Neural Computation 16 (2004) 1283-1297

A Modified Algorithm for Generalized Discriminant Analysis

Wenming Zheng, Li Zhao, Cairong Zou

{wenming_zheng, zhaoli, cairong}@seu.edu.cn

The Engineering Research Center of Information Processing and Application,

Southeast University, Nanjing, Jiangsu 210096, P.R.China

Generalized discriminant analysis (GDA) is an extension of the classical linear discriminant analysis (LDA) from linear domain to a nonlinear domain via the kernel trick. However, in the previous algorithm of GDA, the solutions may suffer from the degenerate eigenvalue problem (i.e., several eigenvectors with the same eigenvalue), which makes them not optimal in terms of the discriminant ability. In this article, we propose a modified algorithm for GDA (MGDA) to solve this problem. The MGDA method aims to remove the degeneracy of GDA and find the optimal discriminant solutions, which maximize the between-class scatter in the subspace spanned by the degenerate eigenvectors of GDA. Theoretical analysis and experimental results on the ORL face database show that the MGDA method achieves better performance than the GDA method.

1 Introduction

Linear discriminant analysis (LDA) (Duda & Hart, 1973) is a well-known feature extraction method in pattern recognition. It finds the set of the optimal vectors that map the high-dimensional samples onto a low-dimensional feature space, where the ratio of the between-class scatter to the within-class scatter of the projected samples is extreme and the projected samples are well separated. Although the LDA method works for linear problem, it fails as for nonlinear case. Baudat and Anouar (2000) extended the LDA method from linear to a nonlinear domain using the kernel trick (Vapnik, 1995; Schölkopf, Smola, & Müller, 1998) and then presented the Generalized Discriminant Analysis (GDA) method. The GDA aims to find the optimal discriminant nonlinear features for the training samples when they are not linearly separable. This is implemented by mapping the input space into a high-dimension (or even infinite-dimension) feature space via a nonlinear kernel function and performing the feature extractions using LDA in this feature space, thus producing the nonlinear features in the input space.

However, in using the Baudat-Anouar algorithm (Baudat & Anouar, 2000) one may face a degenerate eigenvalue problem. This occurs especially in the case of the so-called small sample size (SSS) problem (Chen, Liao, Ko, Lin, & Yu, 2000), where the most discriminant eigenvectors of GDA correspond to the same eigenvalue (i.e., degeneracy of eigenvectorsSchiff; Schiff, 1968).

In this paper, we modify the Baudat-Anouar algorithm and propose a robust and efficient algorithm to overcome the degenerate eigenvalue problem. The proposed

algorithm aims to find the discriminant eigenvectors that maximize the between-class scatter matrix in the subspace spanned by the degenerate eigenvectors of GDA and thus remove the degeneracy of the solutions.

In the next section, we introduce a theorem used for this purpose. Then we review the GDA method. In section 3, we propose the MGDA method and develop the formulation. Section 4 is devoted to the experiments on the ORL face database. The discussion and conclusion are given in the last section.

2 Related Work

A. Related Theorems

Suppose that S_B, S_W and S_T are the between-class scatter matrix, the within-class scatter matrix and the total-class scatter matrix of the training samples, respectively. Let I be the identity matrix.

Theorem 1 (Duchene, & Leclercq, 1988) Let $\varphi_1, \cdots, \varphi_r$ be the first r discriminant eigenvectors of Foley-Sammon Optimal Set of Discriminant Vectors (FSODV; Foley, & Sammon, 1975), Then the $(r+1)$th discriminant direction φ_{r+1} of FSODV is the eigenvector corresponding to maximum eigenvalue of the eigenquation $PS_B\varphi = \lambda S_W \varphi$, where

$$P = I - D^T (DS_W^{-1} D^T)^{-1} DS_W^{-1}, \quad D = [\varphi_1 \quad \varphi_2 \quad \cdots \quad \varphi_r]^T$$

Theorem 2 (Jin, Yang, Hu, & Lou, 2001) Suppose that $\varphi_1, \cdots, \varphi_r$ are the first r discriminant eigenvectors of the statistically uncorrelated optimal discriminant vectors

(UODV; Jin, Yang, Hu, & Lou, 2001). Then the $(r+1)$th discriminant direction φ_{r+1} of UODV is the eigenvector corresponding to maximum eigenvalue of the eigenequation $PS_B\varphi = \lambda S_W\varphi$, where

$$P = I - S_T D^T (DS_T S_W^{-1} S_T D^T)^{-1} DS_T S_W^{-1}, \quad D = [\varphi_1 \quad \varphi_2 \quad \cdots \quad \varphi_r]^T$$

In fact, theorem 1 and theorem 2 can be extended to be a more general form. We give it as the theorem 3. (The proof is given in appendix.)

