Face Recognition By Stepwise Nonparametric Margin Maximum Criterion

Xipeng Qiu and Lide Wu
Department of Computer Science and Engineering
Fudan University, China
xpqiu,ldwu@fudan.edu.cn

Abstract

Linear Discriminant Analysis (LDA) is a popular feature extraction technique in face recognition. However, it often suffers from the small sample size problem when dealing with the high dimensional data. Moreover, while LDA is guaranteed to find the best directions when each class has a Gaussian density with a common covariance matrix, it can fail if the class densities are more general. In this paper, a new nonparametric linear feature extraction method, stepwise nonparametric margin maximum criterion(SNMMC), is proposed to find the most discriminant directions, which does not assume that the class densities belong to any particular parametric family and does not depend on the nonsingularity of the within-class scatter matrix either. On three datasets from ATT and FERET face databases, our experimental results demonstrate that SNMMC outperforms other methods and is robust to variations of pose, illumination and expression.

1. Introduction

The curse of high-dimensionality is a major cause of the practical limitations of many pattern recognition technologies, such as face recognition. Linear discriminant analysis (LDA) [4] is a very popular and powerful method for face recognition [16].

The purpose of LDA is to maximize the between-class scatter S_b while simultaneously minimizing the within-class scatter S_w. It can be formulated by Fisher Criterion [4].

A major drawback of LDA is that it 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[8, 1, 3, 15] have been proposed to address this problem. A common drawback of all these proposed variant LDA approaches is that they all lose some discriminative information in the preprocess of dimensionality reduction.

Another disadvantage of LDA is that it assumes each class has a Gaussian density with a common covariance matrix. LDA guaranteed to find the best directions when the distributions are unimodal and separated by the scatter of class means. However, if the class distributions are multimodal and share the same mean, it fails to find the discriminant direction [4]. Besides, the rank of S_b is $c-1$, where c is the number of class. So the number of extracted features is, at most, $c-1$. However, unless a posteriori probability function are selected, $c-1$ features are suboptimal in Bayes sense, although they are optimal with regard to Fisher criterion [4]. Besides, since nonparametric methods (such as nearest neighbor classification), are often used in face recognition, it is ill-suited to extract features with parametric methods (such as LDA).

In this paper, a new feature extraction method, stepwise nonparametric margin maximum criterion(SNMMC), is proposed. SNMMC is to find the linear transform that maximizes the distance between classes, while to minimize the distances among the samples of a single class. SNMMC can be regarded as an extension of nonparametric discriminant analysis [5], but it doesn't depend on the nonsingularity of the within-class scatter matrix. Moreover, SNMMC finds the important discriminant directions without assuming the class densities belong to any particular parametric family.

The rest of the paper is organized as follows: Section 2 gives the review and analysis of the current existing variant LDA methods. Then we describe stepwise nonparametric margin maximum criterion in Section 3. Experimental evaluations of our method, existing variant LDA methods and the other state-of-art face recognition approaches are presented in Section 4. Finally, we give the conclusions in Section 5.

2. Review and Analysis of Variant LDA Methods

The purpose of LDA is to maximize the between-class scatter while simultaneously minimizing the within-class scatter.

The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined as

$$S_b = \sum_{i=1}^{c} p_i(m_i - m)(m_i - m)^T \qquad (1)$$

$$S_w = \sum_{i=1}^{c} p_i S_i, \tag{2}$$

where c is the number of classes; m_i and p_i are the mean vector and a priori probability of class i, respectively; $m = \sum_{i=1}^{c} p_i m_i$ is the total mean vector; S_i is the covariance matrix of class i.

LDA method tries to find a set of projection vectors $W \in R^{D\times d}$ maximizing the ratio of determinant of S_b to S_w,

$$W = \arg\max_W \frac{|W^T S_b W|}{|W^T S_w W|}, \tag{3}$$

where D and d are the dimensionalities of the data before and after the transformation respectively.

From Eq.(3), the transformation matrix W must be constituted by the d eigenvectors of $S_w^{-1} S_b$ corresponding to its first d largest eigenvalues [4].

