m \boldsymbol{\beta}_m^j \right) \left(\mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \mathbf{X}_m \boldsymbol{\beta}_m^j \right)^T}{\left| \mathrm{N}_{ij}^e \right|} \right) \\
&= tr \left(\mathbf{W}^T \left\{ \sum_i \sum_j \sum_{\mathbf{X}_m \in \mathrm{N}_{ij}^e} \frac{\left(\mathbf{x}_i^j - \mathbf{X}_m \boldsymbol{\beta}_m^j \right) \left(\mathbf{x}_i^j - \mathbf{X}_m \boldsymbol{\beta}_m^j \right)^T}{\left| \mathrm{N}_{ij}^e \right|} \right\} \mathbf{W} \right) \\
&= tr \left(\mathbf{W}^T \mathbf{S}_b^R \mathbf{W} \right)
\end{aligned}$$

where $tr\left(\cdot\right)$ is the notation of trace operator and

$$\mathbf{S}_b^R = \sum_i \sum_j \sum_{\mathbf{X}_m \in \mathrm{N}_{ij}^e} \frac{\left(\mathbf{x}_i^j - \mathbf{X}_m \boldsymbol{\beta}_m^j \right) \left(\mathbf{x}_i^j - \mathbf{X}_m \boldsymbol{\beta}_m^j \right)^T}{\left| \mathrm{N}_{ij}^e \right|} \tag{15}$$

Similarly,

$$\begin{aligned}
&\sum_i \sum_j \left\| \mathbf{y}_i^j - \tilde{\mathbf{Y}}_{\mathbf{ij}} \boldsymbol{\beta}_i^j \right\|^2 \\
&= \sum_i \sum_j \left\| \mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j \right\|^2 \\
&= tr \left(\sum_i \sum_j \left(\mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j \right) \left(\mathbf{W}^T \mathbf{x}_i^j - \mathbf{W}^T \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j \right)^T \right)
\end{aligned}$$

$$
\begin{aligned}
&= tr\left(\mathbf{W}^T\left\{\sum_i\sum_j\left(\mathbf{x}_i^j-\tilde{\mathbf{X}}_{ij}\boldsymbol{\beta}_i^j\right)\left(\mathbf{x}_i^j-\tilde{\mathbf{X}}_{ij}\boldsymbol{\beta}_i^j\right)^T\right\}\mathbf{W}\right)\\
&= tr\left(\mathbf{W}^T\mathbf{S}_w^R\mathbf{W}\right)
\end{aligned}
$$

where

$$\mathbf{S}_w^R=\sum_i\sum_j\left(\mathbf{x}_i^j-\tilde{\mathbf{X}}_{ij}\boldsymbol{\beta}_i^j\right)\left(\mathbf{x}_i^j-\tilde{\mathbf{X}}_{ij}\boldsymbol{\beta}_i^j\right)^T \tag{16}$$

The total nearest subspace margin in the projected subspace changes to:

$$\gamma\left(\mathbf{W}\right)=tr\left(\mathbf{W}^T\left(\mathbf{S}_b^R-\mathbf{S}_w^R\right)\mathbf{W}\right)=\sum_{k=1}^{d}\mathbf{w}_k^T\left(\mathbf{S}_b^R-\mathbf{S}_w^R\right)\mathbf{w}_k \tag{17}$$

To eliminate the freedom, we add the constraint $\mathbf{w}_k^T\mathbf{w}_k=1$.

Thus the goal of MNSMC is to solve the following optimization problem:

$$
\begin{aligned}
&\max\sum_{k=1}^{d}\mathbf{w}_k^T\left(\mathbf{S}_b^R-\mathbf{S}_w^R\right)\mathbf{w}_k\\
&\text{subject to }\mathbf{w}_k^T\mathbf{w}_k=1
\end{aligned} \tag{18}
$$

It is easy to prove the optimal solution $\mathbf{W}$ is composed of the d eigenvectors corresponding to the d largest eigenvalues of Eq. (19).

$$\left(\mathbf{S}_b^R-\mathbf{S}_w^R\right)\mathbf{w}=\lambda\mathbf{w} \tag{19}$$

Compared with LDA, the proposed method does not needs to compute the inverse of $\mathbf{S}_w^R$. Therefore, MNSMC avoids the small sample size (SSS) problem and can extract more features [9].

To show the effectiveness of MNSMC, we recalculate the reconstruction errors of the four images as shown in Fig. 1 in the MNSMC's subspace and illustrate the results in Fig. 2. As can be seen in Fig. 1, only one face image has the smallest intra-class reconstruction error. But in Fig. 2, we can find the intra-class reconstruction errors are significantly smaller and three face images in Fig. 2 have the minimum intra-class reconstruction errors. The experimental results indicate that MNSMC can reduce the impact of the illuminations and improve the performances of SRC and LRC.

Note that, in Eq. (10), when the training number n_i of the ith class is larger than the dimension of $\mathbf{x}_i^j$, the matrix $\tilde{\mathbf{X}}_{ij}^T\tilde{\mathbf{X}}_{ij}$ is singular. Thus the inverse of $\tilde{\mathbf{X}}_{ij}^T\tilde{\mathbf{X}}_{ij}$ can not be calculated directly. In this case, we can apply ridge regression (RR) [22,23] to solve the singularity problem. Different from LSE, RR aims to minimize the following cost function:

$$\min\left(\left\|\mathbf{x}_i^j-\tilde{\mathbf{X}}_{ij}\boldsymbol{\beta}_i^j\right\|^2+\lambda\left\|\boldsymbol{\beta}_i^j\right\|^2\right) \tag{20}$$

where λ is a positive factor to reduce the solution space.