Theorem 3 Suppose that B and R are positive semi-definite matrices, V is a positive matrix. The discriminant criteria is defined as:

$$F(\varphi) = \frac{\varphi^T B\varphi}{\varphi^T V\varphi} \tag{2.1}$$

Let φ_1 be the discriminant eigenvector that maximizes $F(\varphi)$. Suppose that φ_i ($i = 1,2,\cdots,r \geq 1$) are obtained. Let φ_{r+1} be the $(r+1)$th discriminant eigenvector that maximizes $F(\varphi)$ under the following constraints:

$$\varphi_{r+1}^T R\varphi_i = 0, \quad (i = 1,2,\cdots,r) \tag{2.2}$$

Then, φ_{r+1} is the eigenvector corresponding to the largest eigenvalue of the following eigenequation:

$$PB\varphi = \lambda V\varphi \tag{2.3}$$

where

$$P = I - RD^T (DRV^{-1}RD^T)^{-1} DRV^{-1} \tag{2.4}$$

$$D = [\varphi_1 \quad \varphi_2 \quad \cdots \quad \varphi_r]^T \tag{2.5}$$

B. GDA Formulation in Kernel Space

Suppose that X is an n-dimensional sample set with N elements. Let X_l

denote subset of X. Thus, $X = \bigcup_{l=1}^{c} X_l$, where c is the number of the classes. The cardinality of the subsets X_l is denoted by N_l. Thus, we have $\sum_{l=1}^{c} N_l = N$. Let X be mapped into a Hilbert space F through a nonlinear mapping function Φ,

$$\Phi : X \to F, \quad x \to \Phi(x) \tag{2.6}$$

The between-class scatter matrix S_B^{Φ}, the within-class scatter matrix S_W^{Φ} and the total-scatter matrix S_T^{Φ} in F are given as follows:

$$S_B^{\Phi} = \sum_{i=1}^{c} N_i (u_i^{\Phi} - u^{\Phi})(u_i^{\Phi} - u^{\Phi})^T \tag{2.7}$$

$$S_W^{\Phi} = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\Phi(x_i^j) - u_i^{\Phi})(\Phi(x_i^j) - u_i^{\Phi})^T \tag{2.8}$$

$$S_T^{\Phi} = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\Phi(x_i^j) - u^{\Phi})(\Phi(x_i^j) - u^{\Phi})^T \tag{2.9}$$

where x^T represents the transpose of the vector x, $\Phi(x_i^j)$ is the jth sample in the ith class, u_i^{Φ} is the mean of the ith class samples and u^{Φ} is the mean of all samples in F:

$$u_i^{\Phi} = \frac{1}{N_i} \sum_{j=1}^{N_i} \Phi(x_i^j), \quad u^{\Phi} = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \Phi(x_i^j) \tag{2.10}$$

GDAaims to find eigenvalues λ and eigenvectors ω, solutions of the equation

$$S_B^{\Phi} \omega = \lambda S_T^{\Phi} \omega \tag{2.11}$$

The largest eigenvalue of equation (2.11) gives the maximum of the following quotient of the inertia:

$$\lambda = \frac{\omega^T S_B^{\Phi} \omega}{\omega^T S_T^{\Phi} \omega} \tag{2.12}$$

Because the eigenvectors are linear combinations of F elements, there exist coefficients $\alpha_{pq} (p = 1, \cdots, c; q = 1, \cdots, N_p)$ such that

$$\omega = \sum_{p=1}^{c}\sum_{q=1}^{N_p}\alpha_{pq}(\Phi(x_p^q) - u^{\Phi}) \tag{2.13}$$

Assume that a kernel function $k(x_i, x_j)$ can be expressed as the dot product form on the Hilbert space F:

$$k_{ij} = k(x_i, x_j) =< \Phi(x_i), \Phi(x_j) >= (\Phi(x_i))^T \Phi(x_j) \tag{2.14}$$

where $< \Phi(x_i), \Phi(x_j) >$ stands for the dot product of $\Phi(x_i)$ and $\Phi(x_j)$.

For a given classes p and q, this kernel function can be expressed as:

$$(k_{ij})_{pq} = (\Phi(x_p^i))^T \Phi(x_q^j) \tag{2.15}$$

Let $W = (W_l)_{l=1,\cdots,c}$ be a $N \times N$ block diagonal matrix, where W_l is a $N_l \times N_l$ matrix with all terms equal to $1/N_l$. Let $M = (m_{ij})_{i=1,\cdots,N;j=1,\cdots,N}$ be a $N \times N$ matrix with all terms equal to $1/N$ and let

$$\Phi(X) = [\Phi(x_1^1) \quad \cdots \quad \Phi(x_1^{N_1}) \quad \cdots \quad \Phi(x_c^1) \quad \cdots \quad \Phi(x_c^{N_c})] \tag{2.16}$$

Then equation (2.13) can be expressed as

$$\omega = \Phi(X)(I - M)\alpha \tag{2.17}$$

where $\alpha = [\alpha_{11} \quad \cdots \quad \alpha_{1N_1} \quad \cdots \quad \alpha_{c1} \quad \cdots \quad \alpha_{cN_c}]^T$.

