Dual-Space Linear Discriminant Analysis for Face Recognition

Xiaogang Wang and Xiaoou Tang

Department of Information Engineering
The Chinese University of Hong Kong
{xgwang1, xtang}@ie.cuhk.edu.hk

Abstract

Linear Discriminant Analysis (LDA) is a popular feature extraction technique for face recognition. However, it often suffers from the small sample size problem when dealing with the high dimensional face data. Some approaches have been proposed to overcome this problem, but they are often unstable and have to discard some discriminative information. In this paper, a dual-space LDA approach for face recognition is proposed to take full advantage of the discriminative information in the face space. Based on a probabilistic visual model, the eigenvalue spectrum in the null space of within-class scatter matrix is estimated, and discriminant analysis is simultaneously applied in the principal and null subspaces of the within-class scatter matrix. The two sets of discriminative features are then combined for recognition. It outperforms existing LDA approaches.

1. Introduction

LDA is a popular face recognition approach. It determines a set of projection vectors maximizing the between-class scatter matrix (S_b) while minimizing the within-class scatter matrix (S_w) in the projective feature space. However, LDA often suffers from the small sample size problem when dealing with the high dimensional face data. When there are not enough training samples, S_w may become singular, and it is difficult to compute the LDA vectors.

Several approaches have been proposed to address this problem. In a two-stage PCA+LDA approach [1], the data dimensionality is first reduced by Principal Component Analysis (PCA), and LDA is performed in the reduced PCA subspace, in which S_w is non-singular. However, Chen et al. [2] suggested that the null space spanned by the eigenvectors of S_w with zero eigenvalues contains the most discriminative information. A LDA method in the null space of S_w was proposed. It chooses the projection vectors maximizing S_b with the constraint that S_w is zero. But this approach discards the discriminative information outside the null space of S_w. Yu. et al. [3] proposed a direct LDA algorithm. It first removes the null space of S_b, and assumes that no discriminative information exists in this space. Unfortunately, we can show that this assumption is incorrect. We will demonstrate that the optimal discriminant vectors do not necessarily lie in the subspace spanned by the class centers. A common problem with all these proposed LDA approaches is that they all lose some discriminative information in the high dimensional face space.

In this paper, using the probabilistic visual model [4], the eigenvalue spectrum in the null space of S_w is estimated. We then apply discriminant analysis in both the principal and null subspaces of S_w. The two parts of discriminative features are combined in recognition. This dual-space LDA approach successfully resolves the small sample size problem. Compared with conventional LDA approaches, it is more stable and makes use of all the discriminative information in both the principal and null space. The experiments on FERET database clearly demonstrate its efficacy.

2. Linear Discriminant Analysis

2.1. LDA

LDA method tries to find a set of projection vectors W maximizing the ratio of determinant of S_b to S_w,

$$W = \arg\max \left| \frac{W^T S_b W}{W^T S_w W} \right| \tag{1}$$

Let the training set contain L classes and each class X_i has n_i samples. S_w and S_b are defined as,

$$S_w = \sum_{i=1}^{L} \sum_{\vec{x}_k \in X_i} (\vec{x}_k - \vec{m}_i)(\vec{x}_k - \vec{m}_i)^T \ , \tag{2}$$

$$S_b = \sum_{i=1}^{L} n_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T \ , \tag{3}$$

where $\vec{m}$ is the center of the whole training set, $\vec{m}_i$ is the center for the class X_i, and $\vec{x}_k$ is the sample belonging to class X_i. W can be computed from the eigenvectors of $S_w^{-1} S_b$ [7]. However, when the small sample size problem occurs, S_w becomes singular and it is difficult to compute S_w^{-1}. To avoid the singularity of S_w, a two-

1063-6919/04 $20.00 © 2004 IEEE

stage PCA+LDA approach is used in [1]. PCA is first used to project the high dimensional face data into a low dimensional feature space. Then LDA is performed in the reduced PCA subspace, in which S_w is non-singular.