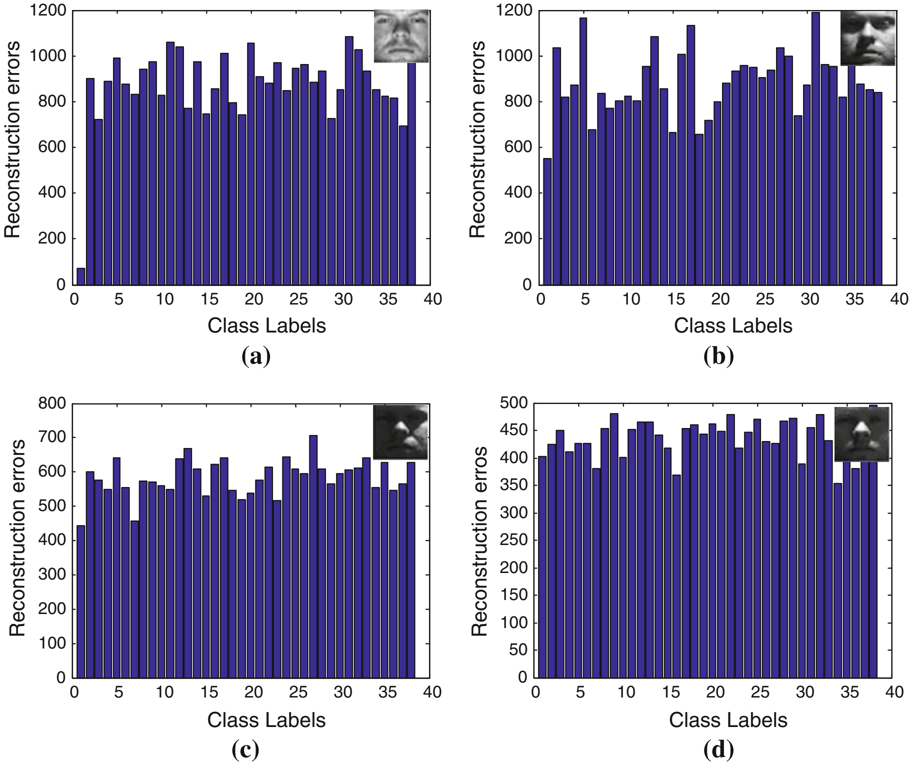


Fig. 2 The reconstruction errors of four *images* of the first class under the different illumination conditions in the MNSMC's subspace

The cost function can be rewritten as:

$$
\begin{aligned}
& \left\| \mathbf{x}_i^j - \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j \right\|^2 + \lambda \left\| \boldsymbol{\beta}_i^j \right\|^2 \\
&= \left(\mathbf{x}_i^j - \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j \right)^T \left(\mathbf{x}_i^j - \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j \right) + \lambda \left(\boldsymbol{\beta}_i^j \right)^T \boldsymbol{\beta}_i^j \\
&= \left(\left(\mathbf{x}_i^j \right)^T \mathbf{x}_i^j + \left(\boldsymbol{\beta}_i^j \right)^T \tilde{\mathbf{X}}_{ij}^T \tilde{\mathbf{X}}_{ij} \boldsymbol{\beta}_i^j - 2 \left(\boldsymbol{\beta}_i^j \right)^T \tilde{\mathbf{X}}_{ij}^T \mathbf{x}_i^j \right) + \lambda \left(\boldsymbol{\beta}_i^j \right)^T \boldsymbol{\beta}_i^j
\end{aligned}
$$

Taking derivatives and equaling them to zero, then the optimal solution of the cost function can be obtained as follows.

$$
\boldsymbol{\beta}_i^j = \left(\tilde{\mathbf{X}}_{ij}^T \tilde{\mathbf{X}}_{ij} + \lambda \mathbf{I} \right)^{-1} \tilde{\mathbf{X}}_{ij}^T \mathbf{x}_i^j \tag{21}
$$

where $\mathbf{I}$ is the identity matrix. It is easy to prove $\tilde{\mathbf{X}}_{ij}^T \tilde{\mathbf{X}}_{ij} + \lambda \mathbf{I}$ is nonsingular.

As can be seen in Eq. (6), similar to MNSMC, LRC also suffers from the singularity problem. Using RR instead of LSE, Eq. (6) can be rewritten as:

$$
\hat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i^T \mathbf{X}_i + \lambda \mathbf{I} \right)^{-1} \mathbf{X}_i^T \mathbf{y} \tag{22}
$$

Table 1 The algorithm of MNSMC

Input: Column sample matrix $\mathbf{X}$, the nearest subspace size k
Output: Transform matrix $\mathbf{W}$
Step 1: Construct $\mathbf{S}_w^R$ and $\mathbf{S}_b^R$ using $\mathbf{X}$.
Step 2: Solve the generalized eigenvectors of $\left(\mathbf{S}_b^R - \mathbf{S}_w^R\right)\mathbf{w} = \lambda\mathbf{w}$ and construct $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d\}$ corresponding to the d largest eigenvalues.
Step 3: Output $\mathbf{W}$.

Table 2 The details of the four databases

Database	Size	Number of classes	Number of samples per class	Number of training sample per class
YALE-B	32 × 32	38	64	10
AR	50 × 40	120	26	5
FKP	55 × 110	100	12	5
CENPARMI	121	10	600	200

Based on RR, we extend LRC to any cases without limits. And the modified LRC is called ridge regression classification (RRC) in this paper.

3.4 The Algorithm of MNSMC

The main steps of MNSMC is summarized in Table 1.

4 Experiments

To evaluate the performance of the proposed method, we compare it with three feature extraction methods, i.e., principal component analysis (PCA) [24], LDA and (MMC) [25] over three classifiers, i.e., nearest neighbor classifier (NNC) [26], SRC and LRC, on four well-known databases. The details of the databases are summarized in Table 2. As a baseline, we directly employ NNC, SRC and LRC to classify the raw data. Then we investigate whether MNSMC can enhance the performances of SRC and LRC. We also compare MNSMC with PCA, LDA and MMC to find which feature subspace is more suitable for SRC and LRC. The NNC is used to compare with SCR and LRC to determine which classifier is more suitable for MNSMC. Note that LDA can extract at most $c - 1$ features due to the characteristic (c is the total number of the classes). Therefore, the dimension of LDA is limited to a value in the figures.