However, when the small sample size problem occurs, S_w becomes singular and S_w^{-1} does not exist. Moreover, if the class distributions are multimodal or share the same mean, it can fail to find the discriminant direction[4]. Many methods have been proposed for solving the above problems. In following subsections, we give more detailed review and analysis of these methods.

2.1. Methods Aimed at Singularity of S_w

In recent years, many researchers have noticed the problem about singularity of S_w and tried to overcome the computational difficulty with LDA.

To avoid the singularity of S_w, a two-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. But this method is obviously suboptimal due to discarding much discriminative information.

Liu *et al.* [8] modified Fisher's criterion by using the total scatter matrix $S_t = S_b + S_w$ as the denominator instead of S_w. It has been proven that the modified criterion is exactly equivalent to Fisher criterion. However, when S_w is singular, the modified criterion reaches the maximum value, namely 1, for any transformation W in the null space of S_w. Thus the transformation W cannot guarantee the maximum class separability $|W^T S_b W|$ is maximized [7]. Besides, this method still need calculate an inverse matrix, which is time consuming. Chen *et al.* [3] suggested that the null space spanned by the eigenvectors of S_w with zero eigenvalues contains the most discriminative information. A LDA method (called NLDA) 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. Thus, it is obviously suboptimal because it maximizes the between-class scatter in the null space of S_w instead of the original input space. Yu *et al.* [15]proposed a direct LDA (DLDA) algorithm, which first removes the null space of S_b. They assume that no discriminative information exists in this space. Unfortunately, it be shown that this assumption is incorrect [13].

2.2. Methods Aimed at Limitations of S_b

When the class conditional densities are multimodal, the class separability represented by S_b is poor. Especially in the case that each class shares the same mean, it fails to find the discriminant direction because there is no scatter of the class means[4].

Notice the rank of S_b is $c-1$, so the number of extracted features is, at most, $c-1$. However, unless a posteriori probability function are selected, $c-1$ features are suboptimal in Bayes sense, although they are optimal with regard to Fisher criterion [4].

In fact, if classification is the ultimate goal, we need only estimate the class density well near the decision boundary[6].

Fukunaga and Mantock [5] presented a nonparametric discriminant analysis (NDA) in an attempt to overcome these limitations presented in LDA. In nonparametric discriminant analysis the between-class scatter S_b is of nonparametric nature. This scatter matrix is generally full rank, thus loosening the bound on extracted feature dimensionality. Also, the nonparametric structure of this matrix inherently leads to the extracted features that preserve relevant structures for classification. Bressan *et al.* [2] explored the nexus between nonparametric discriminant analysis (NDA) and the nearest neighbors (NN) classifier and gave a slight modification of NDA which extends the two-class NDA to a multi-class version.

Although these nonparametric methods overcomes the limitations of S_b, they still depend on the singularity of S_w. The rank of S_w must be no more than $N-c$.

3. Stepwise Nonparametric Margin Maximum Criterion

In this section, we propose a new feature extraction method, stepwise nonparametric margin maximum criterion(SNMMC). SNMMC also uses nonparametric between-class and within-class scatter matrix and does not depend on singularity of within-class scatter matrix. We first propose a nonparametric margin maximum criterion, then a stepwise dimensionality reduction process is presented.

3.1. Nonparametric Margin Maximum Criterion

Our objective is to find a linear transform matrix to make the samples in the same class as compact as possible and the samples belong to the different classes as dispersed as possible.

Assuming a multi-class problem with classes $\omega_i(i = 1, \dots, c)$, we define the extra-class nearest neighbor of a sample $x \in \omega_i$ as

$$x^E = \{x' \notin \omega_i | \, ||x' - x|| \leq ||z - x||, \forall z \notin \omega_i\}. \tag{4}$$

In the same fashion, the intra-class furthest neighbor of the sample $x \in \omega_i$ is defined as

$$x^I = \{x' \in \omega_i | \, ||x' - x|| \geq ||z - x||, \forall z \in \omega_i\}. \tag{5}$$