Fukunnaga [7] has proved that W can also be computed from simultaneous diagonalization of S_w and S_b. First S_w is whitened by

$$\Theta^{-1/2}\Phi^T S_w \Phi\Theta^{-1/2} = I\,, \qquad (4)$$

where Φ, and Θ are the eigenvector matrix and eigenvalue matrix of S_w. To avoid singularity and overfitting for noise, only the eigenvectors with non-zero and non-trivial eigenvalues are selected in the enhanced LDA model proposed by Liu et. al. [6]. Second, apply PCA on class centers of the transformed data. To do this, the class centers are projected onto $\Phi\Theta^{-1/2}$, and the between-class scatter matrix is transformed to K_b,

$$K_b = \Theta^{-1/2}\Phi^T S_b \Phi\Theta^{-1/2}\,. \qquad (5)$$

After computing the eigenvector matrix Ψ and eigenvalue matrix Λ of K_b, the overall projection vectors of LDA can be defined as

$$W = \Phi\Theta^{-1/2}\Psi\,. \qquad (6)$$

2.2. LDA in the Null Space of S_w

The LDA approaches described above are all performed in the principal subspace of S_w, in which $W^T S_w W \neq 0$. However, the null space of S_w, in which $W^T S_w W = 0$, also contains much discriminative information, since it is possible to find some projection vectors W satisfying $W^T S_w W = 0$ and $W^T S_b W \neq 0$, thus the Fisher criteria in Eq. (1) definitely reaches its maximum value. A LDA in the null space of S_w was proposed by Chen et. al. [2]. First, the null space of S_w is computed as,

$$V^T S_w V = 0 \ (V^T V = I\,). \qquad (7)$$

Then S_b is projected to the null space of S_w,

$$\tilde{S}_b = V^T S_b V\,. \qquad (8)$$

Choose the eigenvectors U of $\tilde{S}_b$ with the largest eigenvalues Λ,

$$U^T \tilde{S}_b U = \Lambda\,. \qquad (9)$$

The LDA transformation matrix is defined as $W = VU$.

As the rank of S_w increases, the null space of S_w becomes small, and much discriminative information outside the null space is discarded [2].

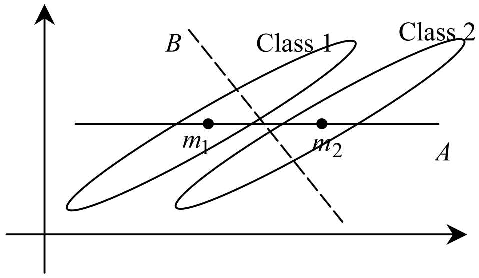

Figure 1. Using direct LDA, the discriminant vector is constrained to the line A passing through the two class centers m_1 and m_2, but according to the Fisher criteria, the optimal discriminant projection should be line B

2.3. Direct LDA

Yu et. al. [3] proposed a direct LDA method. S_b is first diagonalized, and the null space of S_b is removed,

$$Y^T S_b Y = D_b > 0\,, \qquad (10)$$

where Y are eigenvectors and D_b are the corresponding non-zero eigenvalues of S_b. S_w is transformed to

$$K_w = D_b^{-1/2} Y^T S_w Y D_b^{-1/2}\,. \qquad (11)$$

K_w is diagonalized by eigenanalysis,

$$U^T K_w U = D_w\,. \qquad (12)$$

The LDA transformation matrix for classification is defined as,

$$W = Y D_b^{-1/2} U D_w^{-1/2}\,. \qquad (13)$$

In direct LDA, the null space of S_b is first removed. It is assumed that the null space of S_b contains no discriminative information. This assumption is not true. In direct LDA, projection vectors are restricted in the subspace spanned by class centers. However the optimal discriminant vectors do not necessarily lie in the subspace spanned by class centers. This point can be clearly illustrated in Figure 1. For a binary classification problem, using direct LDA, the derived discriminant projection vector is constrained to the line passing through the two class centers. But according to the Fisher criteria, the optimal discriminant vector should be line B.

2.4. Discussion

All these proposed LDA approaches lose some discriminative information in the face space. This point can be further summarized in Figure 2, where **A** is the principal subspace of S_w, and **B** is the principal subspace of S_b. Since the total scatter matrix S_t is equal to the summarization of S_w and S_b [7],

$$S_t = S_w + S_b\,, \qquad (14)$$

1063-6919/04 $20.00 © 2004 **IEEE**

the face space is composed of **A** and **B**. When $\mathbf{B} \subseteq \mathbf{A}$ as shown in Figure 2 (a), LDA in the principal subspace of S_w, contains all the discriminative information in the face space. When $\mathbf{A} \subseteq \mathbf{B}$ as shown in Figure 2 (b), direct LDA, working in subspace **B**, can keep all the discriminative information. When $\mathbf{A} \cap \mathbf{B} = \phi$ as shown in Figure 2 (c), LDA in the null space of S_w can keep all the discriminative information. Finally when **A** and **B** are only partially overlapped as shown in Figure 2 (d), some discriminative information will definitely be lost using any of the existing LDA approaches.

