Null Space Approach of Fisher Discriminant Analysis for Face Recognition

Wei Liu[1], Yunhong Wang[1], Stan Z. Li[2], Tieniu Tan[1]

[1] Institute of Automation, Chinese Academy of Sciences,
100080 Beijing, China
{wliu, wangyh, tnt}@nlpr.ia.ac.cn
[2] Microsoft Research Asia, Beijing Sigma Center,
100080 Beijing, China
szli@microsoft.com

Abstract. The null space of the within-class scatter matrix is found to express most discriminative information for the small sample size problem (SSSP). The null space-based LDA takes full advantage of the null space while the other methods remove the null space. It proves to be optimal in performance. From the theoretical analysis, we present the NLDA algorithm and the most suitable situation for NLDA. Our method is simpler than all other null space approaches, it saves the computational cost and maintains the performance simultaneously. Furthermore, kernel technique is incorporated into discriminant analysis in the null space. Firstly, all samples are mapped to the kernel space through a better kernel function, called Cosine kernel, which is proposed to increase the discriminating capability of the original polynomial kernel function. Secondly, a truncated NLDA is employed. The novel approach only requires one eigenvalue analysis and is also applicable to the large sample size problem. Experiments are carried out on different face data sets to demonstrate the effectiveness of the proposed methods.

1 Introduction

Linear Discriminant Analysis (LDA) has been successfully applied to face recognition. The objective of LDA is to seek a linear projection from the image space onto a low dimensional space by maximizing the between-class scatter and minimizing the within-class scatter simultaneously. Belhumeur [1] compared Fisherface with Eigenface on the *HARVARD* and *YALE* face databases, and showed that LDA was better than PCA, especially under illumination variation. LDA was also evaluated favorably under the *FERET* testing framework [2], [7].

In many practical face recognition tasks, there are not enough samples to make the within-class scatter matrix S_w nonsingular, this is called a small sample size problem. Different solutions have been proposed to deal with it in using LDA for face recognition [1]-[6].

The most widely used methods (Fisherface) [1, 2, 3] applies PCA firstly to reduce the dimension of the samples to an intermediate dimension, which must be guaranteed not more than the rank of S_w so as to obtain a full-rank within-class scatter matrix. Then standard LDA is used to extract and represent facial features. All these methods above do not consider the importance of null space of the within-class scatter matrix, and remove the null space to make the resulting within-class scatter full-rank.

Yang et al. [4] proposed a new algorithm which incorporates the concept of null space. It first removes the null space of the between-class scatter matrix S_b and seeks a projection to minimize the within-class scatter (called Direct LDA / DLDA). Because the rank of S_b is smaller than that of S_w, removing the null space of S_b may lose part of or the entire null space of S_w, which is very likely to be full-rank after the removing operation.

Chen et al. [5] proposed a more straightforward method that makes use of the null space of S_w. The basic idea is to project all the samples onto the null space of S_w, where the resulting within-class scatter is zero, and then maximize the between-class scatter. This method involves computing eigenvalue in a very large dimension since S_w is an $n \times n$ matrix. To avoid the great computational cost, pixel grouping method is used in advance to artificially extract features and to reduce the dimension of the original samples.

Huang et al. [6] introduced a more efficient null space approach. The basic notion behind the algorithm is that the null space of S_w is particularly useful in discriminating ability, whereas, that of S_b is useless. They proved that the null space of the total scatter matrix S_t is the common null space of both S_w and S_b. Hence the algorithm firstly removes the null space of S_t and projects the samples onto the null space of S_w. Then it removes the null space of the between-class scatter in the subspace to get the optimal discriminant vectors.

Although null space-based LDA seems to be more efficient than other linear subspace analysis methods for face recognition, it is still a linear technique in nature. Hence it is inadequate to describe the complexity of real face images because of illumination, facial expression and pose variations. The kernel technique has been extensively demonstrated to be capable of efficiently representing complex nonlinear relations of the input data. Kernel Fisher Discriminant Analysis [8, 9, 10] (KFDA) is an efficient nonlinear subspace analysis method, which combines the kernel technique with LDA. After the input data are mapped into an implicit feature space, LDA is performed to yield nonlinear discriminating features of the input data.