4.1 Parameter Selection

For efficiency, on the YALE-B, AR and FKP databases, PCA is first applied to reduce the dimensionality. Then the experiments are performed on the 150-dimensional PCA subspaces. On the YALE-B, AR and FKP databases, there is only one model parameter, i.e., k heterogeneous nearest subspace. And on the CENPARMI database, the parameter λ is introduced to overcome the singularity problem. In the experiments, the values of the parameters are empirically set as in Table 3.

4.2 Face Recognition

The YALE-B database [27–29] consists of 2432 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.

Table 3 The parameter settings of the four databases

Database	PCA dimensions	k	λ
YALE-B	150	1	None
AR	150	1	None
FKP	150	10	None
CENPARMI	120	1	0.01

Table 4 The average maximal recognition rates on the YALE-B database

	Baseline	PCA	LDA	MMC	MNSMC
NNC	44.3 ± 1.0	43.1 ± 1.0	**80.0 ± 1.4**	79.7 ± 1.4	70.8 ± 2.0
SRC	86.2 ± 1.1	87.7 ± 0.8	85.4 ± 1.2	85.2 ± 1.3	**88.3 ± 1.1**
LRC	81.7 ± 1.0	83.2 ± 1.3	84.1 ± 1.1	87.6 ± 1.0	**91.1 ± 0.9**

Significance of bold are the highest recognition rates of the corresponding classifiers

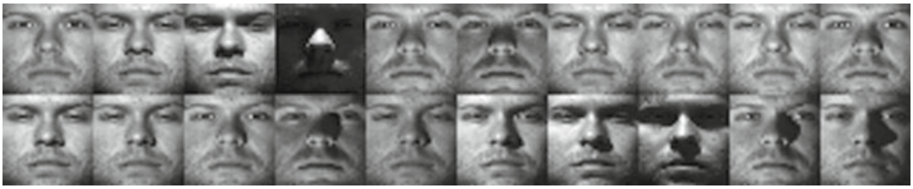


Fig. 3 Sample images of one person from the YALE-B face database

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. Sample images of one person from the YALE-B face database are shown in Fig. 3. We randomly choose 10 samples images from all the subsets and the rest sample images are used for test. This procedure is repeated for 50 times. The average recognition rates and the corresponding standard deviation are indicated in Table 4. And the recognition rates versus the dimensions are illustrated in Figs. 4, 5 and 6.

The AR face database [30,31] contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions and occlusions. The pictures of most persons were taken in two sessions (separated by two weeks). Each section contains 13 color images and 120 individual s (65 men and 55 women) participated in both sessions. The images of these 120 individuals were selected and used in our experiment. We manually cropped the face portion of the image and then normalized it to 50 × 40 pixels. Sample images of one person from the AR face database are shown in Fig. 7. We randomly choose 5 samples images from each person and the rest sample images are used for test. This procedure is repeated for 50 times. The average recognition rates and the corresponding standard deviation are indicated in Table 5.

And the recognition rates versus the dimensions are illustrated in Figs. 8, 9 and 10.

4.3 Finger Knuckle Print Recognition

In PolyU FKP database [32–34], FKP images were collected from 165 volunteers, including 125 males and 40 females. Among them, 143 subjects were 20–30 years old and the

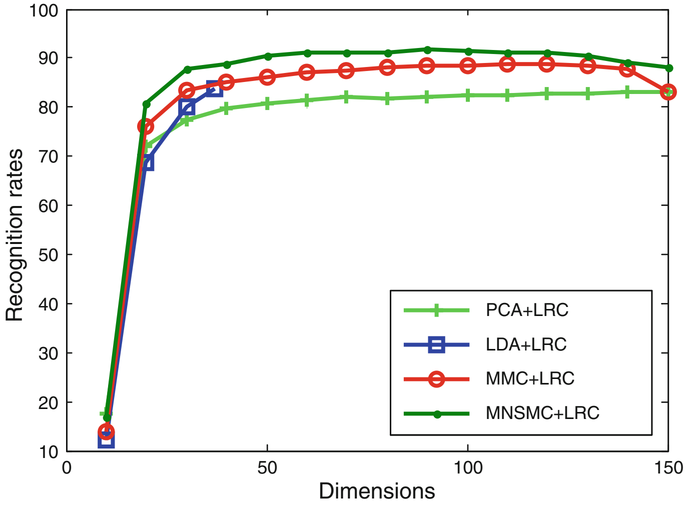


Fig. 4 The recognition rates of 4 methods plus LRC on the YALE-B database

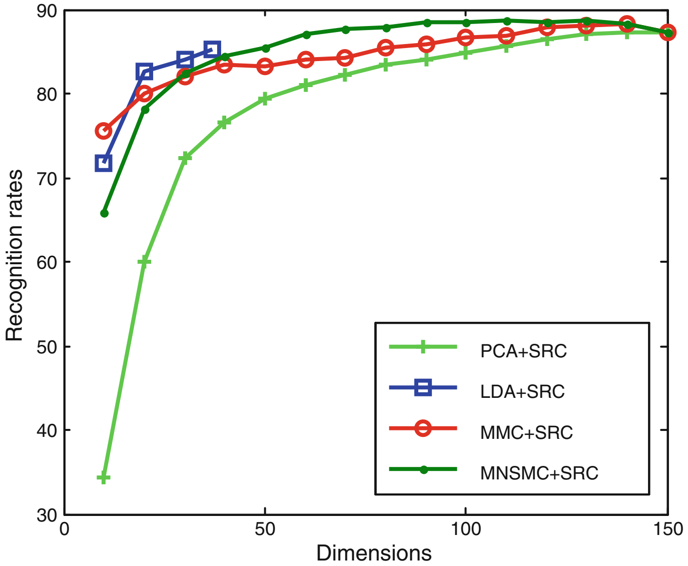


Fig. 5 The recognition rates of 4 methods plus SRC on the YALE-B database

Table 5 The average maximal recognition rates on the AR database

	Baseline	PCA	LDA	MMC	MNSMC
NNC	59.6 ± 1.3	57.4 ± 1.1	**88.6 ± 1.0**	84.5 ± 1.1	85.4 ± 1.3
SRC	87.5 ± 0.7	91.8 ± 0.8	90.9 ± 0.7	92.0 ± 0.9	**92.4 ± 0.5**
LRC	74.8 ± 1.3	72.5 ± 0.4	90.4 ± 1.1	88.4 ± 1.0	**92.5 ± 0.9**

