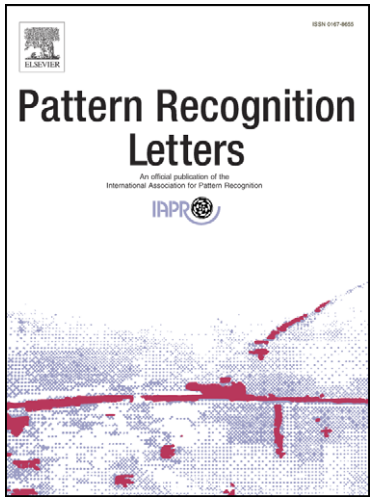


Push-Pull Marginal Discriminant Analysis for Feature Extraction

Zhenghong Gu, Jian Yang, Lei Zhang

PII: S0167-8655(10)00220-5
DOI: 10.1016/j.patrec.2010.07.001
Reference: PATREC 4913

To appear in: *Pattern Recognition Letters*

Received Date: 16 June 2009

Please cite this article as: Gu, Z., Yang, J., Zhang, L., Push-Pull Marginal Discriminant Analysis for Feature Extraction, *Pattern Recognition Letters* (2010), doi: 10.1016/j.patrec.2010.07.001

Push-Pull Marginal Discriminant Analysis for Feature Extraction

Zhenghong Gu, Jian Yang [a], Lei Zhang[b]

[a] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, P. R. China

[b] Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong

*Corresponding author: Tel.: +86-25-8431-7297; fax: +86-25-8431-5510.

E-mail: gzhnjust@gmail.com (Zhenghong Gu) csjyang@mail.njust.edu.cn (Jian Yang)

cslzhang@comp.polyu.edu.hk (Lei Zhang)

Abstract Marginal information is of great importance for classification. This paper presents a new nonparametric linear discriminant analysis method named Push-Pull marginal discriminant analysis (PPMDA), which takes full advantage of marginal information. For two-class cases, the idea of this method is to determine projected directions such that the marginal samples of one class are pushed away from the between-class marginal samples as far as possible and simultaneously pulled to the within-class samples as close as possible. This idea can be extended for multi-class cases and give rise to the PPMDA algorithm for feature extraction of multi-class problems. The proposed method is evaluated using the CENPARMI handwritten numeral database, the Extended Yale face database B and the ORL database. Experimental results show the effectiveness

of the proposed method and its advantage after performance over the state-of-the-art feature extraction methods.

Keywords Feature extraction, linear discriminant analysis, nonparametric methods, Classification

1. Introduction

Discriminant analysis is a popular tool for feature extraction and classification. Parametric discriminant analysis methods such as (Fisher, 1936; Belhumeur et al., 1997; Chen et al., 2005; Etemad and Chellappa, 1996; Etemad and Chellappa, 1997; Swets and Weng, 1996; Loog et al., 2004; Liu et al., 1992; Chen et al., 2000; Yu and Yang, 2001) rely on the assumption that the samples are normally distributed. However, if the distribution is non-normal, features extracted by such parametric version cannot be expected to accurately preserve any complex structure that might be needed for classification (Fukunaga et al. 1983).

To overcome the limitation of parametric methods, Fukunaga et al. (1983) presented a nonparametric discriminant analysis (NDA) method. The term *nonparametric* is not meant that this method completely lack parameters but it doesn't rely on any assumption of prior probability distribution. This method gives a nonparametric of definition between-class scatter matrix. However, it can only deal with two-class problems. Recently, Li et al. (2005) extended the definition of the nonparametric between-class scatter matrix to the multi-class cases and developed a method called nonparametric subspace analysis (NSA). It should be mentioned that the within-class scatter matrix in NSA is still of parametric version. Li et al. (2009) further improved NSA by introducing a nonparametric version of the within-class scatter matrix and then developed a method called nonparametric feature analysis (NFA). Qiu and Wu (2005) proposed a nonparametric margin maximum criterion (NMMC) which suggests an alternative extension of NDA by introducing a different nonparametric version of the within-class scatter matrix.

The NMMC method, relying on the within-class *farthest* neighbor in the construction of the within-class scatter matrix, may encounter the following problem: minimizing the distance between a point and its within-class *farthest* point does not make sense for classification if the farthest point is not on the margin at all. This paper presents a push-pull marginal discriminant analysis (PPMDA) to address the foregoing problem of NMMC. In the PPMDA method, for each sample point, we choose its corresponding within-class sample point to be the sample that is close to the margin and potentially the chosen sample contributes to the increase of the margin, rather than choose the within-class *farthest* sample which is sometimes meaningless for enlarging the margin. The proposed method can be unified under the graph framework (S. Yan, D. Xu et al.)

The remainder of this paper is organized as follows: Section 2 gives a review of LDA and existing nonparametric methods. Section 3 describes our push-pull marginal discriminant analysis. Experimental evaluation of the proposed method using the CENPARMI handwritten numeral database, the Extended Yale face database B and the ORL database are presented in Section 4. Finally, we give the conclusion in Section 5.

2. Related work

The problem can be simply stated as follows. Suppose there are L classes $\{C_1, C_2, ..., C_L\}$. The number of samples in class C_i is N_i $(i = 1, \cdots, L)$ and let $N = \sum_{i=1}^{L} N_i$. The purpose of discriminant analysis is to extract features which best separate the *L* classes by finding an optimal projection. These features are used for later classification.

2.1 Linear Discriminant Analysis

FLDA (Fisher 1936) is a classical linear discriminant analysis which is popular and powerful for face recognition (Duda and Hart, 1973). The parametric form of the scatter matrix of FLDA is based on the Gaussian distribution assumption. The between-class scatter matrix is defined as

$$S_B^{FLDA} = \sum_{i=1}^{L} N_i \left(m_i - m\right)\left(m_i - m\right)^T . \quad (1)$$

And the within-class scatter matrix is defined as

$$S_W^{FLDA} = \sum_{i=1}^{L} \sum_{l=1}^{N_i} \left(x_{il} - m_i\right)\left(x_{il} - m_i\right)^T , \quad (2)$$

where m_i is the mean vector of C_i, m is global mean vector. x_{il} is *l*-th pattern sample of C_i. If S_W is nonsingular, the optimal projection W_{opt} is chosen as the matrix with column vectors $\varphi_1,...,\varphi_d$ to maximize the ratio of the determinant of the between-class scatter matrix to that of within-class scatter matrix, i.e. ,

$$J\left(\varphi\right) = \frac{\varphi^T S_B^{FLDA} \varphi}{\varphi^T S_W^{FLDA} \varphi} . \quad (3)$$

In order to obtain a set of uncorrelated discriminant features, $\varphi_1,...,\varphi_d$ should be subject to the conjugate-orthogonal constraints (Jin et al. 2001). Specifically, W_{opt} is formed by *d* generalized eigenvectors of $S_B^{FLDA} X = \lambda S_W^{FLDA} X$ corresponding to its *d* largest eigenvalues.