In this paper, some elements of state-of-the-art null space techniques will be looked at in more depth and our null space approach is proposed to save the computational cost and maintain the performance simultaneously. Furthermore, we concentrate on the advantages of both the null space approach and the kernel technique. A kernel mapping based on an efficient kernel function, called Cosine kernel, is performed on all the samples firstly. In kernel space, we can find that the total scatter matrix is full-rank, so the procedure of the null space approach is greatly simplified and more stable in numerical computation.

The paper is laid out as follows. In Section 2, the related work on LDA-based algorithms will be reviewed. Next, our null space method (NLDA) will be presented. In Section 4 null space-based KFDA (NKFDA) will be proposed and some experiments will be reported in Section 5. Finally, Section 6 ends with some conclusions.

2 Previous Work

Some assumptions and definitions in mathematics are provided at first. Let n denote the dimension of the original sample space, and c is the number of classes. The between-class scatter matrix S_b and the within-class scatter S_w are defined as below:

$$S_b = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T = \Phi_b \Phi_b^T \ , \quad (1)$$

$$S_w = \sum_{i=1}^{c} \sum_{k \in C_i} (x_k - m_i)(x_k - m_i)^T = \Phi_w \Phi_w^T \ , \quad (2)$$

where N_j is the number of samples in class C_i (i=1,2,…,c), N is the number of all samples, m_j is the mean of the samples in the class C_i, and m is the overall mean of all samples. The total scatter matrix i.e. the covariance matrix of all the samples is defined as:

$$S_t = S_b + S_w = \sum_{i=1}^{N} (x_i - m)(x_i - m)^T = \Phi_t \Phi_t^T \ . \quad (3)$$

LDA tries to find an optimal projection: $W = [w_1, w_2, w_3, ..., w_{c-1}]$, which satisfies

$$J(W) = \arg\max_W \frac{|W^T S_b W|}{|W^T S_w W|} \ , \quad (4)$$

that is just Fisher criterion function.

2.1 Standard LDA and Direct LDA

As well known, W can be constructed by the eigenvectors of $S_w^{-1} S_b$. But this method is numerically unstable because it involves the direct inversion of a likely high-dimensional matrix. The most frequently used LDA algorithm in practice is based on simultaneous diagonalization. The basic idea of the algorithm is to find a matrix W that can simultaneously diagonalize both S_w and S_b, i.e.,

$$W^T S_w W = I \ , W^T S_b W = \Lambda \ . \quad (5)$$

Most algorithms require that S_w be non-singular, because the algorithms diagonalize S_w first. The above procedure will break down when S_w becomes singular. It surely happens when the number of training samples is smaller than the dimension of the sample vector, i.e. the small sample size problem (SSSP). The singularity exists for most face recognition tasks.

An available solution to this problem is to perform PCA to project the n-dimensional image space onto a lower dimensional subspace. The PCA step essentially removes null space from both S_w and S_b. Therefore, this step potentially loses useful information.

In fact, the null space of S_w contains the most discriminative information especially when the projection of S_b is not zero in that direction. The Direct LDA (DLDA) algorithm [4] is presented to keep the null space of S_w.

DLDA removes the null space of S_b firstly by performing eigen-analysis on S_b, then a simultaneous procedure is used to seek the optimal discriminant vectors in the subspace of S_b, i.e.

$$W^T S_b W = I, W^T S_w W = D_w \ . \quad (6)$$

Because the rank of S_b is smaller than that of S_w in majority, removing the null space of S_b may lose part of or the entire null space of S_w, which is very likely to be full-rank after the removing operation. So, DLDA does not make full use of the null space.