Significance of bold are the highest recognition rates of the corresponding classifiers

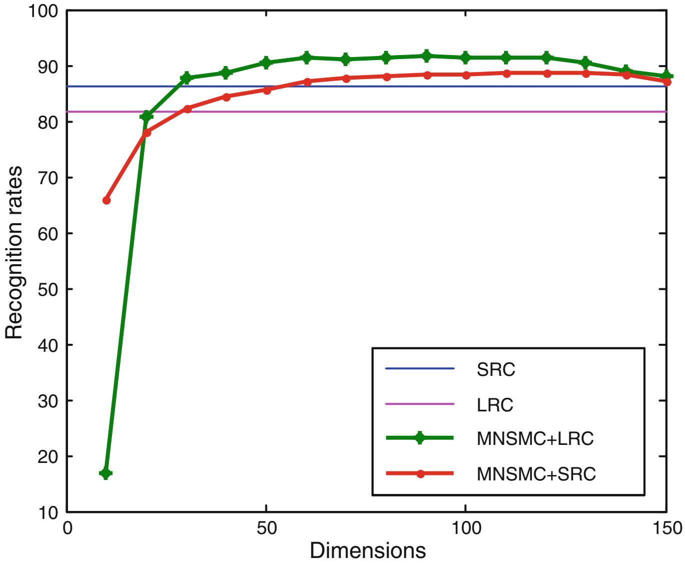


Fig. 6 The recognition rates of MNSMC plus SRC and LRC versus the baselines of SRC and LRC on the YALE-B database

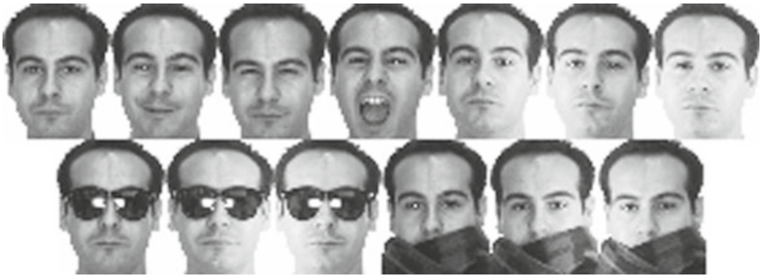


Fig. 7 Sample images of one person from the AR face database

others were 30–50 years old. The samples were collected in two separate sessions. In each session, the subject was asked to provide six 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 four fingers were collected from each subject. In total, the database contains 7,920 images from 660 different fingers. The average time interval between the first and the second sessions was about 25 days. The maximum and minimum time intervals were 96 days and 14 days respectively. All the samples in the database are histogram equalized and resized to 55 × 110. Sample images of one person from the FKP database are shown in Fig. 11. We randomly choose five samples images from each person and the rest sample images are used for test. This procedure is repeated for 50 times. The average recognition rates and the corresponding standard deviation are indicated in Table 6. And the recognition rates versus the dimensions are illustrated in Figs. 12, 13 and 14.

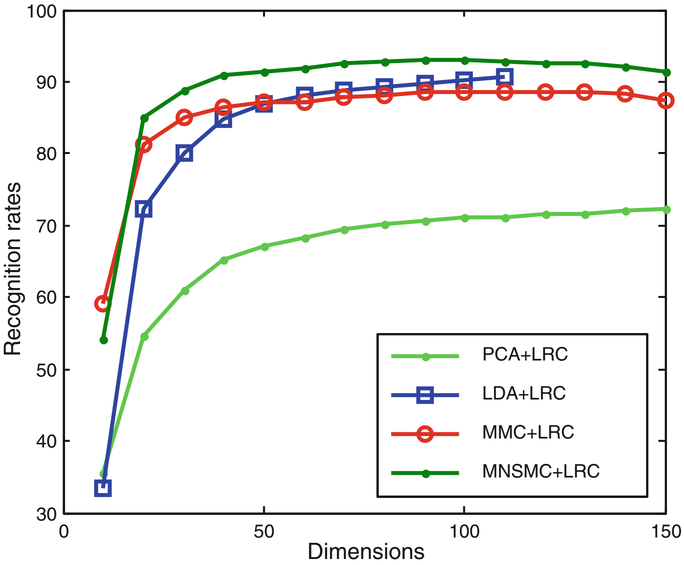


Fig. 8 The recognition rates of 4 methods plus LRC on the AR database

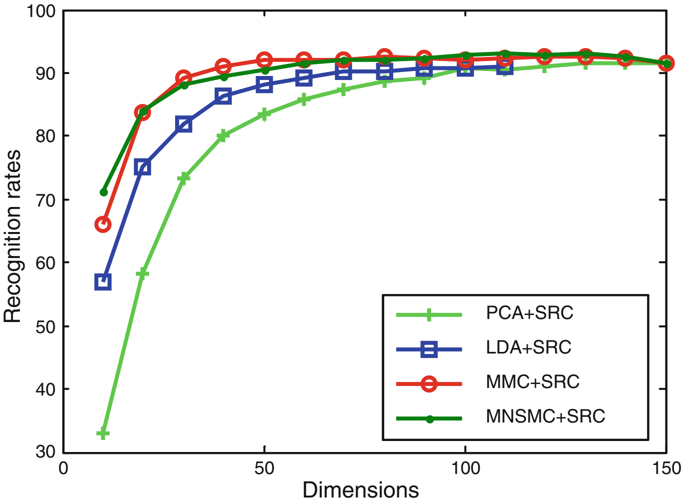


Fig. 9 The recognition rates of 4 methods plus SRC on the AR database

Table 6 The average maximal recognition rates on the FKP database

	Baseline	PCA	LDA	MMC	MNSMC
NNC	86.6 ± 1.4	88.7 ± 1.6	83.5 ± 1.6	88.7 ± 1.6	**92.8 ± 1.5**
SRC	86.6 ± 1.1	91.7 ± 1.2	91.4 ± 1.2	93.0 ± 0.8	**93.8 ± 0.9**
LRC	86.1 ± 1.2	91.5 ± 1.4	92.4 ± 1.3	91.3 ± 1.2	**94.5 ± 1.2**