The the nonparametric extra-class and intra-class differences are defined as

$$\begin{aligned} \Delta^E &= x - x^E, & (6) \\ \Delta^I &= x - x^I. & (7) \end{aligned}$$

. The nonparametric between-class and within-class scatter matrix are defined as

$$\begin{aligned} \hat{S}_b &= \sum_{n=1}^{N} w_n (\Delta_n^E)(\Delta_n^E)^T, & (8) \\ \hat{S}_w &= \sum_{n=1}^{N} w_n (\Delta_n^I)(\Delta_n^I)^T, & (9) \end{aligned}$$

where w_n is the sample weight defined as

$$w_n = \frac{||\Delta_n^I||^\alpha}{||\Delta_n^I||^\alpha + ||\Delta_n^E||^\alpha}, \tag{10}$$

where α is a control parameter between zero and infinity. This sample weight is introduced to deemphasize the samples in the class center and give emphases to the samples near to the other class. The sample that has a larger ratio between the nonparametric extra-class and intra-class differences is given an undesirable influence on the scatter matrix. The weight of sample in Eq.(10) takes a larger value near the classification boundaries and decreases towards zero as we move it to class center. The control parameter α adjusts how fast this happens. α can be chosen by cross-validation.

From the Eq.(6) and (7), we can see that $||\Delta_n^E||$ represents the distance between the sample x_n and its nearest neighbor in the different classes, and $||\Delta_n^I||$ represents the distance between the sample x_n and its furthest neighbor in the same class.

Given a training sample x_n, we define its nonparametric margin as

$$\Theta_n = ||\Delta_n^E||^2 - ||\Delta_n^I||^2, \tag{11}$$

where Δ^E and Δ^I are nonparametric extra-class and intra-class differences and defined in Eq.(6) and (7).

For the sample x_n, the accuracy of the nearest neighbor classification can be evaluated by examining the nonparametric margin. If the nonparametric margin Θ_n is more than zero, x_n will be correctly classified definitely. Otherwise, x_n will be potentially classified to the false class. The larger the nonparametric margin Θ_n is, the more accurately the sample x_n is classified.

Assuming that we extract features by the $D \times d$ linear projection matrix W, the projected sample $x^{new} = W^T x$. The projected nonparametric extra-class and intra-class differences can be written as $\delta^E = W^T \Delta^E$ and $\delta^I = W^T \Delta^I$. So we expect to find the optimal W to make the nonparametric margin $||\delta_n^E||^2 - ||\delta_n^I||^2$ in the projected subspace as large as possible ($W^T W = I$).

$$\widehat{W} = \arg\max_W \sum_{n=1}^{N} w_n (||\delta_n^E||^2 - ||\delta_n^I||^2). \tag{12}$$

This optimization problem can be interpreted as: find the linear transform that maximizes the distance between classes, while minimizing the maximum intra-class distance among the samples of a single class.

Considering that,

$$\begin{aligned} & \sum_{n=1}^{N} w_n (||\delta_n^E||^2 - ||\delta_n^I||^2) \\ = & \sum_{n=1}^{N} w_n (W^T \Delta_n^E)^T (W^T \Delta_n^E) - \sum_{n=1}^{N} w_n (W^T \Delta_n^I)^T (W^T \Delta_n^I) \\ = & \; tr(\sum_{n=1}^{N} w_n (W^T \Delta_n^E)(W^T \Delta_n^E)^T) \\ & -tr(\sum_{n=1}^{N} w_n (W^T \Delta_n^I)(W^T \Delta_n^I)^T) \\ = & \; tr(W^T (\sum_{n=1}^{N} w_n \Delta_n^E (\Delta_n^E)^T) W) \\ & -tr(W^T (\sum_{n=1}^{N} w_n \Delta_n^I (\Delta_n^I)^T) W) \\ = & \; tr(W^T \hat{S}_b W) - tr(W^T \hat{S}_w W) \\ = & \; tr(W^T (\hat{S}_b - \hat{S}_w) W), \end{aligned} \tag{13}$$

where $tr(\cdot)$ is the trace of matrix, $\hat{S}_b$ and $\hat{S}_w$ are the nonparametric between-class and within-class scatter matrix, as defined in Eq.(8) and (9).