2.2 Null Space-based LDA

From Fisher's criterion that is objective function (4), we can find that: In standard LDA, W is seeked such that (5), so the form of the optimal solution provided by standard LDA is

$$optimum_{LDA} = \max_W \left|W^T S_b W\right| / \left|W^T S_w W\right| = \left|\Lambda\right| = opt \max / 1 \ . \quad (7)$$

In DLDA, W is seeked such that (6), so the form of the optimal solution provided by DLDA is

$$optimum_{DLDA} = \max_W \left|W^T S_b W\right| / \left|W^T S_w W\right| = 1 / \left|D_w\right| = 1 / opt \min \ . \quad (8)$$

Compared with above LDA approaches, a more reasonable method (Chen [5]), we called Null Space-based LDA, has been presented. In Chen's theory, null space-based LDA should reach below:

$$optimum_{Null} = \max_W \left|W^T S_b W\right| / \left|W^T S_w W\right| = opt \max / 0 \ . \quad (9)$$

That means the optimal projection W should satisfy

$$W^T S_w W = 0,\ W^T S_b W = \Lambda \ , \quad (10)$$

i.e. the optimal discriminant vectors must exist in the null space of S_w.

In a performance benchmark, we can conclude that null space-based LDA generally outperforms LDA (Fisherface) or DLDA since

$$optimum_{Null} = \infty \geq optimum_{DLDA} \geq optimum_{LDA} \ . \quad (11)$$

Because the computational complexity of extracting the null space of S_w is very high because of the high dimension of S_w. So in [5] a pixel grouping operation is used in advance to extract geometric features and to reduce the dimension of the samples. However, the pixel grouping preprocess is irresponsible and may arouse a loss of useful facial features.

3 Our Null Space Method (NLDA)

In this section, the essence of null space-based LDA in the SSSP is revealed by theoretical justification, and the most suitable situation of null space methods is discovered. Next, we propose the NLDA algorithm, which is conceptually simple yet powerful in performance.

3.1 Most Suitable Situation

For the small sample size problem (SSSP) in which $n>N$, the dimension of null space of S_w is very large, and not all null space contributes to the discriminative power. Since both S_b and S_w are symmetric and semi-positive, we can prove, as mentioned in [6], that

$$N(S_t) = N(S_b) \bigcap N(S_w) . \qquad \textbf{(12)}$$

From the statistical perspective, the null space of S_b is of no use in its contribution to discriminative ability. Therefore, the useful subspace of null space of S_w is

$$\hat{N}(S_w) = N(S_w) - N(S_t) = N(S_w) \bigcap \overline{N(S_t)} . \qquad \textbf{(13)}$$

The sufficient and necessary condition so that null space methods work is

$$\hat{N}(S_w) \neq \Phi \Rightarrow N(S_w) \supset N(S_t) \Rightarrow \dim N(S_w) > \dim N(S_t) \Rightarrow$$
$$rank(S_t) > rank(S_w) . \qquad \textbf{(14)}$$

In many cases,

$$rank(S_t) = \min\{n, N-1\}, rank(S_w) = \min\{n, N-c\} , \qquad \textbf{(15)}$$

the dimension of discriminative null space of S_w can be evaluated from (12):

$$\dim \hat{N}(S_w) = rank(S_t) - rank(S_w) . \qquad \textbf{(16)}$$

If $n \leq N-c$, due to $rank(S_t) = n \leq rank(S_w) = N-c$, the necessary condition (14) is not satisfied so that we can not extract any null space. That means any null space-based method does not work in the large sample size case.

If $N-c < n < N-1$, due to $rank(S_t) = n > rank(S_w) = N-c$, the dimension of effective null space can be evaluated from (16): $\dim \hat{N}(S_w) = n - N + c < c-1$. Hence, the number of discriminant vectors would be less than c-1, and some discriminatory information maybe lost.

Only when $n \geq N-1$ (SSSP), for $rank(S_t) = N-1 > rank(S_w) = N-c$, we derive $\dim \hat{N}(S_w) = c-1$.The dimension of extracted null space is just c-1, which coincides with the number of ideal features for classification. Therefore, we can conclude that null space methods are always applicable to any small sample size problem.