Significance of bold are the highest recognition rates of the corresponding classifiers

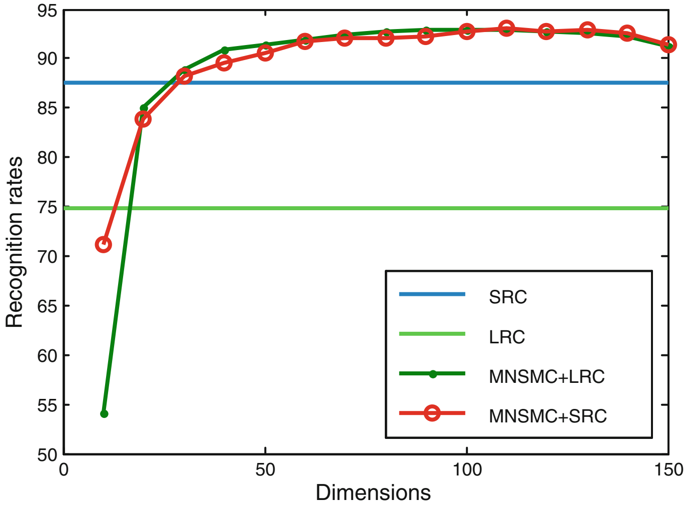


Fig. 10 The recognition rates of MNSMC plus SRC and LRC versus the baselines of SRC and LRC on the AR database

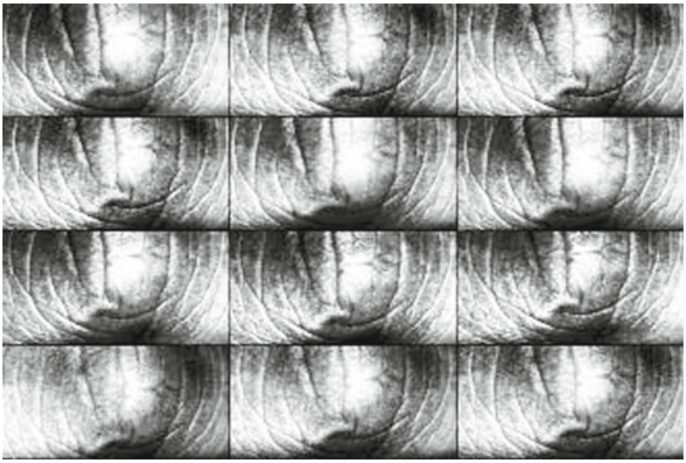


Fig. 11 Sample images of the right index finger from one individual

4.4 Handwritten Numeral Recognition

The Concordia University CENPARMI handwritten numeral database [35] is used to test the performance of MNSMC plus RRC. The database contains 10 numeral classes and each class has 600 samples. In our experiment, we randomly choose 200 samples of each class for training, the remaining 400 samples for testing. Thus, the total number of training samples is 2,000 while the total number of testing samples is 4000. Since the dimensionality of one digit (121) is less than training samples (200) of one class, LRC fails under this circumstance. So RRC is employed in this experiment. The recognition rates versus the dimensions are illustrated in Figs. 15, 16 and 17.

Table 7 The recognition rates on the four subsets using different methods and classifiers

	Subset2			Subset3			Subset4			Subset5		
	NNC	SRC	LRC	NNC	SRC	LRC	NNC	SRC	LRC	NNC	SRC	LRC
PCA	90.1	100	100	41.5	98.6	98.6	15.8	74.0	76.2	8.2	26.1	30.4
LDA	**99.8**	100	100	**94.3**	99.8	98.6	**22.8**	56.3	62.8	7.6	20.7	25.6
MMC	95.8	100	100	66.2	99.6	99.6	12.5	61.4	67.2	7.4	17.6	22.3
MNSMC	99.3	**100**	**100**	73.8	**100**	**100**	16.8	**83.4**	**88.3**	**10.2**	**33.6**	**39.8**

Significance of bold are the highest recognition rates of the corresponding classifiers