There are three disadvantages of FLDA. First, FLDA is optimal in Bayes sense if all classes share the Gaussian distribution with the same covariance matrix and different means. Otherwise, its performance cannot be guaranteed. Second, the number of its features has an upper limit of *L*-1 since the rank of the between-class scatter matrix is at

most *L*-1. Third, the features extracted by such scatter matrices fail to preserve marginal structures which are proven to be important for classification (Fukunaga et al. 1983).

2.2. Nonparametric Discriminant Analysis

Nonparametric discriminant analysis (Fukunaga et al. 1983) is presented to overcome the first two disadvantages of FLDA by introducing a nonparametric version of the between-class scatter matrix by *k*-nearest neighbor (kNN) techniques. In the nonparametric discriminant analysis, the between-class scatter matrix is defined as

$$S_B^{NDA} = \sum_{l=1}^{N_i} w(i,j,l)\left(x_{il} - m_{jl}\right)\left(x_{il} - m_{jl}\right)^T + \sum_{l=1}^{N_j} w(j,i,l)\left(x_{jl} - m_{il}\right)\left(x_{jl} - m_{il}\right)^T , \quad (4)$$

where x_{il} denotes the *l*-th pattern sample of C_i and m_{jl} is the local mean of x_{il} in C_j. We call m_{jl} the C_j-local mean of x_{il}. m_{jl} is defined as

$$m_{jl} = \frac{1}{k}\sum_{p=1}^{k} y_{jl}^p , \quad (5)$$

where y_{jl}^p is the *p*-th nearest neighbor of the pattern sample x_{il} from C_j, $w(i,j,l)$ is weighting function defined as

$$w(i,j,l) = \frac{\min\left\{d^\alpha\left(x_{il}, m_{il}\right), d^\alpha\left(x_{il}, m_{jl}\right)\right\}}{d^\alpha\left(x_{il}, m_{il}\right) + d^\alpha\left(x_{il}, m_{jl}\right)} , \quad (6)$$

where α is a parameter ranging from zero to infinity. Samples which are far away from the margin tend to have larger magnitudes. These large magnitudes exert a considerable influence on between-class scatter matrix and may distort the marginal information. The weighting function is used to emphasize the sample near the margin (The weighting functions of NSA, NFA and NMMC are similar, as we will see below). But the nonparametric discriminant analysis is only suitable for two-class problems. Li et al.

(2005) extended the nonparametric discriminant analysis for dealing with multi-class problems. In their nonparametric subspace analysis (NSA), the nonparametric between-class scatter matrix is defined as follows:

$$S_B^{NSA} = \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{l=1}^{N_i} w(i,j,l)\left(x_{il} - m_{jl}\right)\left(x_{il} - m_{jl}\right)^T. \quad (7)$$

We can regard NSA as a semi-parametric method, since the within-class scatter matrix of NSA is of parametric version, which is the same as FLDA. Thus, this method still encounters the singularity of S_W when the training sample size is small. To avoid this singularity, nonparametric feature analysis (NFA) and nonparametric margin maximum criterion (NMMC) were presented. In these two methods, two nonparametric versions of the within-class scatter matrix are given respectively.

2.3 Nonparametric Feature Analysis

Li et al. (2009) developed an enhanced nonparametric method called Nonparametric Feature Analysis (NFA) by introducing a nonparametric version of within-class scatter matrix which is generally full of rank. This method, therefore, can overcome the singularity of within-class scatter matrix. In NFA, the nonparametric between-class scatter and within-class scatter matrices are respectively defined as follows

$$S_B^{NFA} = \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{p=1}^{k_2} \sum_{l=1}^{N_i} w(i,j,p,l)\left(x_{il} - y_{jl}^p\right)\left(x_{il} - y_{jl}^p\right)^T, \quad (8)$$

$$S_W^{NFA} = \sum_{i=1}^{L} \sum_{p=1}^{k_1} \sum_{l=1}^{N_i} \left(x_{il} - y_{il}^p\right)\left(x_{il} - y_{il}^p\right)^T. \quad (9)$$

Differing from NSA, the within-class scatter matrix of NFA is of nonparametric version. Moreover, NFA constructs the nonparametric between-class scatter and within-class scatter matrices directly using the K nearest neighbors, rather than their local mean.

2.4 Nonparametric Margin Maximum Criterion

Qiu et al. proposed a nonparametric margin maximum criterion (NMMC) method (Qiu and Wu, 2005). The basic idea of NMMC is to find the within-class *farthest* neighbor and the between-class nearest neighbor of each sample point, and then based on them to construct the between-class and within-class scatter matrices. Like NFA, NMMC is a complete nonparametric discriminant analysis method in that the between-class and within-class scatter matrices are both constructed in a nonparametric manner.

It looks for the between-class nearest neighbor of a sample $x \in C_i$ denoted as y

$$y = \left\{ y \notin C_i \mid \|y - x\| \le \|x' - x\|, \forall x' \notin C_i \right\}, \tag{10}$$

and the within-class *furthest* neighbor of x as

$$z = \left\{ z \in C_i \mid \|z - x\| \ge \|x' - x\|, \forall x' \in C_i \right\}. \tag{11}$$

The nonparametric between-class scatter matrix in NMMC is defined as

$$S_B^{NMMC} = \sum_{i=1}^{N} w(i)\left(x_i - y_i\right)\left(x_i - y_i\right)^T. \tag{12}$$

The nonparametric within-class scatter matrix in NMMC is defined as

$$S_W^{NMMC} = \sum_{i=1}^{N} w(i)\left(x_i - z_i\right)\left(x_i - z_i\right)^T. \tag{13}$$

The nonparametric margin maximum criterion is

$$W_{opt} = \arg\max_W tr\left(W^T\left(S_B^{NMMC} - S_W^{NMMC}\right)W\right). \tag{14}$$

Obviously, this criterion can work even when S_W is singular. By this criterion, we can get an optimal projection matrix W_{opt}.

3. Push-Pull Marginal Discriminant Analysis

3.1 PPMDA for two-class cases

The NMMC method, relying on the within-class farthest neighbor in the construction of the within-class scatter matrix, may encounter the following problem: minimizing the distance between a point and its within-class farthest neighbor does not make sense for classification in some cases. As shown in Figure 1, reducing the distance between a sample x and its within-class furthest neighbor z has no effect on the classification of the two-class samples.