Especially when n is equal to N-1, S_t is full-rank and $N(S_t)$ is null. By (13) we have $\hat{N}(S_w) = N(S_w)$, it follows all null space of S_w contributes to the discriminative power. Hence, we conclude the most suitable situation for null space-based methods:

$$n = N-1 . \qquad \textbf{(17)}$$

3.2 NLDA

Combining (12)-(16), we develop our null space method.

algorithm I:

1. Remove the null space of S_t.

   Perform PCA to project the n-dimensional image space onto a low dimensional subspace, i.e. perform eigen-analysis on S_t., the dimension of the extracted subspace is usually N-1. The projection P, whose columns are all the eigenvectors of S_t corresponding to the nonzero eigenvalues, are calculated firstly, and then the within-class scatter and between-class scatter in the resulting subspace are obtained.

$$P^T S_t P = D_t,\ \ P^T S_w P = S_w^{'},\ \ P^T S_b P = S_b^{'}\ .$$

2. Extract the null space of $S_w{}'$.

   Diagonalize $S_w{}'$, we have

$$V^T S_w^{'} V = D_w\ ,$$

   where $V^T V = I, D_w$ is diagonal matrix sorted in increasing order. Discard those with eigenvalues sufficiently far from 0, keep c-1 eigenvectors of $S_w{}'$ in most cases. Let Y be the first c-1 columns of V, which is the null space of $S_w{}'$, we have

$$Y^T S_w^{'} Y = 0\ ,\ Y^T S_b^{'} Y = S_b^{''}\ .$$

3. Diagonalize $S_b{}''$ (usually a (c-1)×(c-1) matrix) which is full-rank.

   Perform eigen-analysis:

$$U^T S_b^{''} U = \Lambda\ ,$$

   where $U^T U = I, \Lambda$ is diagonal matrix sorted in decreasing order.

   The final projection matrix is:

$$W = PYU\ ,$$

W is usually an n×(c-1) matrix, which diagonalizes both the numerator and the denominator of Fisher's criterion to (c-1)×(c-1) matrices as (10) , especially leads to a denominator of 0 matrix.

It is notable that the third step of Huang [6]' algorithm is used to remove the null space of $S_b{}''$. In fact, we are able to prove that it is full-rank once through the previous two steps.

Lemmas $S_b{}''$ is full-rank, $S_b{}''$ is defined in step2 of algorithm I.

Proof:

From step1 and 2, we derive that $S_b^{''} = Y^T S_b^{'} Y = Y^T S_b^{'} Y + Y^T S_w^{'} Y = Y^T (S_b^{'} + S_w^{'}) Y = Y^T P^T (S_b + S_w) PY = Y^T P^T S_t PY = Y^T D_t Y$, for any vector α whose dimension is equal to that of $S_b{}''$, $\alpha^T S_b^{''} \alpha = \alpha^T Y^T D_t Y \alpha = (D_t^{1/2} Y\alpha)^T (D_t^{1/2} Y\alpha) \geq 0$, so $S_b{}''$ is semi-positive. Suppose there exists α such that $\alpha^T S_b^{''} \alpha = 0$, then $D_t^{1/2} Y\alpha = 0$. By step1, we know D_t is full-rank, thus $Y\alpha = 0$. And by step2, we derive that Y is full-rank in columns since it is the extracted null space. Hence $\alpha = 0$, iff. $\alpha^T S_b^{''} \alpha = 0$. Therefore $S_b{}''$ is a positive matrix which is of course full-rank.□

The third step is optional. Although it maximizes the between-class scatter in the null subspace, which appears to achieve best discriminative ability, it may incur overfitting. Because projecting all samples onto the null space of S_w is powerful enough in its clustering ability to achieve good generalization performance, step3 of algorithm *I* should be eliminated in order to avoid possible overfitting.

NLDA algorithm:

1. Remove the null space of S_t, i.e.

$$P^T S_t P = D_t \ , \ P^T S_w P = S_w^{'} \ ,$$

   P is usually $n\times(N\text{-}1)$.

2. Extract the null space of $S_w^{'}$, i.e.

$$Y^T S_w^{'} Y = 0 \ ,$$

   Y is the null space, and is usually $(N\text{-}1)\times(c\text{-}1)$.

The final NLDA projection matrix is:

$$W = PY \ ,$$

PY is the discriminative subspace of the whole null space of S_w and is really useful for discrimination. The number of the optimal discriminant vectors is usually c-1, which just coincides with the number of ideal discriminant vectors [1]. Therefore, removing step3 is a feasible strategy against overfitting.

Under situation (17), S_t is full-rank and step1 of the NLDA algorithm is skipped. The NLDA projection can be extracted by performing eigen-analysis on S_w directly. The procedure of NLDA under this situation is most straightforward and only requires one eigen-analysis. We can discover that NLDA will save much computational cost under the most suitable situation it is applicable to.