Let K be a $N \times N$ matrix defined on the class elements by $(K_{pq})_{p=1,\cdots,c;q=1,\cdots,c}$:

$$K = (K_{pq})_{p=1,\cdots,c;q=1,,\cdots c} \tag{2.18}$$

where (K_{pq}) is a $N_p \times N_q$ matrix in the feature space F:

$$K_{pq} = (k_{ij})_{pq}, \quad i = 1,\cdots,N_p, \quad j = 1,\cdots,N_q \tag{2.19}$$

Using the notations above, the equations (2.11), (2.12) and (2.18) are equivalent to the following expressions, respectively:

$$B\alpha = \lambda T\alpha \tag{2.20}$$

$$\lambda = \frac{\alpha^T B\alpha}{\alpha^T T\alpha} \tag{2.21}$$

$$K = (\Phi(X))^T \Phi(X) \tag{2.22}$$

where
$$B = (I - M)^T K(W - M)(W - M)^T K(I - M) \tag{2.23}$$

$$T = (I - M)^T K(I - M)(I - M)^T K(I - M) \tag{2.24}$$

Baudat and Anouar (2000) give a method to resolve the eigenequation (2.20) by using the eigenvectors decomposition method (for more details of the algorithm, see Baudat & Anouar, 2000). Although that method can resolve the eigenvectors and the corresponding eigenvalues of the eigenequation (2.20) very well, there is a weakness containing in it which still has not been overcome yet. This is the degenerate perturbation problem of the eigenvectors (i.e., several eigenvectors with the same eigenvalue). This problem often occurs in many cases such as the small sample size problem: Denote the null space of S_W^Φ by $S_W^\Phi(0)$. As for small sample size problem, suppose the solution $\omega \in S_W^\Phi(0)$, then we have $S_W^\Phi \omega = 0$. Note that $S_T^\Phi = S_W^\Phi + S_B^\Phi$, thus we have

$$\omega^T S_T^\Phi \omega = \omega^T S_W^\Phi \omega + \omega^T S_B^\Phi \omega = \omega^T S_B^\Phi \omega \tag{2.25}$$

From equations (2.17) and (2.25), we have

$$\alpha^T T\alpha = \alpha^T B\alpha \Rightarrow \frac{\alpha^T B\alpha}{\alpha^T T\alpha} = 1 \tag{2.26}$$

From equation (2.26), we obtain that all the eigenvectors in the null space of the within-class scatter matrix share the same maximal eigenvalue (= 1).

3 Modified Algorithm for GDA

The GDA method provides an efficient technique to calculate the discriminant eigenvectors in the feature space F. However, in many cases such as the small sample size problem, some of the eigenvectors solved by GDA may be degenerate.

Suppose that $\alpha_k^{(1)},\cdots,\alpha_k^{(t)}$ ($t>1$) are the degenerate eigenvectors of the eigenequation (2.20) sharing the same eigenvalue λ_k, i.e.

$$B\alpha_k^{(i)} = \lambda_k T\alpha_k^{(i)},\ \ i=1,\cdots,t \tag{3.1}$$

From the equations (2.11) and (2.17), equation (3.1) is equivalent to the expressions:

$$S_B^{\Phi}\omega_k^{(i)} = \lambda_k S_T^{\Phi}\omega_k^{(i)},\ \ i=1,\cdots,t \tag{3.2}$$

where

$$\omega_k^{(i)} = \Phi(X)(I-M)\alpha_k^{(i)},\ \ i=1,\cdots,t \tag{3.3}$$

Thus, $\omega_k^{(i)}$ ($i=1,\cdots,t$) are the eigenvectors corresponding to the same eigenvalue λ_k of the eigenequation (2.11). Let Ω denote the subspace spanned by the eigenvectors $\omega_k^{(i)}$ ($i=1,\cdots,t$). Then any vector in Ω is the eigenvector of the eigenequation (2.11) corresponding to the eigenvalue λ_k. Thus, the solutions of GDA may be unstable with respect to changes in training data (model variance) and probably are not optimal in terms of the discriminant ability. The MGDA method overcomes this problem by limiting the attention to the subspace Ω to find the eigenvectors with the best discriminant ability.