Furthermore, conventional LDA approaches suffer from the overfitting problem. Projection vectors are tuned to the training set with the existence of noise. As suggested in [8], an eigenvector will be very sensitive to small perturbation if its eigenvalue is close to another eigenvalue of the same matrix, so the eigenvectors of S_w with very small eigenvalues are unstable. In Eq. (6) and (13), LDA in the principal subspace of S_w and direct LDA all need to be whitened using the inverse of eigenvalues of S_w. Since the trivial eigenvalues sensitive to noise are not well estimated because of the small sample size problem, they can substantially change the projection vectors. If data vector is whitened on noisy eigenvectors, overfitting will happen.

For LDA in the null space of S_w, the rank of S_w, $r(S_w)$, is bounded by $\min(M-L, N)$, where M is the total training sample number, L is the class number, and N is the dimensionality of the face data. $r(S_w)$ is almost equal to this bound because of the existence of noise. As shown by experiments in [2], when the training sample number is large, the null space of S_w becomes small, so much discriminative information outside it will be lost.

3. Dual-Space LDA

3.1. Probabilistic model

In LDA, the main difficulty for the small sample size problem is that S_w, especially the eigenvalue spectrum in the null space of S_w, is not well estimated. In order to estimate the eigenvalue spectrum in the null space of S_w, we use the probabilistic visual model proposed in [4]. In classification, the likelihood of an input vector x belonging to class X_j is often estimated with a Gaussian density,

$$P(x|X_j) = \frac{1}{(2\pi)^{N/2}|\Sigma_j|}\exp\left[-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)\right], \quad (15)$$

where Σ_j is the covariance matrix for X_j, and N is the dimensionality of the face vector. When there are not enough samples in each class, Σ_j is replaced by S_w [5],

$$P(x|X_j) = \frac{1}{(2\pi)^{N/2}|S_w|}\exp\left[-\frac{1}{2}(x-m_j)^T S_w^{-1}(x-m_j)\right]. \quad (16)$$

It is called Mahalanobis likelihood [5], characterized by a Mahalanobis distance,

$$d(x) = \tilde{x}^T S_w^{-1} \tilde{x}, \quad (17)$$

where $\tilde{x} = x - m_j$. S_w can be diagonalized as,

$$S_w = \Phi \left(\begin{array}{cccccc} \lambda_1 & & & & & \\ & \ddots & & & 0 & \\ & & \lambda_{N_T} & & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{array} \right) \Phi^T, \quad (18)$$

where $\Phi = [\phi_1, \ldots, \phi_N]$ are the eigenvector matrix of S_w. The Mahalanobis distance can be estimated as the sum of two independent parts, "distance-in-feature-space" (DIFS) and "distance-from-feature space" (DFFS), corresponding to the principal subspace $F = \{\phi_i\}_{i=1}^K$, spanned by the K eigenvectors with the largest eigenvalues, and its orthogonal complement $\overline{F} = \{\phi_i\}_{i=K+1}^N$,

$$d(x) = \sum_{i=1}^{K} \frac{y_i^2}{\lambda_i} + \frac{\varepsilon^2(\tilde{x})}{\rho}, \quad (19)$$

where y_i is the component projecting $\tilde{x}$ to the ith eigenvector, and λ_i is the corresponding eigenvalue in F. $\varepsilon^2(\tilde{x})$ is the PCA reconstruction error of $\tilde{x}$ in $\overline{F}$. The eigenvalues in $\overline{F}$ are not well estimated and have zero values. As suggested in [4], they are simply the estimated noise spectrum, and tend to be the "flattest" portion of eigenvalue spectrum. So they can be estimated by a single value ρ,

$$\rho = \frac{1}{N-K}\sum_{i=K+1}^{N} \lambda_i^* . \quad (20)$$

The unknown λ_i^* in $\overline{F}$ are estimated by fitting a nonlinear function to the available portion of the eigenvalue spectrum in F.