4 Null Space-based Kernel Fisher Discriminant Analysis

The key idea of Kernel Fisher Discriminant Analysis (KFDA) [8, 9, 10] is to solve the problem of LDA in an implicit feature space *F*, which is constructed by the kernel trick:

$$\phi : x \in R^n \rightarrow \phi(x) \in F \ . \qquad \textbf{(18)}$$

The important feature of kernel techniques is that the implicit feature vector ϕ needn't be computed explicitly, while the inner product of any two vectors in *F* need to be computed based a kernel function.

In this section, we will present a novel method (NKFDA) in which kernel technique is incorporated into discriminant analysis in the null space.

4.1 Kernel Fisher Discriminant Analysis (KFDA)

The between-class scatter S_b and the within-class scatter S_w in *F* are computed as (1) and (2). But at this time, we replace x_j by $\phi(x_j)$ as samples in *F*. Consider performing LDA in the implicit feature space *F*. It caters for maximizing the Fisher criterion function (4).

Because any solution $w \in F$ must lie in the span of all the samples in *F*, there exist coefficients α_i, *i*=1,2…*N*, such that

$$w = \sum_{i=1}^{N} \alpha_i \phi_i \ . \tag{19}$$

Substitute w in (4), the solution of (4) can be obtained by solve a new Fisher problem:

$$J(\alpha) = \arg\max_{\alpha} \frac{\left|\alpha^T K_b \alpha\right|}{\left|\alpha^T K_w \alpha\right|} \ , \tag{20}$$

where K_b and K_w (Liu [8]) are based on new samples:

$$\zeta_i = (k(x_1, x_i), k(x_2, x_i), ..., k(x_N, x_i))^T, 1 \le i \le N \ . \tag{21}$$

As for the kernel function, Liu [13] proposed a novel kernel function called Cosine kernel, which is based on the original polynomial kernel, has been demonstrated to improve the performance of KFDA. It is defined as below:

$$k(x, y) = (\phi(x) \cdot \phi(y)) = (a(x \cdot y) + b)^d \ , \tag{22}$$

$$\tilde{k}(x, y) = \frac{k(x, y)}{\sqrt{k(x, x)k(y, y)}} \ . \tag{23}$$

In our experiments, Cosine kernel (a=10^{-3}/*sizeof* (image), *b*=0, *d*=2) is adopted and shows good performance in face recognition. Cosine measurement should be more reliable than inner production measurement due to a better similarity representation in the implicit feature space.

4.2 NKFDA

Here we define a kernel sample set (corresponding to the kernel space in *N* dimensions) $\{\zeta_i\}_{1\le i\le N}$ .The optimal solution of (4) is equivalent to that of (20), so the original problem can be entirely converted to the problem of LDA on the kernel sample set.

In section 3, we know that NLDA will save much computational cost under the most suitable situation. The null space projection can be extracted from the within-class scatter directly. Our objective is just to transform the dimension of all the samples from *n* to *N*-1 through the kernel mapping, so that NLDA can work under the most suitable situation. Any method that can transform raw samples to (*N*-1)-dimensional data without adding or losing main information, can exploit the merit of NLDA.

In (19), all the training samples in *F*, $\{\phi_i\}_{1\le i\le N}$, are used to represent *w*. Define the kernel matrix *M*,

$$M = (k(x_i, x_j))_{1\le i,j\le N} = (k_{i,j})_{1\le i,j\le N} \ , \tag{24}$$

assume $\Phi = (\phi_1, \phi_2, ..., \phi_N)$, then $M = \Phi^T \Phi$. In mathematics,

$$rank(\Phi) = rank(M) \ . \tag{25}$$

Because $rank(M) < N$ holds, especially when the training data set is very large, it follows that $rank(M) << N$ [11][12], we conclude that

$$rank(\Phi) << N \ . \tag{26}$$

Due to (26), we may assume ϕ_N is not a basis vector of $\{\phi_i\}_{1\le i\le N}$ without loss of generality, and consequently we can rewrite (19) as follows:

$$w = \sum_{i=1}^{N-1} \alpha_i \phi_i \ , \tag{27}$$

subsequently, K_b and K_w are recomputed, we derive :

$$\zeta_i = (k(x_1, x_i), k(x_2, x_i), ..., k(x_{N-1}, x_i))^T \ . \tag{28}$$

Now the dimension of our defined kernel space is N-1. My objective is just to enable NLDA work on the (N-1) -dimensional kernel sample set.