Suppose that $\tilde{\omega}_k^{(i)} \in \Omega$ ($i=1,\cdots,t$) are the eigenvectors of the eigenequation (2.11) with best discriminant ability, where

$$\widetilde{\omega}_k^{(i)} = \sum_{j=1}^{t} \omega_k^{(j)} \gamma_{ij} \ (i = 1, \cdots, t) \tag{3.4}$$

Let $A = [\alpha_k^{(1)} \ \cdots \ \alpha_k^{(t)}]$ and $\gamma_i = [\gamma_{i1} \ \cdots \ \gamma_{it}]^T$. Then equation (3.4) can be rewritten as:

$$\widetilde{\omega}_k^{(i)} = \Phi(X)(I - M)A\gamma_i, \ \ i = 1, \cdots, t \tag{3.5}$$

According to the physical meaning of discriminant analysis (Fukunaga, 1990), the projections of the training samples projected by the solutions $\widetilde{\omega}_k^{(i)}$ ($i = 1, \cdots, t$) should have maximal between-class scatter in order to get better discriminant ability. To do this, we should define new discriminant criteria to replace the previous discriminant criteria (equation 2.12). Such a criteria can be expressed as:

$$J_1(\omega) = \frac{\omega^T S_B^{\Phi} \omega}{\omega^T \omega}, \quad \text{where } \omega \in \Omega \tag{3.6}$$

By using the new criteria, these discriminant eigenvectors can be generated in the following forms: the first eigenvector $\widetilde{\omega}_k^{(1)}$ is the one that maximizes $J_1(\omega)$ in Ω; Suppose that the first r discriminant eigenvectors $\widetilde{\omega}_k^{(i)} = \Phi(X)(I - M)A\gamma_i$ $(i = 1, 2, \cdots, r < t)$ are obtained, then the $(r+1)$th eigenvector $\widetilde{\omega}_k^{(r+1)}$ is the one that maximizes $J_1(\omega)$ in Ω under the following orthogonal constraints:

$$(\widetilde{\omega}_k^{(r+1)})^T \widetilde{\omega}_k^{(i)} = 0, i = 1, 2, \cdots, r \tag{3.7}$$

From the equation (3.5), we obtain that to find $\widetilde{\omega}_k^{(1)}$ is equivalent to find the coefficient γ_1 that maximizes $J_2(\gamma)$, where

$$J_2(\gamma) = \frac{\gamma^T \widetilde{B} \gamma}{\gamma^T R \gamma} \tag{3.8}$$

$$\widetilde{B} = A^T B A \tag{3.9}$$

$$R = A^T (I - M)^T K (I - M) A \tag{3.10}$$

From discriminant analysis (Duda, & Hart, 1973; Fukunaga, 1990): γ_1 is the eigenvector of the following generalized eigenequation corresponding to the largest eigenvalue:

$$\widetilde{B}\gamma = \lambda R\gamma \tag{3.11}$$

By the same method, to find the $(r+1)$th eigenvector $\widetilde{\omega}_k^{(r+1)}$ that maximizes $J_1(\omega)$ under the orthogonal constraints of equation (3.7) in Ω is equivalent to find the coefficient γ_{r+1} that maximizes $J_2(\gamma)$ under the following constraints:

$$\gamma_{r+1}^T R\gamma_i = 0\,, i = 1,2,\cdots,r \tag{3.12}$$

From theorem 3, we obtain that γ_{r+1} is the eigenvector corresponding to the largest eigenvalue of the following eigenequation:

$$P\widetilde{B}\gamma = \lambda R\gamma \tag{3.13}$$

where

$$P = I - RD^T(DRD^T)^{-1}D \tag{3.14}$$

$$D = [\gamma_1 \quad \gamma_2 \quad \cdots \quad \gamma_r]^T \tag{3.15}$$

The coefficient γ_i $(i = 1,\cdots,t)$ is normalized by requiring that the corresponding eigenvector $\widetilde{\omega}_k^{(i)}$ $(i = 1,\cdots,t)$ is normalized in F, i.e.

$$(\widetilde{\omega}_k^{(i)})^T \widetilde{\omega}_k^{(i)} = 1 \tag{3.16}$$

Using equations (2.22) and (3.5), we have

$$\gamma_i^T A^T (I-M)^T K(I-M)A\gamma_i = 1 \tag{3.17}$$

Thus the coefficient γ_i is divided by $\sqrt{\gamma_i^T A^T (I-M)^T K(I-M)A\gamma_i}$ in order to get normalized eigenvector $\widetilde{\omega}_k^{(i)}$.