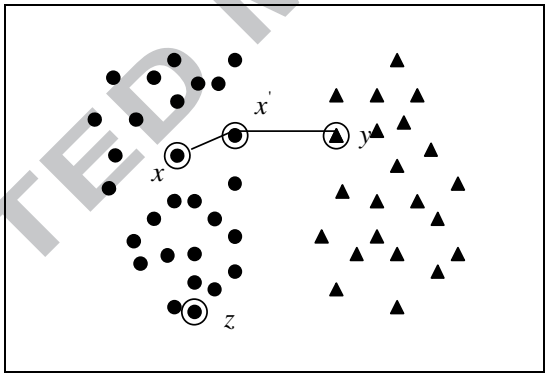


Figure 1 Illumination of neighbors in two-class cases. For the sample x in C_1, its between-class nearest neighbor is y in C_2, its within-class furthest neighbor is z in C_1, the between-class nearest neighbor of y is x' in C_1.

In this paper, we propose a nonparametric method called Push-Pull marginal discriminant analysis (PPMDA). Look at Figure 1. For the sample x in C_1, we find its between-class nearest neighbor y in C_2. Then with respect to y, we find its between-class nearest

neighbor $x^{'}$ in C_1. We can see that $x^{'}$ and y are marginal samples. Intuitively, for increasing the class margin, we push $x^{'}$ away from y and simultaneously pull $x^{'}$ close to x.

When the two classes are overlapped, using only the nearest neighbor might fail to characterize a proper margin. To overcome this problem, we can use *k* nearest neighbors (kNNs) for marginal characterization. Specifically, as illustrated in Figure 2, for sample $x \in C_1$, we find its C_2-kNNs instead of nearest neighbor. We denote the local mean of C_2-kNNs as m_2. For m_2, we then find its C_1-kNNs. The local mean of C_1-kNNs is denoted as m_1. If a proper k is chosen, we can guarantee that m_1 and m_2 are not in the overlapped field. Thus we can increase the margin by pushing m_1 away from m_2 and simultaneously pulling m_1 to x.

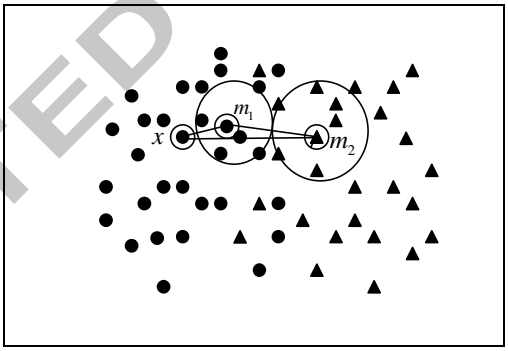


Figure 2 Illustration of two overlapped classes. For the sample $x \in C_1$, its C_2-kNNs are within the right circle. The local mean of C_2-kNNs is m_2, the C_1-kNNs of m_2 are within the left circle. The local mean of C_1-kNNs is m_1.

Formally, given two classes C_i and C_j $(i \neq j)$, we begin with the samples in C_i. For sample $x_{il} \in C_i$, we find its C_j-kNNs, then we can get its C_j-local mean m_{jl} which is computed by Eq. (5). Then we can get the C_i-local mean $m_{il}^{'}$ of m_{jl} in the same way. We define one-side between-class scatter as $\sum_{l=1}^{N_i}\left\|m_{il}^{'}-m_{jl}\right\|^2$ and one-side within-class scatter as $\sum_{l=1}^{N_i}\left\|m_{il}^{'}-x_{il}\right\|^2$. In an average sense, pushing $m_{il}^{'}$ away from m_{jl} as far as possible and pulling $m_{il}^{'}$ to x_{il} as close as possible is equivalent to maximizing the ratio of one-side between-class scatter to one-side within-class scatter.

Now let's consider the problem in the transformed space. After the linear transform

$$\tilde{x} = W^T x, \text{ where } W = (\varphi_1, ..., \varphi_d) \tag{15}$$

For simplicity, let's first consider a one-dimensional linear transform $\tilde{x} = \varphi^T x$. After this transform, x_{il}, m_{jl} and $m_{il}^{'}$ in observed space are mapped into $\tilde{x}_{il} = \varphi^T x_{il}$, $\tilde{m}_{jl} = \varphi^T m_{jl}$ and $\tilde{m}_{il}^{'} = \varphi^T m_{il}^{'}$.

Let's define one-side between-class scatter in transformed space as follows

$$\begin{aligned}\sum_{l=1}^{N_i}\left(\tilde{m}_{il}^{'}-\tilde{m}_{jl}\right)^2 &= \sum_{l=1}^{N_i}\left(\varphi^T m_{il}^{'}-\varphi^T m_{jl}\right)\left(\varphi^T m_{il}^{'}-\varphi^T m_{jl}\right)^T \\ &= \varphi^T\left[\sum_{l=1}^{N_i}\left(m_{il}^{'}-m_{jl}\right)\left(m_{il}^{'}-m_{jl}\right)^T\right]\varphi \\ &= \varphi^T S_B^{ij}\varphi,\end{aligned} \tag{16}$$

where

$$S_B^{ij} = \sum_{l=1}^{N_i} \left(m_{il}^{'} - m_{jl} \right) \left(m_{il}^{'} - m_{jl} \right)^T \tag{17}$$

is one-side between-class scatter matrix.

Similarly, we can define one-side within-class scatter in transformed space as follows

$$\begin{aligned} \sum_{l=1}^{N_i} \left(\tilde{m}_{il}^{'} - \tilde{x}_{il} \right)^2 &= \sum_{l=1}^{N_i} \left(\varphi^T m_{il}^{'} - \varphi^T x_{il} \right) \left(\varphi^T m_{il}^{'} - \varphi^T x_{il} \right)^T \\ &= \varphi^T \left[\sum_{l=1}^{N_i} \left(m_{il}^{'} - x_{il} \right) \left(m_{il}^{'} - x_{il} \right)^T \right] \varphi \\ &= \varphi^T S_W^{ij} \varphi , \end{aligned} \tag{18}$$

where

$$S_W^{ij} = \sum_{l=1}^{N_i} \left(m_{il}^{'} - x_{il} \right) \left(m_{il}^{'} - x_{il} \right)^T \tag{19}$$

is one-side within-class scatter matrix.

To maximize the ratio of one-side between-class scatter to one-side within-class scatter, we can choose the following criterion

$$J(\varphi) = \frac{\varphi^T S_B^{ij} \varphi}{\varphi^T S_W^{ij} \varphi} . \tag{20}$$

The optimal solution of Eq. (20) is actually the generalized eigenvector φ of $S_B^{ij} X = \lambda S_W^{ij} X$ corresponding to the largest eigenvalue.

Symmetrically, let's take the problem from the other side and begin with the samples in C_j. Similarly, we can define the other-side between-class scatter and the other-side within-class scatter. Just like the one-side case, we can get the other-side between-class

and within-class scatter matrices S_B^{ji} and S_W^{ji}. So the other-side between-class and within-class scatter in the transformed space are $\varphi^T S_B^{ji} \varphi$ and $\varphi^T S_W^{ji} \varphi$, respectively.