Input: 1) training samples $\{x_i\}_{1\le i\le N}$ and label set $\{C_j\}_{1\le j\le c}$

2) the kernel function and its parameters: $k(x, y)$

Algorithm:

1. For i = 1,2,…,N

do kernel mapping on each training sample:

$$K(x_i) = (k(x_1, x_i), k(x_2, x_i), ..., k(x_{N-1}, x_i))^T \ .$$

For a new sample x, whose corresponding point in the kernel space is

$$K(x) = (k(x_1, x), k(x_2, x), ..., k(x_{N-1}, x))^T \ .$$

2. Calculate class mean and within-class scatter:

$$m_j = \sum_{i\in C_j} K(x_i) / N_j \ ,$$

$$K_w = \sum_{j=1}^{c} \sum_{i\in C_j} (K(x_i) - m_j)(K(x_i) - m_j)^T \ .$$

3. Extract the null space Y of K_w (N-1×N-1), such that

$Y^T K_w Y = 0$, Y is usually in (N-1)×(c-1).

Output: The resulting mapping on the raw sample set:

$$\Psi(x) = (Y^T K) \cdot (x) = Y^T \cdot K(x) \ .$$

Fig. 1. NKFDA algorithm

As shown in Fig. 1, NKFDA algorithm outputs the mapping Ψ which is a nonlinear dimensionality reduction mapping (n dimensions reduce to c-1). For any sample (whether it is a prototype or a query), Ψ provides a universal mapping to transform the raw sample point into a lower dimensional space. Such a technique can be applied with a reasonable implementation of generalization.

It's noticeable that our method NKFDA also cannot deal with the case that only one sample per person is available for training since KFDA can not achieve that.

For the large sample size problem ($n<<N$), S_w is full-rank so that we can not extract any null space. That means any null space-based method does not work in the large sample size case. However, after the kernel mapping, NLDA can work on the kernel sample set. Hence the kernel mapping extends the ability of null space approaches to the large sample size problem.

5 Experiments

To demonstrate the efficiency of our method, extensive experiments are done on the *ORL* face database, the *FERET* database and the mixture database. All LDA methods were compared on the same training sets and testing sets, including Fisherface proposed in [1, 2, 3], Direct LDA proposed in [4], and our methods: NLDA and NKFDA.

5.1 ORL Database

There are 10 different images for each subject in the *ORL* face database composed of 40 distinct subjects. All the subjects are in up-right, frontal position. The size of each face image is 92×112. The first line of Fig. 2 shows 6 images of the same subject.

We listed the recognition rates with different number of training samples. The number of training samples per subject, k, increases from 2 to 9. In each round, k images are randomly selected from the database for training and the remaining images of the same subject for testing. For each k, 20 tests were performed and these results were averaged. Table 1 shows the average recognition rates (%). Without any pre-processing, we choose 39 (i.e. c-1) as the final dimensions. Our methods NLDA, NKFDA show an encouraging performance.

Table 1. Recognition rates on the ORL database

k	LDA	DLDA	NLDA	NKFDA
2	76.65	80.10	85.47	82.89
3	87.09	87.54	90.91	89.13
4	91.68	91.50	93.86	93.15
5	93.17	94.65	95.45	95.13
6	95.79	96.50	97.13	96.72
7	96.85	97.12	97.54	97.21
8	98.25	99.15	98.95	98.95
9	99.00	99.95	99.15	99.38

5.2 FERET Database

We have to test our method on more complex and challenging datasets such as the *FERET* database. We selected 70 subjects from the *FERET* database [7] with 6 up-right, frontal-view images of each subject. The face images involve much more variations in lighting, facial expressions and facial details. The second line of Fig. 2 shows one subject from the selected data set.