The MGDA procedure can be summarized in the following steps:

1. Compute the discriminant vectors by using the Baudat-Anouar's algorithm. If no degeneracy occurs, then finish the algorithm and the solutions of MGDA equal to those of GDA; else go to step 2.
2. Select the discriminant vectors $\alpha_k^{(i)}$ ($i = 1, \cdots, t$) corresponding to the same eigenvalue λ_k, where t is the number of the degenerate vectors.
3. Compute the matrices $\widetilde{B}$ and R (see equations 3.9 and 3.10).
4. Compute the eigenvector γ_1 using system 3.11.
5. Compute the eigenvectors γ_i ($i = 2, \cdots, t$) using system 3.13.
6. Compute eigenvectors $\widetilde{\omega}_k^{(i)}$ using γ_i ($i = 1, \cdots, t$) (see equation 3.5) and normalize them (see equation 3.17).

4 Experiments

We test the MGDA method on the Olivetti Research Lab. (ORL) face database in Cambridge (Online Available: http://www.cam-orl.co.uk/facedatabase.html). The ORL database contains 40 distinct subjects, where each one contains 10 different images taken at different times, varying lighting slightly. All the images are taken against a dark homogeneous background and the persons are in upright, frontal position, with tolerance for some tilting and rotation. Figure 1 shows ten face images of one subject in the face database. The original face images are all sized 112×92 pixels with a 256-level gray scale. For each image, we use a two-level wavelet transform (Chien & Wu, 2002) and get a low-pass image of 28×23 size pixels.

Then we normalize the intensity values of the low-pass image with a linear function. After doing that, the dimension of the image vector is 644. The mean and standard deviation of Kurtosis of the images are 2.08 and 0.39, respectively. The polynomial and gaussian kernels used in the experiments are defined as (Baudat & Anouar, 2000):

Polynomial kernel: $k(x, y) = (< x, y >)^d$, where d is the polynomial degree;

Gaussian kernel: $k(x, y) = \exp(-\|x - y\|^2 / \sigma)$, where the parameter σ has to be chosen.

Figure 1: Ten face images for one subject in ORL face database

4.1 Face Recognition

Two examples for face recognition based on the nearest neighbor classifier are performed in this experiment. The first example is similar with that done by Yang (2002), which aims to compare the performance of GDA and MGDA with other methods in face recognition. We use the leave-one-out strategy to perform this experiment: To classify an face image, we remove it from the whole face image set, and the discriminant vectors are computed using the training set of the remainder 399

images. The test image and the training images then are projected to a reduced space using the computed discriminant vectors of GDA and MGDA, respectively. Table 1 shows the experimental results. We can see from Table 1 that the MGDA method achieves the error rate as low as the Kernel Fisherface method (= 1.25%), which is the lowest error rate among the methods reported by Yang (2002). We also see that the MGDA method achieves better performance than the GDA method in this example.

Table 1 Performance of various systems

Method	Reduced Space	Error Rate (%)
Eigenface (Yang, 2002)	40	2.50 (10/400)
Fisherface (Yang, 2002)	39	1.50 (6/400)
SVM (Yang, 2002)	N/A	3.00 (12/400)
Kernel Eigenface (Yang, 2002)	40	2.00 (8/400)
Kernel Fisherface (Yang, 2002)	39	1.25 (5/400)
GDA (Polynomial kernel with $d = 2$)	39	2.25 (9/400)
MGDA (Polynomial kernel with $d = 2$)	39	1.5 (6/400)
GDA (Gaussian kernel with $\sigma = 10000$)	39	1.5 (6/400)
MGDA (gaussian kernel with $\sigma = 10000$)	39	1.25 (5/400)

The second example aims to further to compare the performance of GDA and MGDA in face recognition. Ten images per subject are randomly partitioned into five training images and five test images for a total of 200 training images and 200 test

images. There is no overlap between the two sets. Two trials of tests are performed by swapping the training and the test sets. We treat each test as a Bernoulli random test and the average test error rate as the probability of the wrong classification over all the 400 tests. Table 2 shows the average test error rate and the standard deviation for each method, where MLDA is a particular case of GDA using the polynomial kernel with the degree $d = 1$. As can be seen form Table 2, the MGDA achieves lower average test error rate and standard deviation than GDA over all the tests.

Table 2: Average test error rate for GDA and MGDA

Method	LDA	MLDA	GDA [1)]	MGDA [1)]	GDA [2)]	MGDA [2)]
Average Test Error Rate(%)	10.0	5.0	4.75	3.0	4.25	2.75
Standard Deviation	6.0	4.3589	4.2541	3.4117	4.0345	3.2707

Note: [1)] polynomial kernel with $d = 2$ is used; [2)] gaussian kernel with $\sigma = 10000$ is used.

4.2 Stability Test

This experiment aims to compare the performance of stability between GDA and MGDA. We select five images randomly from per subject as training samples and use the other five images as test images, thus obtain 200 training images and 200 test images. The gaussian kernel with $\sigma = 10000$ and the nearest neighbor classifier are used in this experiment.