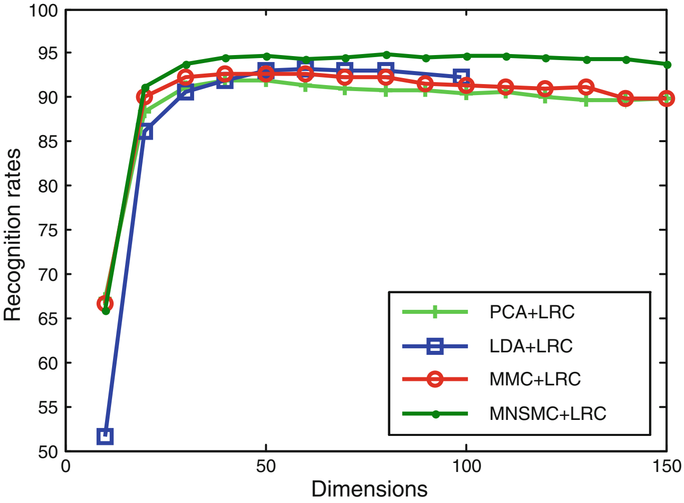


Fig. 12 The recognition rates of 4 methods plus LRC on the FKP database

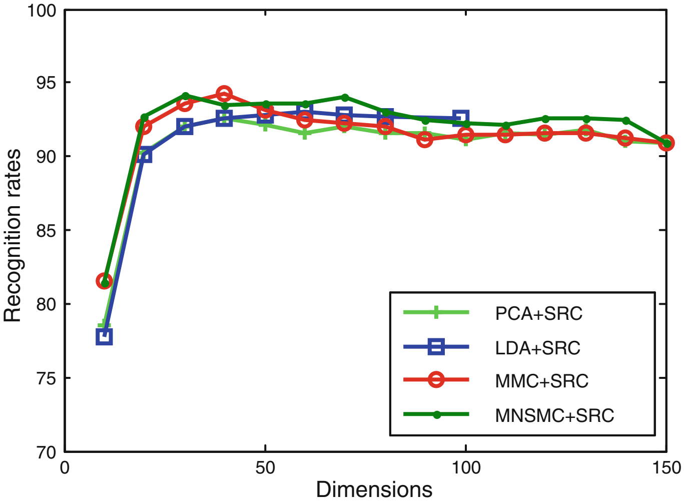


Fig. 13 The recognition rates of 4 methods plus SRC on the FKP database

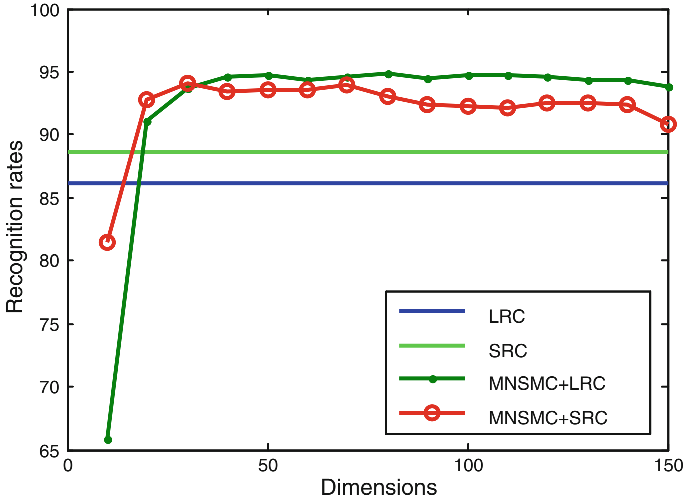


Fig. 14 The recognition rates of MNSMC plus SRC and LRC versus the baselines of SRC and LRC on the FKP database

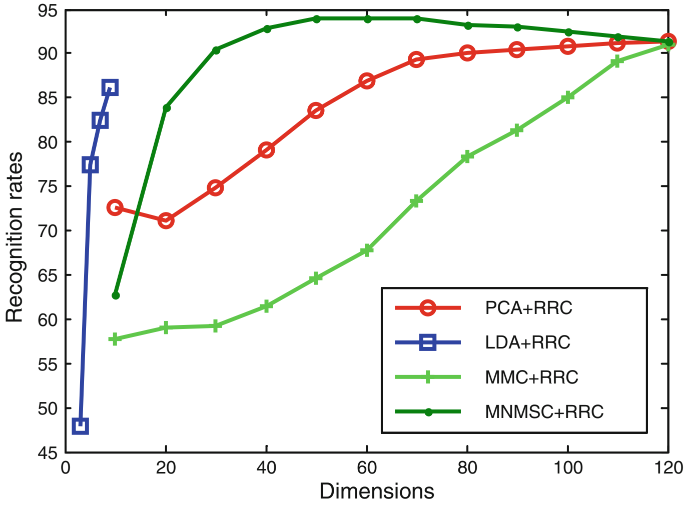


Fig. 15 The recognition rates of 4 methods plus LRC on the CENPARMI database

4.5 Face Recognition Under illumination and noise conditions

Further experiments on the YALE-B face database are conducted to investigate the performance of MNSMC under illumination and noise condition.

The YALE-B was divided into five subsets according to the illumination directions. Sample images of one person from the subsets are shown in Fig. 18. We follow the evaluation protocol as reported in [28,36]. Training is conducted using subset 1 and the system is validated on the remaining subsets. The experimental results are listed in Table 7.

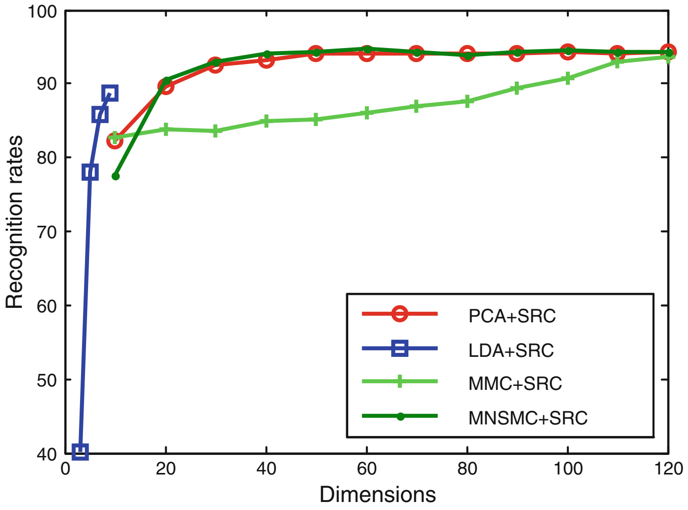


Fig. 16 The recognition rates of 4 methods plus SRC on the CENPARMI database

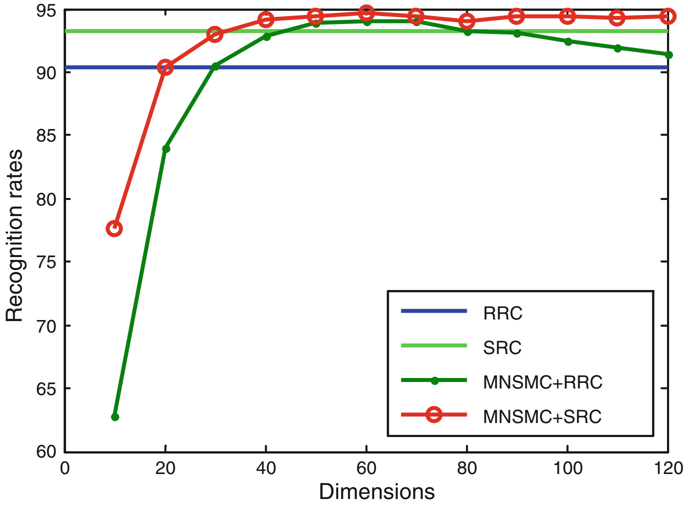


Fig. 17 The recognition rates of MNSMC plus SRC and LRC versus the baselines of SRC and LRC on the CENPARMI database