Now the within-class scatter matrix can be estimated as,

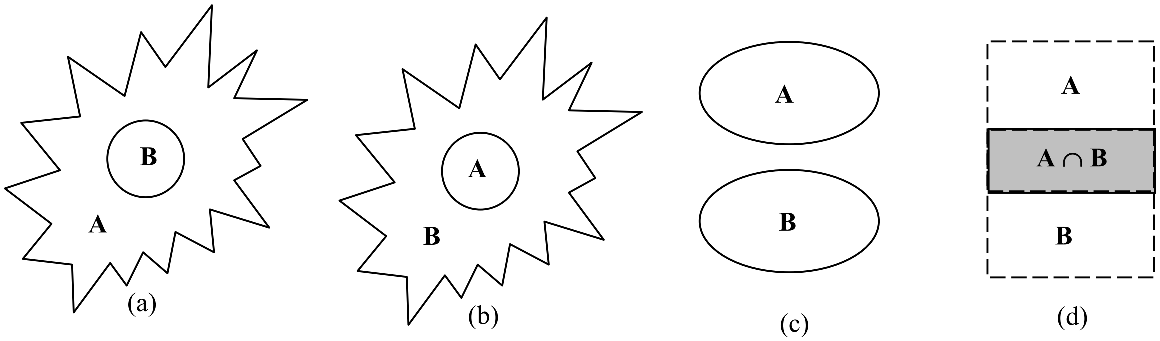

Figure 2. **A** is the principal subspace of S_w, **B** is the principal subspace of S_b, and **A** ∪ **B** is the whole face space. In case (a), **B** ⊆ **A**, LDA in the principal space of S_w can keep all the discriminative information. In case (b), **A** ⊆ **B,** direct LDA can keep all the discriminative information. In case (c), **A** ∩ **B** = ϕ, LDA in the null space of S_w can keep all the discriminative information. In case (d), **A** and **B** are only partially intersect, and so discriminative information in data space will definitely be lost in conventional LDA approaches.

$$\hat{S}_w = \Phi \left(\begin{array}{cccccc} \lambda_1 & & & & & \\ & \ddots & & & 0 & \\ & & \lambda_{K_T} & & & \\ & & & \rho & & \\ & 0 & & & \ddots & \\ & & & & & \rho \end{array} \right) \Phi^T \qquad (21)$$

$\hat{S}_w$ is now non-singular.

3.2. Discriminant Analysis in Dual Space

Both LDA and the Mahalanobis distance in (19) evaluate the distance from the input pattern to the class center after projecting to the subspace. Recall that LDA in the principal subspace of S_w can be divided into two steps. In the first whitening step, face data is projected onto Φ and normalized by $\Theta^{-1/2}$. Most of the within-class variation concentrates on several largest eigenvectors in Φ. Since Θ represents the energy distribution of within-class variation, the whitening process effectively reduces the within-class variation. This process is essentially the same as the Mahalanobis distance in F. Then, PCA is applied on the whitened class centers to find the dominant direction that best separates the class centers. This process further reduces the noise and compacts the discriminative features onto a small number of principal components.

It is also easy to see that the null space of S_w is identical to $\overline{F}$. LDA in the null space of S_w essentially extracts the discriminative features in $\overline{F}$. $\overline{F}$ has removed most of the within-class variation, and LDA further separates the class centers by PCA.

The relationship between the Mahalanobis distance and LDA implies that LDA can be simultaneously applied to the principal and null subspaces of S_w, and the two parts of discriminative features can be combined to make full use of the discriminative information in the face space. Based on this observation, we develop a dual-space LDA algorithm:

At training stage,

1. Compute S_w and S_b from the training set.
2. Apply PCA to S_w, and compute the principal subspace F, with K eigenvectors $V = [\phi_1, \ldots, \phi_K]$, and its complementary subspace $\overline{F}$. Estimate the average eigenvalue ρ in $\overline{F}$.
3. All of the class centers are projected to F and normalized by the K eigenvalues. S_b is transformed to

$$K_b^P = \Lambda^{-1/2} V^T S_b V \Lambda^{-1/2}, \qquad (22)$$

where Λ is the eigenvalue matrix for F. Apply PCA to K_b^P, and compute l_P eigenvectors Ψ_P with the largest eigenvalues. The l_P discriminative vectors in F are defined as

$$W_P = V \Lambda^{-1/2} \Psi_P . \qquad (23)$$

4. Project all the class centers to $\overline{F}$ and compute the reconstruction difference as

$$A_r = A - VV^T A = \left(I - VV^T\right) A \qquad (24)$$

where $A = [\vec{m}_1, \ldots, \vec{m}_L]$ is the class centers matrix. In fact, A_r is the projection of A into $\overline{F}$. In $\overline{F}$, S_b is transformed to