Our purpose is to maximize the ratio of (both-side) between-class scatter to within-class scatter. We can choose the following criterion

$$J(\varphi) = \frac{\varphi^T \left(S_B^{ij} + S_B^{ji} \right) \varphi}{\varphi^T \left(S_W^{ji} + S_W^{ji} \right) \varphi} . \tag{21}$$

3.2 Extension to multi-class cases

For each pair of C_i and C_j $(i \neq j)$, we can compute the one-side between-class and within-class scatter. In the transformed space, the one-side between-class and within-class scatter are $\varphi^T S_B^{ij} \varphi$ and $\varphi^T S_W^{ij} \varphi$, respectively. Our purpose is to maximize the ratio of (all-side) between-class scatter to (all-side) within-class scatter. We can choose the following criterion

$$J(\varphi) = \frac{\varphi^T S_B^{PPMDA} \varphi}{\varphi^T S_W^{PPMDA} \varphi}, \tag{22}$$

where

$$S_B^{PPMDA} = \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{l=1}^{N_i} \left(m_{il}^{'} - m_{jl} \right) \left(m_{il}^{'} - m_{jl} \right)^T , \tag{23}$$

$$S_W^{PPMDA} = \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{l=1}^{N_i} \left(m_{il}^{'} - x_{il} \right) \left(m_{il}^{'} - x_{il} \right)^T . \tag{24}$$

Like FLDA, for multi-class problems, only one projection axis φ is not enough for discrimination. So we generally need to find a set of projection axis. Similar to the way adopted by FLDA to get multiple projection axes, we can calculate the generalized

eigenvectors $\varphi_1,...,\varphi_d$ of $S_B^{PPMDA}X = \lambda S_W^{PPMDA}X$ corresponding to the *d* largest eigenvalues and use them as projection axis to produce a transform matrix $W = (\varphi_1,...,\varphi_d)$, where *d* is the number of chosen projection axes. The linear transformation $\tilde{x} = W^T x$ forms a feature extractor which reduces the dimension of original feature vectors to *d*.

3.3 PPMDA Algorithm

In summary of the description above, the Push-Pull marginal discriminant analysis (PPMDA) algorithm is given below:

Step 1. For each sample $x_{il} \in C_i$ ($l = 1,2,...,N_i$, N_i is the number of samples in class C_i, $i = 1,2,...,L$), find its k nearest neighbors in C_j and compute the local mean vector m_{jl} ($j = 1,2...,L, j \neq i$) by Eq.(5). For each m_{jl}, find its k nearest neighbors in C_i and compute the local mean vector $m_{il}^{'}$.

Step 2. Based on the obtained local mean vectors, construct the between-class and within-class scatter matrices S_B^{PPMDA} and S_W^{PPMDA} using Eqs. (23) and (24). Compute the generalized eigenvector $\varphi_1,...,\varphi_d$ of $S_B^{PPMDA}X = \lambda S_W^{PPMDA}X$ corresponding to the largest d eigenvalues. Let $W = (\varphi_1,...,\varphi_d)$.

Step 3. For a given sample x, its feature vector $\tilde{x}$ is obtained by $\tilde{x} = W^T x$.

It should be noted that S_W may be singular in small sample size cases. We borrow the idea in PCA+LDA (Belhumeur et al. 1997) and discriminant eigenfeatures (Swets et al. 1996) and use PCA to reduce the dimension of input space firstly so that S_W is nonsingular in the PCA-transformed space. Then we perform PPMDA in the PCA-

transformed space. Further, we can regularize the within-class scatter matrix to avoid overfitting

$$S_W^{PPMDA} \leftarrow S_W^{PPMDA} + \alpha I \,, \tag{25}$$

where I is the identity matrix and $\alpha = 0.001 \times trace(S_W)$.

Finally, we would like to analyze the computational complexity of PPMDA. In the construction of the between-class and within-class scatter matrices S_B^{PPMDA} and S_W^{PPMDA}, for each training sample, we need to find its k nearest neighbors within each class. Therefore, compared to the FLDA method, an additional computational cost of PPMDA is required for the nearest neighbor search. The naive (linear) search of the k neighbors of one point within C_i has a running time of $O(kN_iD)$, where N_i is the number of samples in C_i and D is of dimension of the pattern vectors. So the computational complexity for nearest neighbor search in PPMDA is $O(kN^2D)$, where is N is total number of training samples, $N = \sum_{i=1}^{c} N_i$. The naive search algorithm only suits for small sample size cases. For large sample size cases, more advanced nearest neighbor search algorithms with lower computational complexity can be used instead (Vaidya, 1989; Arya, 1998).

3.5 FLDA: A Special Case of PPMDA

Assume the number of training samples per class is same i.e., $N_i = N/L, (i = 1, \cdots, L)$. We choose k as the number of training samples per class. In this case, we can prove that PPMDA is equivalent to FLDA.

For the sample $x_{il} \in C_i$, its C_j-local mean m_{jl} is exactly the mean vector m_j of C_j. Similarly, the C_i-local mean $m_{il}^{'}$ of m_{jl} is exactly the mean vector m_i of C_i. The global mean vector is

$$m = \frac{\sum_{i=1}^{L} N_i m_i}{N_i L} = \frac{1}{L}\sum_{i=1}^{L} m_i \,. \tag{26}$$

When $k = N_i = N/L, (i = 1, \cdots, L)$, the Eq. (23) can be derived as follows

$$S_B^{PPMDA} = \sum_{i=1}^{L}\sum_{\substack{j=1 \\ j\neq i}}^{L}\sum_{l=1}^{N_i}\left(m_{il}^{'} - m_{jl}\right)\left(m_{il}^{'} - m_{jl}\right)^T$$

$$= \sum_{i=1}^{L}\sum_{\substack{j=1 \\ j\neq i}}^{L} N_i \left(m_i - m_j\right)\left(m_i - m_j\right)^T$$

$$= \sum_{i=1}^{L}\sum_{j=1}^{L} N_i \left(m_i - m_j\right)\left(m_i - m_j\right)^T \text{ (note that } m_i - m_j = 0 \text{ when } i = j\text{)}$$

$$= \sum_{i=1}^{L}\sum_{j=1}^{L} N_i \left(m_i - m + m - m_j\right)\left(m_i - m + m - m_j\right)^T$$

$$= \sum_{i=1}^{L}\sum_{j=1}^{L} N_i \begin{bmatrix} \left(m_i - m\right)\left(m_i - m\right)^T + \left(m_i - m\right)\left(m - m_j\right)^T \\ + \left(m - m_j\right)\left(m_i - m\right)^T + \left(m - m_j\right)\left(m - m_j\right)^T \end{bmatrix}$$

$$= L\sum_{i=1}^{L} N_i \left(m_i - m\right)\left(m_i - m\right)^T + L\sum_{j=1}^{L} N_i \left(m_j - m\right)\left(m_j - m\right)^T$$

$$= 2L\sum_{i=1}^{L} N_i \left(m_i - m\right)\left(m_i - m\right)^T \,. \tag{27}$$