In the next set of experiments we contaminate the face images in the subset2 with salt and pepper noise [37]. Figure 19 reflects face images distorted with various degrees of salt and pepper noise. The experimental results can be found in Table 8.

4.6 Discussions

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

(1) Since the Yale-B and AR databases contain illuminations, occlusions and expressions, the recognition rates are not quite good when classifiers are directly applied on the raw

Table 8 The recognition rates on the subsets2 using different methods and classifications

Density	20 %			40 %			60 %			80 %		
	NNC	SRC	LRC	NNC	SRC	LRC	NNC	SRC	LRC	NNC	SRC	LRC
PCA	90.0	100	100	80.3	95.3	93.2	43.6	76.4	51.1	18.2	20.1	16.2
LDA	**99.1**	100	100	86.6	98.7	96.4	40.8	85.1	61.5	11.4	23.3	15.8
MMC	97.4	100	100	88.3	99.2	96.6	42.6	83.7	56.8	13.8	21.2	15.3
MNSMC	98.3	**100**	**100**	**92.1**	**100**	**98.2**	**50.7**	**87.4**	**65.2**	**23.7**	**26.5**	**25.8**

Significance of bold are the highest recognition rates of the corresponding classifiers

Table 9 The average maximal recognition rates on the CENPARMI database

	Baseline	PCA	LDA	MMC	MNSMC
NNC	88.3 ± 0.3	88.3 ± 0.3	87.4 ± 0.4	90.0 ± 0.4	**92.3 ± 0.5**
SRC	93.2 ± 0.3	93.0 ± 0.8	88.7 ± 0.7	92.8 ± 0.7	**94.5 ± 0.4**
RRC	90.4 ± 0.4	90.9 ± 0.4	86.2 ± 0.4	91.7 ± 0.6	**94.2 ± 0.5**

Significance of bold are the highest recognition rates of the corresponding classifiers

data. More importantly, we find that not all the feature extraction methods are helpful to the classifiers. As can be seen in Tables 4 and 9, the recognition rates of LDA plus

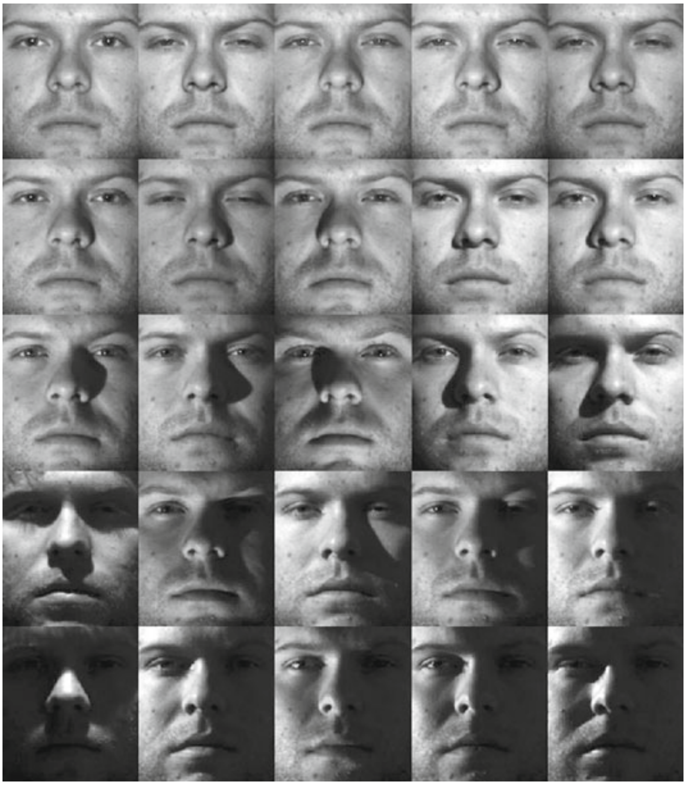


Fig. 18 Each *row* represents typical *images* from subsets 1, 2, 3, 4 and 5, respectively

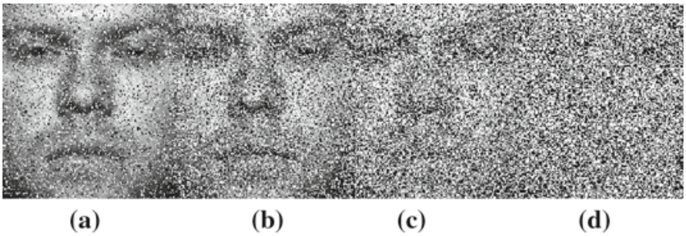


Fig. 19 Face image with **a** 20 %, **b** 40 %, **c** 60 %, **d** 80 % salt and pepper noise density

SRC are lower than the corresponding baselines. That means selecting an appropriate feature extraction method is very important to a specific classifier.

(2) Since MNSMC is designed according to the classification rule of SRC and LRC, it matches SRC and LRC perfectly. As can be seen in Figs. 6, 10, 14, and 17, the recognition rates of MNSMC plus SRC or LRC are consistently higher than the corresponding baselines. Meanwhile, we observe MNSMC plus LRC outperforms MNSMC plus SRC on the YALE-B, AR and FKP databases. As we introduced above, SRC and LRC are based on different reconstruction strategies. In the proposed method, MNSMC shares the same reconstruction strategy with LRC. Thus, MNSMC plus LRC performs better in most of the experiments.