The eye locations are fixed by geometric normalization. The size of face images is normalized to 92×112, and 69 (i.e. c-1) features are extracted. Training and test process are similar to those on the *ORL* database. Similar comparisons between those methods are performed. This time k changes between 2 to 5, and the corresponding averaging recognition rates (%) are shown in table 2.

Table 2. Recognition rates on the FERET database

k	LDA	DLDA	NLDA	NKFDA
2	56.04	63.25	75.20	72.21
3	76.95	76.71	85.64	83.60
4	87.23	88.30	92.79	93.85
5	94.80	94.71	97.34	98.29

5.3 Mixture Database

To test NLDA and NKFDA on large datasets, we construct a mixture database of 125 persons and 985 images, which is a collection of three databases: (a). The *ORL* database (10×40). (b). The *YALE* database (11×15, the third line of Fig. 2 shows one subject). (c). The *FERET* database (6×70). All the images are resized to 92×112. There are facial expression, illumination and pose variations.

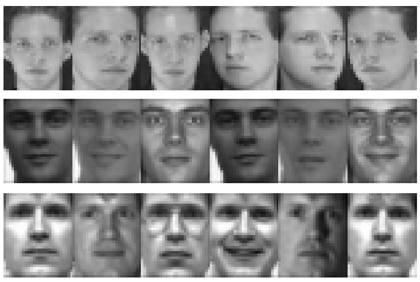

Fig. 2. Samples from the mixture database

The mixture database is divided into two non-overlapping set for training and testing. The training dataset consists of 500 images: 5 images, 6 images and 3 images per person are randomly selected from the *ORL*, the *YALE* database and the *FERET* subset respectively. The remaining 485 images are used for testing. In order to reduce the influence of some extreme illumination, histogram equalization is applied to the images as pre-processing. We compare the proposed method with Fisherface and DLDA, and the experimental results are shown in Fig. 3. It can be seen that NKFDA largely outperforms the other three when over 100 features are used, and a recognition rate of 91.65% can be achieved at the feature dimension of 124 (i.e. c-1).

5.4 Discussion

From the above three experiments, we can find that NKFDA is better than NLDA for large number of training samples (such as larger than 300), while worse than NLDA in the case of small training sample size (such as smaller than 200), and superior to DLDA in most situations. Consequently, NKFDA is more efficient in larger sample size, for the greater the sample size, the more accurately kernels can describe the nonlinear relations of samples.

As to computational cost, the most time-consuming procedure, eigen-analysis, is performed on three matrices (one of $N \times N$, and two of $(N\text{-}c) \times (N\text{-}c)$) in Fisherface method, on two matrices ($c \times c$ and $(c\text{-}1) \times (c\text{-}1)$) in DLDA, on two matrices ($N \times N$, $(N\text{-}1) \times (N\text{-}1)$) in NLDA, and on one matrice ($(N\text{-}1) \times (N\text{-}1)$) in NKFDA. Our method NKFDA only performs one eigen-analysis to achieve efficiency and good performance.

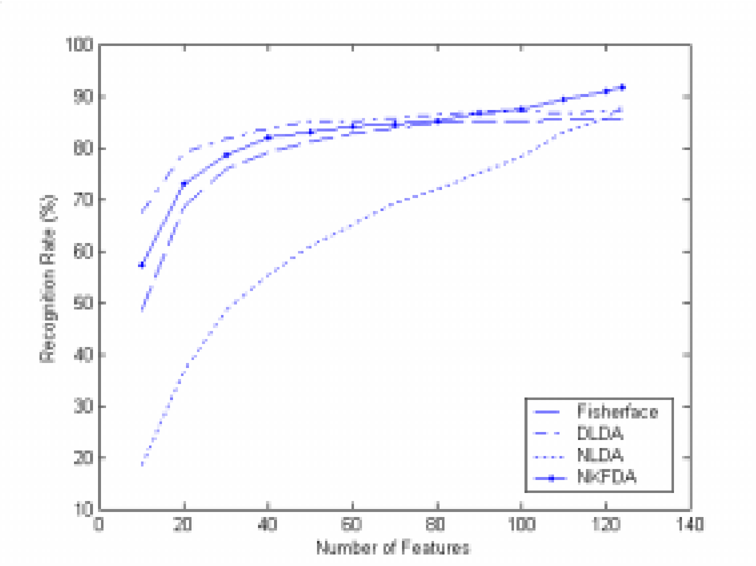

Fig. 3. Comparison of four methods

6 Conclusion

In this paper, we present two new subspace methods (NLDA, NKFDA) based on the null space approach and the kernel technique. Both of them effectively solve the small sample size problem and eliminate the possibility of losing discriminative information.