$$K_b^C = \left(I - VV^T\right) S_b \left(I - VV^T\right). \tag{25}$$

Compute l_C eigenvectors Ψ_C of K_b^C with the largest eigenvalues. The l_C discriminative vectors in the $\overline{F}$ are defined as

$$W_C = \left(I - VV^T\right) \Psi_C . \tag{26}$$

At the recognition stage,

1. All the face class centers $\{m_j\}$ in the gallery and the probe face data x_t are projected to the discriminant vectors in F and $\overline{F}$ to get,

$$\vec{a}_j^P = W_P^T m_j , \tag{27}$$

$$\vec{a}_j^C = W_C^T m_j . \tag{28}$$

$$\vec{a}_t^P = W_P^T x_t , \tag{29}$$

$$\vec{a}_t^C = W_C^T x_t . \tag{30}$$

2. Class is found to minimize the distance measure

$$d(\Delta) = \left\| \vec{a}_j^P - \vec{a}_t^P \right\|^2 + \left\| \vec{a}_j^C - \vec{a}_t^C \right\|^2 / \rho . \tag{31}$$

This dual-space LDA algorithm has several advantages over conventional LDA algorithms. First, it takes advantage of all the discriminative information in the full face space while other LDA approaches all lose some discriminative information one way or the other. In the principal and null subspaces of S_w, LDA vectors can be computed using different criterions,

$$\begin{cases} W_P = \arg\max \dfrac{\left| W_P^T S_b W_P \right|}{\left| W_P^T S_w W_P \right|} , \\ W_P^T S_w W_P \neq 0 \end{cases} \tag{32}$$

$$\begin{cases} W_C = \arg\max \left| W_C^T S_b W_C \right| \\ W_C^T S_w W_C = 0 \end{cases} . \tag{33}$$

Both of the two sets of features contain discriminative information for recognition. However, the two subspaces have different metric scales. The principal subspace of S_w has been whitened, so projection vectors in W_P are not orthornormal. It is not suitable, at least not optimal, to combine the distances in the two subspaces directly. In dual-space LDA, the null space of S_w is also whitened by the average eigenvalues. In Eq.(34), $\left\| \vec{a}_j^P - \vec{a}_t^P \right\|^2$ and $\left\| \vec{a}_j^C - \vec{a}_t^C \right\|^2 / \rho$ are computed under the same metric scale measure, and the distances in the two subspaces are similarly whitened by the eigenvalue spectrum of S_w.

The second advantage of the new algorithm is that it is more stable and insensitive to noise than existing LDA approaches. Since eigenvectors of S_w with very small eigenvalues are unstable and sensitive to small perturbation, we avoid computing these unstable eigenvectors by grouping them into $\overline{F}$. In addition, the eigenvalue spectrum of S_w is better estimated, thus it avoids whitening with very small eigenvalues.

Finally, this approach can be viewed as an improvement to the Mahalanobis likelihood computed from the subspace estimation. It is more effective for classification and more efficient in computation. Besides effective reducing the within-class variation like the Mahalanobis distance, it further distances class centers, and removes some noise disturbance by compacting the discriminative features. It is also much faster than the Mahalanobis distance. Computing the reconstruction error $\varepsilon^2(x)$ in Eq. (20) is expensive. Its computational cost is comparable to the correlation between the two original high dimensional face data vectors. Our approach only needs to compute the distances between vectors of $l_P + l_C$ dimensions.

4. Experiment

In this section, we apply the dual-space LDA to face recognition and compare with conventional approaches. by experiments on the data sets from the FERET face database [9]. The high dimensional image intensity vector is used as input pattern for classification. All the 1195 people from the FERET Fa/Fb data set are used in the experiment. There are two face images for each person. 495 people are used for training, and the remaining 700 people are used for testing. For each testing people, one face image is in the gallery and the other is for probe.

First, we compare the dual-space LDA with the Mahalanobis distance estimated by the probabilistic visual model. Figure 3 reports their Top 1 recognition accuracies with different feature numbers. The feature number for the dual-space LDA is the summation of discriminant feature numbers in F and $\overline{F}$. The feature number for the Mahalanobis distance is the dimensionality (*K*) of F. Three distance measures, DIFS, DFFS and DIFS+DFFS, for the Mahalanobis distance, are evaluated. Notice that only for DIFS. the feature number relates to computation cost. Even for a small *K*, the computation cost of DFFS and DIFS+DFFS are very high, since they need to compute the reconstruction error

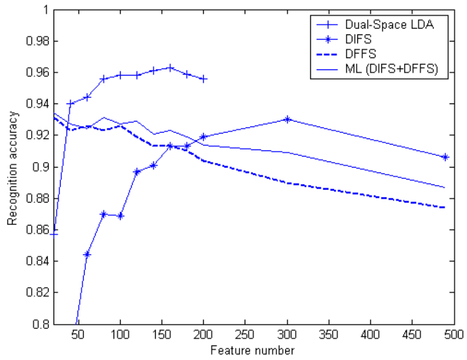

Figure 3. Recognition accuracy comparison between the Dual-Space LDA and the Mahalanobis distances.