Suppose that ω_i and $\tilde{\omega}_i$ ($i = 1, \cdots, 39$) are the 39 discriminant eigenvectors computed by GDA and MGDA, respectively. Let $Z_{GDA} = [\omega_1 \quad \cdots \quad \omega_{39}]$ and let

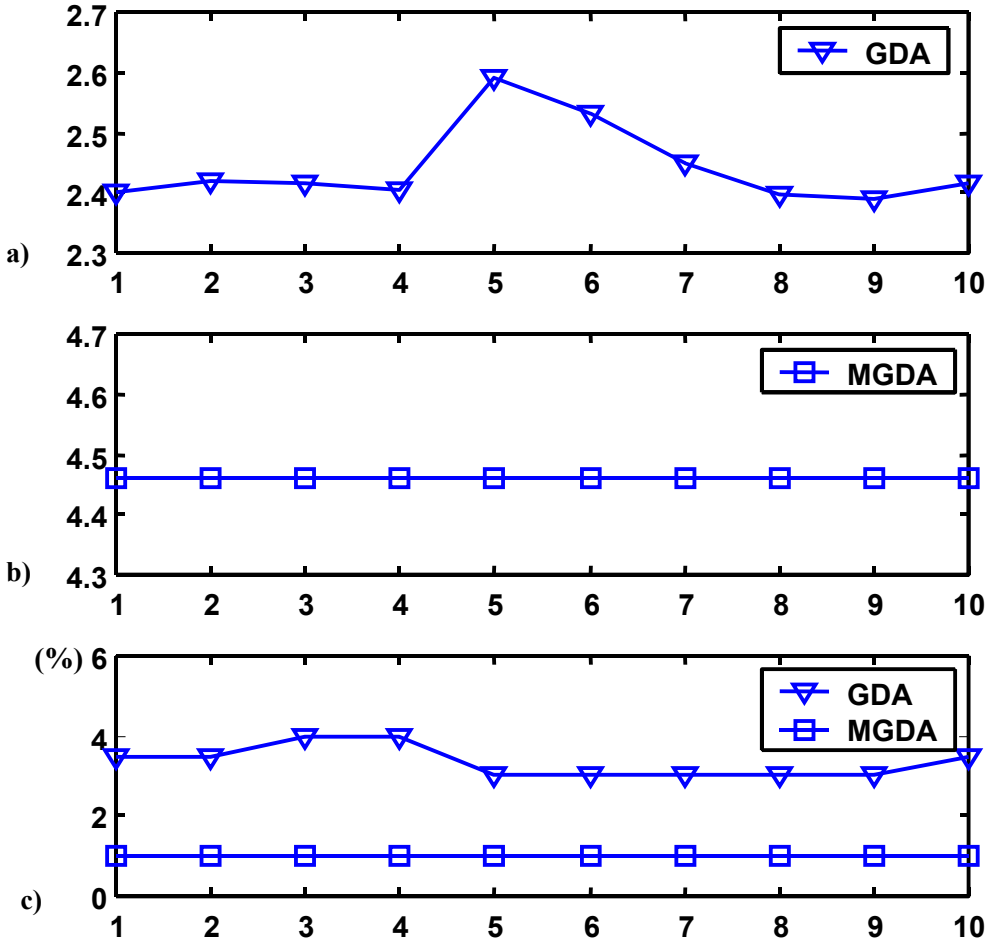

Figure 2: The comparison of the stability between GDA and MGDA. (a) The value of $\Psi(Z_{GDA})$ in each trial; (b) The value of $\Psi(Z_{MGDA})$ in each trial; (c) The test error rate of GDA and MGDA in each trial.

$Z_{MGDA} = [\tilde{\omega}_1 \quad \cdots \quad \tilde{\omega}_{39}]$. We re-order the training images in each subject and then re-compute the projected matrices Z_{GDA} and Z_{MGDA}. Ten trials are repeated in this experiment. In each trial, we write down the value of $\Psi(Z_{GDA})$ and $\Psi(Z_{MGDA})$, respectively, where $\Psi(Z) = tr(Z^T S_B^{\Phi} Z)$. We also use the 200 test images to perform face recognition using the projected matrices Z_{GDA} and Z_{MGDA}, respectively. Figure

2 shows the experimental results of the ten trials. From Figure 2, we can see that the values of $\Psi(Z_{MGDA})$ and the corresponding test error rate are stable over all the ten trials for MGDA. However, it is not the case for GDA. As for GDA, the values of $\Psi(Z_{GDA})$ have slight change over all the trials and the test error rates also have small change in some trials.