When $k = N_i = N/L, (i = 1, \cdots, L)$, Eq. (24) can be derived as follows

$$S_W^{PPMDA} = \sum_{i=1}^{L}\sum_{\substack{j=1 \\ j\neq i}}^{L}\sum_{l=1}^{N_i}\left(m_{il}^{'} - x_{il}\right)\left(m_{il}^{'} - x_{il}\right)^T$$

$$= (L-1)\sum_{i=1}^{L}\sum_{l=1}^{N_i}\left(x_{il}-m_i\right)\left(x_{il}-m_i\right)^T . \tag{28}$$

We then have

$$J(\varphi) = \frac{\varphi^T S_B^{PPMDA} \varphi}{\varphi^T S_W^{PPMDA} \varphi}$$

$$= \frac{2L\varphi^T \left[\sum_{i=1}^{L} N_i \left(m_i - m\right)\left(m_i - m\right)^T\right]\varphi}{(L-1)\varphi^T \left[\sum_{i=1}^{L}\sum_{l=1}^{N_i}\left(x_{il}-m_i\right)\left(x_{il}-m_i\right)^T\right]\varphi}$$

$$\Leftrightarrow \frac{\varphi^T \left[\sum_{i=1}^{L} N_i \left(m_i - m\right)\left(m_i - m\right)^T\right]\varphi}{\varphi^T \left[\sum_{i=1}^{L}\sum_{l=1}^{N_i}\left(x_{il}-m_i\right)\left(x_{il}-m_i\right)^T\right]\varphi}$$

$$\Leftrightarrow \frac{\varphi^T S_B^{FLDA} \varphi}{\varphi^T S_W^{FLDA} \varphi} . \tag{29}$$

Therefore, the PPMDA method is equivalent to FLDA when each class has the same number of training samples and the nearest neighbor parameter *k* is chosen as the number of training samples per class.

3.6 Advantages of PPMDA over others

In contrast to previously mentioned nonparametric methods, PPMDA pays more attention to the marginal samples which are significant for classification. The construction of the between-class scatter matrix of PPMDA fully depends on marginal samples, while the construction of the within-class scatter matrix is also related to marginal samples. The nature of the scatter matrices of PPMDA inherently leads to features which can preserve marginal structures for classification.

On the other hand, our PPMDA method doesn't need the complicated weighting function. The other methods, such as NSA, NFA and NMMC all need a complicated weighting function. Note that in the weighting function, a parameter is needed to be evaluated. The choice of the parameter must affect the performance of these methods. The proposed PPMDA method, however, does not need the weighting function. So, the proposed method is simpler to be implemented.

4. Experiments

In this section, the push-pull marginal discriminant analysis (PPMDA) method is evaluated using the CENPARMI handwritten numeral database, the ORL database, and the Extended Yale face database B and compared with PCA (Turk et al. 1991), FLDA, Nonparametric Margin Maximum Criterion(NMMC), Principal Nonparametric Subspace Analysis (PNSA), Principal Nonparametric Feature Analysis (PNFA). A nearest neighbor (NN) classifier is employed for classification. The justification for using the NN classifier can be traced to Bressan et al. (2003)'s work, where the connection between nonparametric discriminant analysis (NDA) and the nearest neighbor (NN) classifier is revealed. NDA is to maximize the distance between classes meanwhile minimize the distance among the members of a single class. Given a sample x, the rule of NN classifier is ratio of the between-class distance and within-class distance of x, if the ratio is more than one, x will be correctly classified. Therefore the NN classifier is suitable for NDA. Following the same spirit, the NN classifier is also suitable for the proposed PPMDA method, since PPMDA makes full use of the nearest neighbor rule in its model construction.

Two criteria are involved to evaluate the performance of different feature extraction methods: one is the recognition rate and the other is the verification rate. For the former, we report the recognition rate versus the variation of feature dimensions. For the later, we use the Receiver Operating Characteristic (ROC) curves which plots the face verification rate (FVR) versus the false accept rate (FAR), to show the verification performance of different methods.

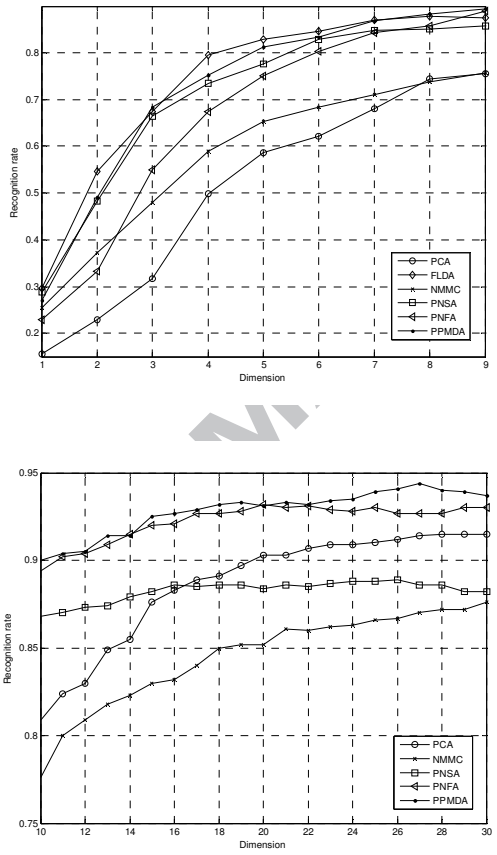
Note that NSA based on the principal space of the within-class scatter matrix is called Principal NSA (PNSA) (Li et al. 2005). NFA based on the principal space of the within-class scatter matrix is called Principal NFA (PNFA) (Li et al. 2009).

4.1 Experiment using the CENPARMI handwritten numeral database

The experiment was done on Concordia University CENPARMI handwritten numeral database. The database contains 6000 samples of 10 numeral classes (each class has 600 samples). In our experiment, we choose the first 200 samples of each class for training, the remaining 400 samples for testing. Thus, the total number of training samples is 2000 while the total number of testing samples is 4000.

PCA, FLDA, NMMC, PNSA, PNFA, and the proposed PPMDA are used respectively, for feature extraction based on the original 121-dimensional Legendre moment features (Liao et al. 1996). Note for PNSA, PNFA, PPMDA, K=6. Figure 3(a) shows the recognition rate when the dimension varies from 1 to 9, and Figure 3(b) shows the recognition rate when the dimension varies from 10 to 30 (The number of features of

FLDA has an upper limit of *L*-1 since the rank of the between-class scatter matrix is at most *L*-1, so the maximal dimension is extremely lower than the other methods. Here L-1 is 9, so we cannot see FLDA in (b)). The ROC curve of each method is shown in Figure 4. The maximal recognition rate of each method and the corresponding dimension are listed in Table 1.