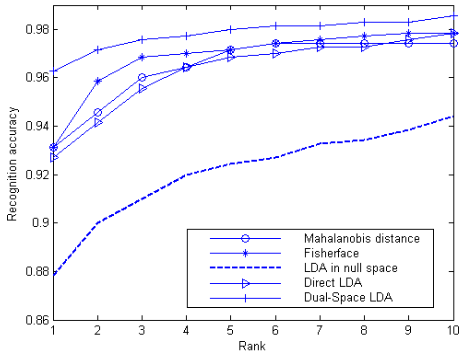

Figure 4 Accumulative matching scores for the dual-space LDA, Mahanobis distance, Fisherface, LDA in null space, and direct LDA on the FERET database.

$\varepsilon^2(x)$. Dual-space LDA outperforms DIFS significantly at the same computational cost. It achieves over 96% recognition accuracy. The best performance for the three Mahalanobis distance measures is about 93%. This is over 40% reduction in recognition error rate.

The dual-space LDA also outperforms conventional LDA approaches. Figure 4 reports the accumulative matching scores comparing the Dual-Space LDA with Fisherface, which uses two stage PCA+LDA, LDA in the null space of S_w, and direct LDA. The novel method has reduced 50% error rate than conventional approaches.

5. Conclusion

In this paper, a dual-space LDA approach for high dimensional data classification is proposed. Compared with existing LDA approaches, it is more stable and makes use of all the discriminative information in the face space. Experiments on the FERET face database have shown that the method is much more effective that existing LDA methods. In future study, we will investigate the application of this dual-space approach to the unified subspace analysis [10] and face sketch recognition [11], and further compare with the random sampling LDA, which combine the two LDA subspaces at the decision level [12].

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region (Project no. CUHK 4190/01E and CUHK 4224/03E).

Reference

[1] P. N. Belhumeur, J. Hespanda, and D. Kiregeman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection,” *IEEE Trans. on PAMI,* Vol. 19, No. 7, pp. 711-720, July 1997.

[2] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, “A New LDA-Based Face Recognition System Which can Solve the Small Sample Size Problem,” Journal of Pattern Recognition, Vol. 33, No. 10, pp. 1713-1726, Oct. 2000.

[3] H. Yu and J. Yang, “A Direct LDA Algorithm for High-Dimensional Data - with Application to Face Recognition,” *Pattern Recognition,* Vol. 34, pp. 2067-2070, 2001.

[4] B. Moghaddam and A. Pentland, “Probabilistic Visual Learning for Object Representation,” *IEEE Trans. on PAMI*, Vol. 19, No. 7, pp. 775-779, July, 1997.

[5] W. Hwang and J. Weng, "Hierarchical Discriminant Regression," *IEEE Trans. on PAMI*, Vol. 22, No. 11, pp. 1277-1293, Nov. 2000.

[6] C. Liu and H. Wechsler, “Enhanced Fisher Linear Discriminant Models for Face Recognition,” *Proceedings of ICPR,* Vol. 2, pp. 1368-1372, 1998.

[7] K. Fukunnaga, “Introduction to Statistical Pattern Recognition,” Academic Press, second edition, 1991.

[8] G. W. Stewart, “Introduction to Matrix Computations,” Academic Press, New York, 1973.

[9] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET Evaluation," in *Face Recognition: From Theory to Applications*, H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie, and T.S. Huang, Eds., Berlin: Springer-Verlag, 1998.

[10] X. Wang and X. Tang, “Unified Subspace Analysis for Face Recognition,” in *Proceedings of ICCV*, pp. 679-686, Nice, France, Oct. 2003.

[11] X. Tang, and X. Wang, “Face Sketch Synthesis and Recognition,” in *Proceedings of ICCV*, pp. 687-694, Nice, France, Oct. 2003.

[12] X. Wang and X. Tang, “Random Sampling LDA for Face Recognition,” in *Proceedings of CVPR*, Washington D.C., USA, 2004.