5. Disscussion and Conclusion

GDA method is the generalization for LDA method as nonlinear discrimination analysis via the kernel trick. Baudat and Anouar (2000) provide an algebra formulation and the eigenvalue resolution, which can give an exact solution for GDA even if some points, such as the choice of kernel function, require further investigation. However, the further study for GDA shows us that the resolution method for GDA Baudat and Anouar (2000) developed may suffer from instability and inaccuracy due to the degenerate perturbation of the solutions, especially to the small sample size problem where the most discriminant eigenvectors in the null space of the within-class scatter matrix share the same maximal eigenvalue (= 1). In this paper, we have developed a modified algorithm for GDA (MGDA), to overcome this problem. The performance of MGDA is just the same as GDA when no degeneracy occurs. If degeneracy occurs, our theoretical analysis and the experiments based on the ORL face database show that the MGDA method still keeps good performance and is superior to the GDA method.

Besides, in the first example in section 4.1, we see that the MGDA method gets exactly the same performance as the Kernel Fisherface method. In fact, we can see that the Kernel Fisherface algorithm developed by Yang (2002) and the GDA algorithm by Baudat and Anouar (2000) are two different algorithms based on the same criteria (equation 2.12) which solves the discriminant analysis problems in feature space. However, both methods did not overcome the possible degenerate problem of the solutions. Thus, the best performance of MGDA could be exactly equal to that of Kernel Fisherfaces.

Appendix A

Proof of theorem 3:

Proof. Let $\widetilde{\varphi}_{r+1}$ be the direction of φ_{r+1}, and $\widetilde{\varphi}_{r+1}$ satisfies the following constraint

$$\widetilde{\varphi}_{r+1}^T V \widetilde{\varphi}_{r+1} = 1 \tag{A.1}$$

Thus, from equation (2.2), we have

$$\widetilde{\varphi}_{r+1}^T R \varphi_i = 0, \quad i = 1, \cdots, r \tag{A.2}$$

In order to compute $\widetilde{\varphi}_{r+1}$, we use the method of Lagrange multipliers to transform the criteria (2.1) including all the constrains

$$L(\widetilde{\varphi}_{r+1}) = \widetilde{\varphi}_{r+1}^T B \widetilde{\varphi}_{r+1} - \lambda[\widetilde{\varphi}_{r+1}^T V \widetilde{\varphi}_{r+1} - 1] - \sum_{i=1}^{r} \mu_i \widetilde{\varphi}_{r+1}^T R \varphi_i \tag{A.3}$$

where λ and μ_i, $i = 1,2,\cdots,r$ are Lagrange multipliers.

The optimization is performed by setting the partial derivative of $L(\widetilde{\varphi}_{r+1})$ with respect to $\widetilde{\varphi}_{r+1}$ equal to zero:

$$2B\widetilde{\varphi}_{r+1} - 2\lambda V\widetilde{\varphi}_{r+1} - \sum_{i=1}^{r} \mu_i R\varphi_i = \mathbf{0} \tag{A.4}$$

Multiplying the left-hand side of equation (A.4) by $\widetilde{\varphi}_{r+1}^T$, and from equation (A.2), we obtain that:

$$\lambda = \frac{\widetilde{\varphi}_{r+1}^T B\widetilde{\varphi}_{r+1}}{\widetilde{\varphi}_{r+1}^T V\widetilde{\varphi}_{r+1}} = F(\widetilde{\varphi}_{r+1}) \tag{A.5}$$

Thus, λ represents the expression $F(\widetilde{\varphi}_{r+1})$ to be maximized.

Multiplying the left-hand side of equation (A.4) by $\varphi_j^T RV^{-1}$, $j = 1,2,\cdots,r$, we obtain a set of r expression:

$$2\varphi_j^T RV^{-1}B\widetilde{\varphi}_{r+1} - \sum_{i=1}^{r} \mu_i \varphi_j^T RV^{-1}R\varphi_i = 0\,, \quad j = 1,2,\cdots,r \tag{A.6}$$

or in another form:

$$\begin{aligned}
&2\varphi_1^T RV^{-1}B\widetilde{\varphi}_{r+1} - \sum_{i=1}^{r} \mu_i \varphi_1^T RV^{-1}R\varphi_i = 0.\\
&2\varphi_2^T RV^{-1}B\widetilde{\varphi}_{r+1} - \sum_{i=1}^{r} \mu_i \varphi_2^T RV^{-1}R\varphi_i = 0.\\
&\cdots\\
&2\varphi_r^T RV^{-1}B\widetilde{\varphi}_{r+1} - \sum_{i=1}^{r} \mu_i \varphi_r^T RV^{-1}R\varphi_i = 0.
\end{aligned} \tag{A.7}$$

i.e.