(a) (b)

Figure 3 The recognition rates of PCA, FLDA, NMMC, PNSA, PNFA and PPMDA versus the variation of dimensions on the CENPARMI handwritten numeral database; (a) low dimensions; (b) high dimensions

Table 1 The maximal recognition rates (%) of PCA, FLDA, NMMC, PNSA, PNFA and PPMDA and the corresponding dimensions on the CENPARMI handwritten numeral database

Method	PCA	FLDA	NMMC	PNSA	PNFA	PPMDA
Maximal Recognition Rate	91.5	87.8	87.6	88.9	93.2	94.4
Dimension	28	8	30	26	19	27

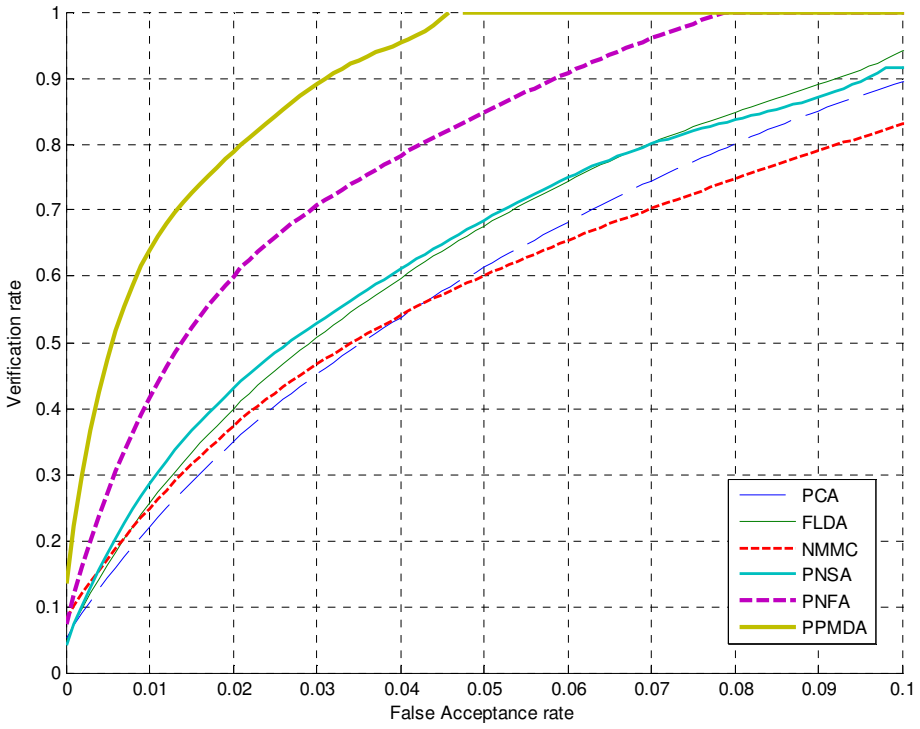


Figure 4 ROC curves of each method on the CENPARMI handwritten numeral database

Figure 3(a) shows that PPMDA almost outperforms PCA, NMMC, PNSA and PNFA in lower dimensions. When the dimension varies from 1 to 7, FLDA is almost best, just slightly more effective than PPMDA. But when the dimension varies from 8 and 9, PPMDA is best. Figure 3(b) shows that PPMDA outperforms the other four methods especially when dimension varies from 24 to 30. Table 1 shows the best recognition rate

of our PPMDA method is 94.4% when the dimension is 27. Figure 4 shows PPMDA achieves better verification performance than the other five methods. In particular, when FAR is 0.047, PPMDA achieves a verification rate of 100% which is over 10% higher than the other methods.

4.2 Experiment using the Extended Yale database B

The Yale face database B (Georghiades et al. 2001) contains 5760 single light source images of 10 subjects each seen under 576 viewing conditions (9 poses*64 illumination conditions). It was updated to the extended Yale face database B (Lee et al. 2005) contains 38 human subjects under 9 poses and 64 illumination conditions. All the image data for test used in the experiments are manually aligned, cropped, and then re-sized to 168*192 images (Lee et al. 2005). All test images are under pose 00 (The pose number is 00-08). Some sample images of one person are shown in Figure 5. In our experiment, we resize each image to 42*48 pixels and further pre-process it using histogram equalization. In our test, we use the first 16 images per subject for training, the remaining 48 images for testing. PCA, FLDA, NMMC, PNSA, PNFA, and the proposed PPMDA are used for feature extraction. Note for PNSA, PNFA, PPMDA, K=2. The recognition rate over the variation of dimensions is plotted in Figure 6. The ROC curve of each method is plotted in Figure 7. The maximal recognition rate of each method and the corresponding dimension are listed in Table 2.

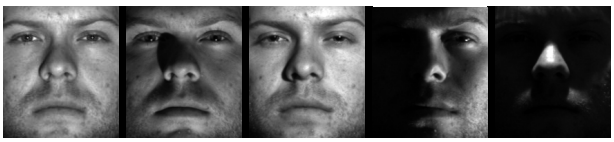


Figure 5 Samples of a person under pose 00 and different illuminations, which are cropped images in the extended Yale face database B

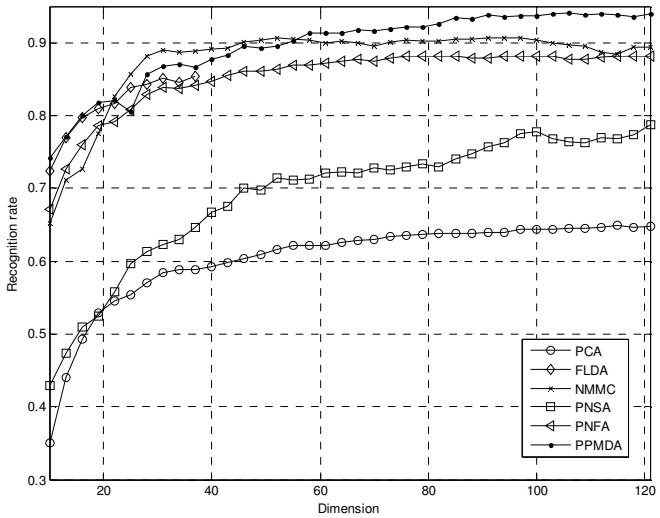


Figure 6 The recognition rates of PCA, FLDA, NMMC, PNSA, PNFA and PPMDA versus the variation of dimensions on the extended Yale face database B