(3) Compared with PCA, LDA and MMC, MNSMC is the best feature extraction method for SRC and LRC. As can be seen in Figs. 4, 5, 8, 9, 12, 13, 15, and 16, MNSMC plus SRC and LRC consistently outperform other combinations. As MNSMC considers the reconstruction error, it perfectly fits the classification rule of SRC and LRC. Therefore, MNSMC performs better than other feature extraction methods.

(4) The experimental results also indicate, compared with SRC and LRC, NNC is not very suitable for MNSMC. Technically, NNC assumes a sample and its nearest neighbor are in the same class. However, MNSMC considers the nearest subspace rather than the nearest neighbor. Therefore, MNSMC may not match NNC perfectly.

5 Conclusions

In this paper, according to the classification rule of SRC and LRC, a new feature extractor based on MNSMC is proposed. By maximizing the inter-class reconstruction error and minimizing the intra-class reconstruction error simultaneously, the proposed method improves the performances of SRC and LRC significantly. Our method avoids the SSS problem and can extract more features than LDA. Moreover, based on RR, we develop RRC to overcome the potential singular problem of LRC. The experimental results on the YALE-B, AR, FKP and the CENPARMI handwritten numeral databases show the effectiveness of the proposed method.

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

References

1. Li SZ (1998) Face recognition based on nearest linear combinations. In: Proceedings of IEEE international conference on computer vision and pattern recognition, pp 839–844

2. Li SZ, Lu J (1999) Face recognition using nearest feature line method. IEEE Trans Neural Netw 10(2):439–443
3. Li SZ, Chan KL, Wang CL (2000) Performance evaluation of the nearest feature line method in image classification and retrieval. IEEE Trans Pattern Anal Mach Intell 22(11):1335–1339
4. Chien J-T, Wu C-C (2002) Discriminant wavelet faces and nearest feature classifiers for face recognition. IEEE Trans Pattern Anal Mach Intell 24(12):1644–1649
5. Wright J, Yang A, Ganesh A, Sastry S, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227
6. Naseem I, Togneri R, Bennamoun M (2010) Linear regression for face recognition. IEEE Trans Pattern Anal Mach Intell 32(11):2106–2112
7. Jolliffe IT (1986) Principal component analysis. Springer, New York
8. Belhumeur PN, Hespanda J, Kiregeman D (1997) Eigenfaces vs Fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
9. Li H, Jiang T, Zhang K (2003) Efficient and robust feature extraction by maximum margin criterion. In: Proceedings of advances in neural information processing systems, pp 97–104
10. He X, Yan S, Hu Y, Niyogi P, Zhang H-J (2005) Face recognition using Laplacian faces. IEEE Trans Pattern Anal Mach Intell 27(3):328–340
11. He X, Cai Deng, Yan S, Zhang HJ (2005a) Neighborhood preserving embedding. In: Proceedings of the 10th IEEE international conference on computer vision, pp 1208-1213
12. Yan S, Xu D, Zhang B, Zhang H, Yang Q, Lin S (2007) Graph embedding and extension: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51
13. Yang J, Zhang D, Yang J, Niu B (2007) Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics. IEEE Trans Pattern Anal Mach Intell 29(4):650–664
14. Chen H-T, Chang H-W, Liu T-L (2005) Local discriminant embedding and its variants. In: IEEE conference on computer vision and pattern recognition (CVPR 2005), pp 846–853
15. Wang F, Wang X, Zhang D, Zhang CS, Li T (2009) marginFace: a novel face recognition method by average neighborhood margin maximization. Pattern Recognit 42(11):2863–2875
16. Candès E, Tao T (2006) Near optimal signal recovery from random projections: universal encoding strategies?. IEEE Trans Inf Theory 52(12):5406–5425
17. Donoho D (2006) For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Commun Pure Appl Math 59(6):797–829
18. Candès E, Romberg J, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun Pure Appl Math 59(8):1207–1223
19. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference and prediction. Springer, New York
20. Seber GAF (2003) Linear regression analysis. Wiley-Interscience, Hoboken
21. Ryan TP (1997) Modern regression methods. Wiley-Interscience, Hoboken
22. Hoerl AE, Kennard RW (1970) Ridge regression: applications to nonorthogonal problems. Technometrics 12(1):69–82
23. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67
24. Jolliffe IT (1986) Principal component analysis. Springer, New York
25. Li H, Jiang T, Zhang K (2003) Efficient and robust feature extraction by maximum margin criterion. In: Proceedings of Advances in Neural Information Processing Systems, pp 97–104
26. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
27. Lee K, Ho J, Kriegman D (2005) Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans Pattern Anal Mach Intell 27(5):684–698
28. Georghiades A, Belhumeur P, Kriegman D (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660
29. The extended YALE-B database: http://www.zjucadcg.cn/dengcai/Data/FaceData.html
30. Martinez AM, enavente RB (1998) The AR Face Database. CVC Technical Report, no. 24
31. Martinez AM, enavente RB (2003) The AR Face Database. http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html
32. Zhang L, Zhang L, Zhang D, Zhu H (2010) Online finger-knuckle-print verification for personal authentication. Pattern Recognit 43(7):2560–2571
33. Zhang L, Zhang L, Zhang D (2009) Finger-knuckle-print: a new biometric identifier. In: Proceedings of the IEEE international conference on image processing
34. The FKP database. http://www.comp.polyu.edu.hk/~biometrics/FKP.htm

35. Liao SX, Pawlak M (1996) On image analysis by moments. IEEE Trans Pattern Anal Mach Intell 18(3):254–266
36. Weilong C, Joo ME, Wu S (2006) Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. IEEE Trans Syst Man Cybernet 36(2):458–464
37. Gonzalez RC, Woods RE (2007) Digital image processing. Pearson Prentice Hall, Upper Saddle River