$$2\begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_r^T \end{bmatrix} RV^{-1}B\widetilde{\varphi}_{r+1} - \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_r^T \end{bmatrix} RV^{-1}R \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_r^T \end{bmatrix}^T \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix} = 0 \tag{A.8}$$

Let

$$\boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_r]^T \tag{A.9}$$

Considering the matrix notation (2.5), the previous set of r equations (A.8) can be represented in a single matrix relation:

$$DRV^{-1}RD^T\boldsymbol{\mu} = 2DRV^{-1}B\widetilde{\varphi}_{r+1} \tag{A.10}$$

Thus, we obtain

$$\boldsymbol{\mu} = 2(DRV^{-1}RD^T)^{-1}DRV^{-1}B\widetilde{\varphi}_{r+1} \tag{A.11}$$

It is obvious that the following equation holds:

$$\sum_{i=1}^{r}\mu_i R\varphi_i = RD^T\boldsymbol{\mu} \tag{A.12}$$

Thus, equation (A.4) can be written as the following form:

$$2B\widetilde{\varphi}_{r+1} - 2\lambda V\widetilde{\varphi}_{r+1} - RD^T\boldsymbol{\mu} = 0 \tag{A.13}$$

Substituting (A.11) into (A.13), we have

$$2B\widetilde{\varphi}_{r+1} - 2\lambda V\widetilde{\varphi}_{r+1} - RD^T[2(DRV^{-1}RD^T)^{-1}DRV^{-1}B\widetilde{\varphi}_{r+1}] = 0 \tag{A.14}$$

or in another form

$$B\widetilde{\varphi}_{r+1} - RD^T(DRV^{-1}RD^T)^{-1}DRV^{-1}B\widetilde{\varphi}_{r+1} = \lambda V\widetilde{\varphi}_{r+1} \tag{A.15}$$

i.e.

$$[I - RD^T(DRV^{-1}RD^T)^{-1}DRV^{-1}]B\widetilde{\varphi}_{r+1} = \lambda V\widetilde{\varphi}_{r+1} \tag{A.16}$$

Considering the matrix notation (2.4), the equation (A.16) can be expressed as

$$PB\widetilde{\varphi}_{r+1} = \lambda V\widetilde{\varphi}_{r+1} \tag{A.17}$$

Because that $\widetilde{\varphi}_{r+1}$ is the direction of φ_{r+1}, we therefore obtain that

$$PB\varphi_{r+1} = \lambda V\varphi_{r+1} \tag{A.18}$$

Besides, it is noted that φ_{r+1} is unitary vector, thus we have

$$\varphi_{r+1}^T\varphi_{r+1} = 1 \tag{A.19}$$

From the equations (A.18) and (A.19), we know that φ_{r+1} is the eigenvector corresponding to the largest eigenvalue λ of the eigenequation (A.18). □

References

Baudat G., & Anouar F. (2000), "Generalized discriminant analysis using a kernel approach," Neural Computation, vol.12, pp.2385-2404.

Chen, L.F., Liao, H.Y.M., Ko, M.T., Lin, J.C., & Yu, G.J. (2000), "A new LDA-based face recognition system which can solve the small sample size problem," Pattern Recognition, 33, pp.1713-1726.

Chien, J.T. & Wu, C.C. (2002), "Discriminant Waveletfaces and Nearest Feature Classifiers for Face Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 12, pp.1644-1649.

Duchene, J., & Leclercq, A. (1988), "An Optimal Transformation for Discriminant and Principal Component Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.10, No.6, pp. 978-983.

Duda, R.O., & Hart, P.E. (1973), "Pattern Classification and Scene Analysis", New York: John Wiley & Sons, Inc.

Foley, D.H., & Sammon, J.W., JR. (1975), "An Optimal Set of Discriminant Vectors," IEEE Trans. on Computer, vol. C-24, No. 3, pp.281-289.

Fukunaga, K. (1990), "Introduction to Statistical Pattern Recognition", Academic Press, Inc.

Jin, Z., Yang, J.Y., Hu, Z.S., & Lou, Z., (2001), "Face recognition based on the uncorrelated discriminant transformation," Pattern Recognition, 34, pp. 1405-1416.

Schiff L. (1968), "Quantum Mechanics", 3rd ed., McGraw-Hill, New York.

Schölkopf, B., Smola, A., & Müller, K.R. (1998), "Nonlinear component analysis as a

kernel eigenvalue problem," Neural Computation, vol.10, pp.1299-1319.

Vapnik, V. (1995), "The Nature of Statistical Learning Theory: Springer, 1995.

Yang, M.H. (2002), "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods", Proceedings of the Fifth International conference on Automatic Face and Gesture Recognition (FG 2002), Washington D. C., May, IEEE Computer Society, pp.215-220.