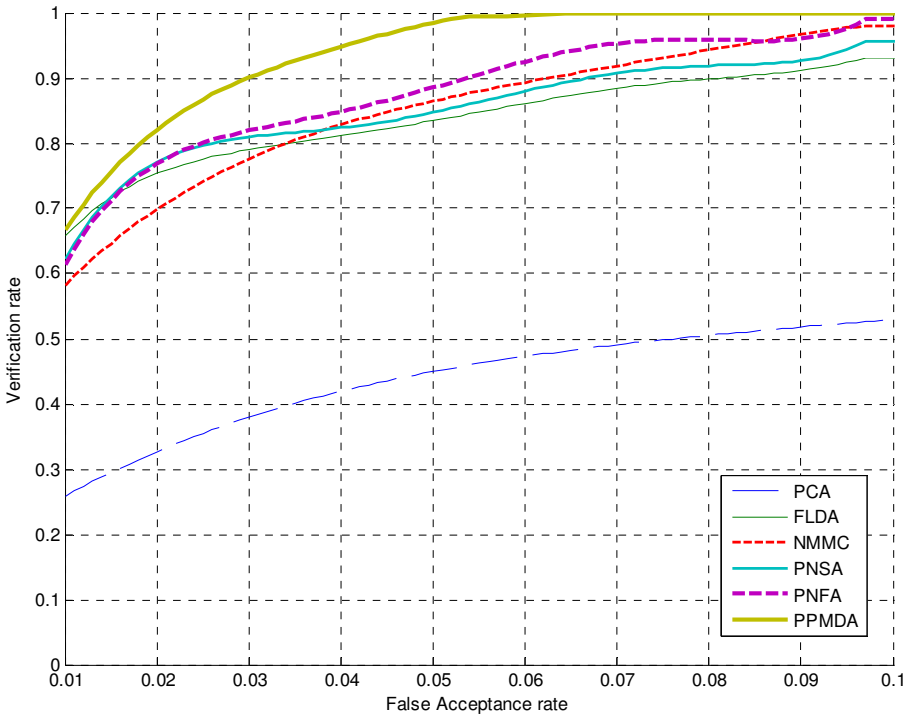


Figure 7 ROC curves of each method on the extended Yale face database B

Table 2 The maximal recognition rates (%) of PCA, FLDA, NMMC, PNSA, PNFA and PPMDA and the corresponding dimensions on the extended Yale face database B

Method	PCA	FLDA	NMMC	PNSA	PNFA	PPMDA
Maximal Recognition Rate	64.9	85.4	90.7	87.2	89.5	94.1
Dimension	115	37	52	100	112	106

Figure 6 shows that when the dimension varies from 20 to 40, NMMC achieves very good results. But when the dimension is over 60, PPMDA obviously outperforms other five methods. Table 2 shows the best results of each method. Our PPMDA method achieves the recognition rate of 94.1%, when the dimension is 106. Figure 7 shows that PPMDA achieves the best verification performance among all of the six methods. Particularly, when FAR is 0.05, the FVR of PPMDA is 98.19% which is about 10% higher than the other methods.

4.3 Experiment using the ORL database

The ORL database (http://www.cam-orl.co.uk) contains images from 40 individuals, each providing 10 different images. For some subjects, the images were taken at different times. The facial expressions (open or closed eyes, smiling or non-smiling) and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there is also some variation in the scale of up to about 10%. All images are grayscale and normalized to a resolution of 92×112 pixels.

In our experiments, we split the whole database into two parts evenly. One part is used for training and the other part is for testing. In order to make full use of the available data and to evaluate the generalization power of algorithms more accurately, we adopt a cross-validation strategy and run the system 50 times. In each time, five face images from each person are randomly selected as training samples. The rest is for testing. PCA, FLDA, NMMC, PNSA, PNFA and the proposed PPMDA are used for feature extraction. Note that for PNSA, PNFA, PPMDA, we choose K=1. Finally, a nearest neighbor classifier is employed for classification with cosine distance. The average recognition rate across 50 tests of each method over the variation of dimensions is plotted in Figure 8. The ROC curve of each method is plotted in Figure 9. The maximal recognition rate of each method and the corresponding dimension are listed in Table 3. Figure 8 and Table 3 reveal that when the number of samples per class is small, PPMDA consistently outperforms the other five methods irrespective of variation in dimensions. Figure 9 demonstrates again the advantage of PPMDA in terms of the verification rate.

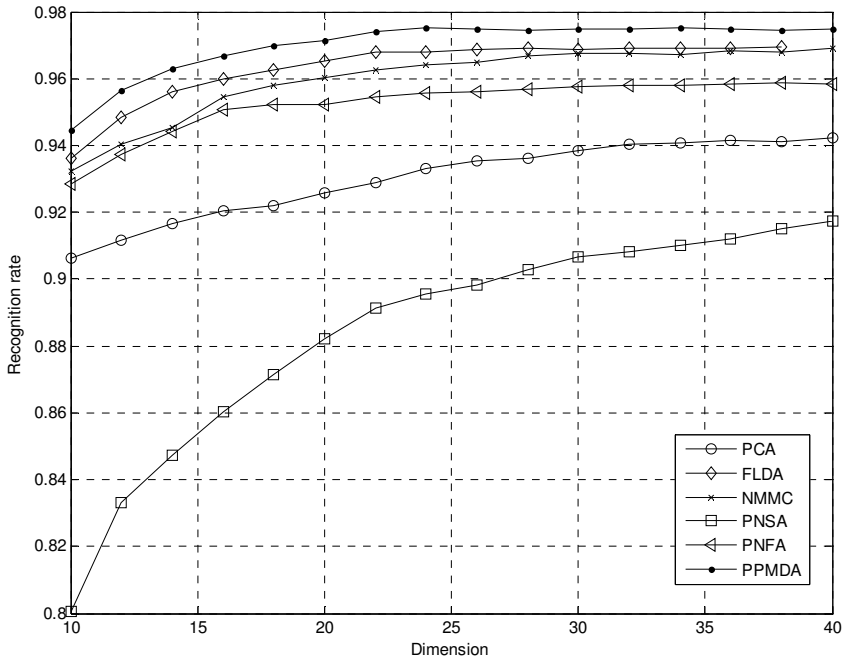


Figure 8 The average recognition rates of PCA, FLDA, NMMC, PNSA, PNFA and PPMDA on ORL database