The main contributions of this paper are summarized as follows: (a) The essence of null space-based LDA in the SSSP is revealed by theoretical justification, and the most suitable situation of null space method is discovered. (b) Propose the NLDA algorithm, which is simpler than all other null space methods and saves the computational cost and maintains the performance simultaneously. (c) A more efficient Cosine kernel function is adopted to enhance the capability of the original polynomial kernel. (d) Present the NKFDA algorithm, which performs only one eigen-analysis and is more stable in numerical computation. (e) NKFDA is also applicable to the large sample size problem, and is superior to NLDA when the sample size is very large.

Acknowledgement This work is supported by research funds from the *Natural Science Foundation of China* (Grant No. 60332010). The authors thank the *Olivetti Research Laboratory* in Cambridge (UK) and the *FERET* program (USA) and the *YALE* University for their devotion of face databases, also thank Qingshan Liu and Zhouyu Fu for their constructive suggestions. Great appreciations especially to Yilin Dong for her encourage.

References

1. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection". IEEE Trans. Pattern Analysis and Machine Intelligence. Volume 19. Issue 7. (1997) 711-720
2. Swets, D.L., Weng, J.: "Using Discriminant Eigenfeatures for Image Retrieval". IEEE Trans. Pattern Analysis and Machine Intelligence. Volume 18. Issue 8. (1996) 831-836
3. Zhao, W., Chellappa, R., Phillips, P.J.: "Subspace Linear Discriminant Analysis for Face Recognition". Technical Report CAR-TR-914, Center for Automation Research, University of Maryland (1999)
4. Yang, J., Yu, H., Kunz, W.: "An Efficient LDA Algorithm for Face Recognition". In: Proceedings of the Sixth International Conference on Control, Automation, Robotics and Vision (2000)
5. Chen, L.F., Liao, H.Y.M., Lin, J.C., Ko, M.T., Yu, G.J.: "A New LDA-based Face Recognition System Which Can Solve the Small Sample Size Problem". Pattern Recognition. Volume 33. Issue 10. (2000) 1713-1726
6. Huang, R., Liu, Q.S., Lu, H.Q., Ma, S.D.: "Solving the Small Sample Size Problem of LDA". In: Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02), Quebec, Canada (2002)
7. Phillips, P.J., Moon, H., Rizvi, S., Rauss, P.: "The FERET Evaluation Methodology for Face Recognition Algorithms". IEEE Trans. Pattern Analysis and Machine Intelligence. Volume 22. Issue 10. (2000) 1090—1104
8. Liu, Q.S., Huang, R., Lu, H.Q., Ma, S.D.: "Face Recognition Using Kernel Based Fisher Discriminant Analysis". In: Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington DC, USA (2002)
9. Baudat, G., Anouar, F.: "Generalized Discriminant Analysis Using a Kernel Approach". Neural Computation. Volume 12. Issue 10. (2000) 2385-2404
10. Mika, S., Ratsch, G., Weston, J.: "Fisher Discriminant Analysis with Kernels". In: Proceedings of Neural Networks for Signal Processing Workshop, Madison, USA (1999)
11. Wu, Y., Huang T.S., Toyama, K.: "Self-Supervised Learning for Object based on Kernel Discriminant-EM Algorithm". In: Proceedings of IEEE International Conference on Computer Vision (ICCV'01), Vancouver BC, Canada (2001)
12. Baudat, G., Anouar, F.: "Kernel-based Methods and Function Approximation". In: Proceedings of International Joint Conference on Neural Networks, Washington DC, USA (2001)
13. Liu, Q.S., Lu, H.Q., Ma, S.D.: "Improving Kernel Fisher Discriminant Analysis for Face Recognition". IEEE Trans. Circuits and Systems for Video Technology. Vol. 14(1) (2004)