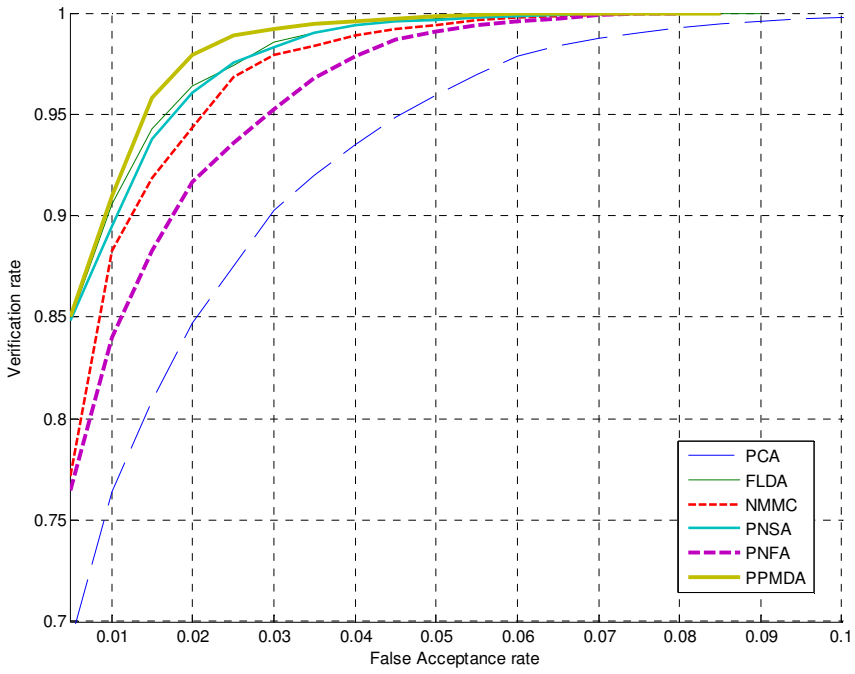


Figure 9 ROC curves of each method on ORL face database

Table 3 The maximal recognition rates (%) of PCA, FLDA, NMMC PNSA, PNFA and PPMDA and the corresponding dimensions on the ORL database

Method	PCA	FLDA	NMMC	PNSA	PNFA	PPMDA
Maximal Recognition Rate	94.21	96.95	96.89	91.74	95.86	97.54
Dimension	40	37	40	40	38	24

5. Conclusions

We present a new nonparametric discriminant analysis method called Push-Pull marginal discriminant analysis (PPMDA) in this paper. This method takes full advantage of marginal information to construct the within-class and between-class scatter matrices, and then uses a class margin related criterion to determine an optimal transform matrix such that the marginal samples of one class are pushed away from the between-class marginal samples as far as possible and simultaneously pulled to the within-class samples as close as possible. The proposed method is applied to character and face recognition and is evaluated using the CENPARMI handwritten numeral database, the Extended Yale face database B and the ORL database. Experimental results show the effectiveness of the proposed method and its performance advantage over others. This effectiveness also verifies the importance of marginal samples for classification.

Acknowledgments: The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions. This work was partially supported by the Program for New Century Excellent Talents in University of China, the NUST Outstanding Scholar Supporting Program, the National Science Foundation of China under Grants No. 60973098 and 60632050.

Reference

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., and Wu, A. 1998. An optimal algorithm for approximate nearest neighbor searching, Journal of the ACM, 45(6):891-923

Belhumeur, P.N., Hespanda, J.P., Kiregeman D.J., 1997. Eigenfaces versus Fisherfaces: Recognition

Using Class Specific Linear Projection, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720.

Bressan, M., Vitri`a. J., 2003.Nonparametric discriminant analysis and nearest neighbor classification,

Pattern Recognition Letters, 24:2743C2749.

Chen, H.T., Chang, H.W., Liu, T.U., 2005. Local Discriminant Embedding and Its Variants, Proc. IEEE

Conf. Computer Vision and Pattern Recognition, vol.2, pp.846-853.

Chen, L.F, Liao, H.Y.M., Lin, J.C., Ko, M.T, Yu, G.J., 2000. A new LDAbased face recognition system

which can solve the small sample size problem, Pattern Recognition, 33(10):1713–1726.

Cortes, C. and Vapnik, V. Support vector networks. *Machine Learning*, 20:273–297, 1995.

Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. New York: Wiley.

Etemad, K., Chellappa, R., 1996. Face Recognition Using Discriminant Eigenvectors, Proc. IEEE Int'l

Conf. Acoustics, Speech, and Signal Processing, vol. 4, pp. 2148-2151.

Etemad, K., Chellappa, R., 1997. Discriminant Analysis for Recognition of Human Face Images, J.

Optical Soc. Am. A, vol. 14, no. 8, pp. 1724-1733.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems, in Annals of Eugenics,

vol. 7, part 11, pp. 179-188,.

Fukunaga, K., MANTOCK, J. M.,1983. Nonparametric Discriminant Analysis, IEEE Transactions on

Pattern Analysis and Machine Intelligence, VOL. PAMI-S, NO. 6,

Georghiades, A.S., Belhumeur, P.N., Kriegman,D.J., 2001, From Few to Many: Illumination Cone

Models for Face Recognition under Variable Lighting and Pose, IEEE Trans. Pattern Anal.

Mach.Intelligence, volume 23,number 6, pp.643-660

Jin, Z., Yang, J.Y., Hu, Z.S, Lou, Z., 2001. Face Recognition based on uncorrelated discriminant

transformation, Pattern Recognition, 33(7), 1405-1416.

Lee, K.C., Ho, J., Driegman, D., Acquiring Linear Subspaces for Face Recognition under Variable

Lighting, 2005, IEEE Trans. Pattern Anal. Mach. Intelligence, volume 27, number 5,pp 684-698

Li, Z.F., Lin, D.H., Tang, X.O., 2005. Nonparametric Subspace Analysis for Face Recognition, Proc.

IEEE Conf. Computer Vision and Pattern Recognition.

Li, Z.F., Lin, D.H., Tang, X.O., 2009. Nonparametric Discriminant Analysis for Face Recognition. IEEE

Transactions on Pattern Analysis and Machine Intelligence, VOL.31, NO.4.

Liao, S.X., Pawlak, M., 1996,On image analysis by moments, IEEE Trans. Pattern Anal. Machine Intell., 18(3), 254-266.

Liu, K., Cheng, Y.Q., Yang, J.Y., 1992. A generalized optimal set of discriminant vectors, Pattern Recognition, 25(7):731C739.

Loog, M., Duin, R.P.W., 2004. Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 6,pp. 732-739.

Qiu, X.P., Wu, L.D., 2005.Face Recognition By Stepwise Nonparametric Margin Maximum Criterion, In Proc. of IEEE Conference on Computer Vision (ICCV 2005), Beijing (China).

S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph Embedding and extension: a general framework for dimensionality reduction, IEEE Trans. PAMI. 29(1) 2007, 40-51.

Swets, D.L., Weng, J.J, 1996. Using Discriminant Eigenfeatures for Image Retrieval, IEEE Trans.

Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 831-836.

Turk, M.A., Pentland, A.P., 1991, Face Recognition Using Eigenfaces, Proc. IEEE Conf. on Computer

Vision and Pattern Recognition, pp. 586-591.

Vaidya, P. M., 1989. An O(n log n) Algorithm for the All-Nearest-Neighbors Problem, Discrete and Computational Geometry 4 (1): 101–115,

Yu, H., Yang, J., 2001. A direct LDA algorithm for high dimensional data with application to face

recognition, Pattern Recognition, 34:2067–2070.