So Eq.(12) is equivalent to

$$\widehat{W} = \arg\max_W tr(W^T (\hat{S}_b - \hat{S}_w) W). \tag{14}$$

We call Eq.(14) the nonparametric margin maximum criterion(NMMC).

A similar work to us is maximum margin criterion (MMC) proposed in [7]. But the intrinsic ideas between the two algorithm is somewhat different. Moreover, MMC lacks for the reasonable interpretations.

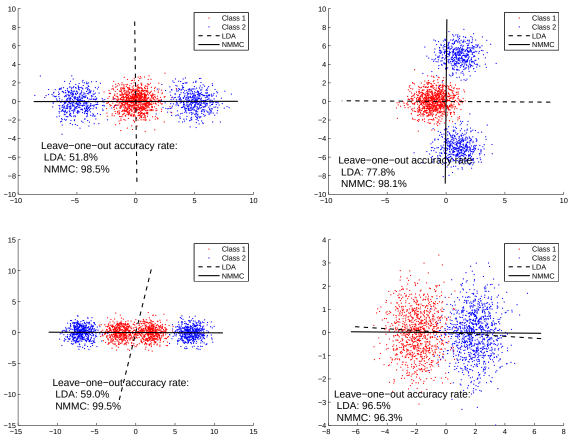

Figure 1: First projected directions of NNDA (solid) and LDA (dashed) projections, for four artificial datasets. Leave-one-out accuracy rates of NN classification are also given.

The projection matrix $\widehat{W}$ must be constituted by the d eigenvectors of $(\hat{S}_b - \hat{S}_w)$ corresponding to its first d largest eigenvalues.

Figure 1 gives comparisons between NMMC and Fisher LDA.

3.2. Stepwise Dimensionality Reduction

However, there is a potential risk in the nonparametric margin maximum criterion. Our objective is to make the samples in the same class as compact as possible and the samples belong to the different classes as dispersed as possible in the projection subspace. However, in the analysis of the nonparametric margin maximum criterion, notice that we calculate nonparametric extra-class and intra-class differences (Δ^E and Δ^I) in original high dimensional space, then project them to the low dimensional space ($\delta^E = W^T\Delta^E$ and $\delta^I = W^T\Delta^I$), which does not exactly agree with the nonparametric extra-class and intra-class differences in projected subspace except for the orthonormal projection case, so we have no warranty on distance preservation. A solution for this problem is to find the projection matrix $\widehat{W}$ by stepwise dimensionality reduction method. In each step, we recalculate the nonparametric extra-class and intra-class differences in its current dimensionality. Thus, we keep the consistency of the nonparametric extra-class and intra-class differences in the process of stepwise dimensionality reduction.

Figure 2 gives the algorithm of stepwise nonparametric margin maximum criterion.

- Give D dimensional samples $\{x_1, \cdots, x_N\}$, we expect to find d dimensional discriminant subspace.
- Suppose that we find the projection matrix $\widehat{W}$ via T steps, we reduce the dimensionality of samples to d_t in step t, and d_t meet the conditions: $d_{t-1} > d_t > d_{t+1}$, $d_0 = D$ and $d_T = d$.
- For $t = 1, \cdots, T$
  1. Calculate the nonparametric between-class $\hat{S}_b^t$ and within-class scatter matrix $\hat{S}_w^t$ in the current d_{t-1} dimensional space;
  2. Calculate the projection matrix $\widehat{W}_t$; $\widehat{W}_t$ is $d_{t-1} \times d_t$ matrix.
  3. Project the samples by the projection matrix $\widehat{W}_t$, $x = \widehat{W}_t^T \times x$.
- The final transform matrix $\widehat{W} = \prod_{t=1}^{T} \widehat{W}_t$.

Figure 2: Stepwise Nonparametric Margin Maximum Criterion

3.3. Discussions

SNMMC has an advantage that there is no need to calculate the inverse matrix, so it is a more efficient and stable method.

However, a drawback of SNMMC is the computational inefficiency in finding the neighbors when the original data space is high dimensionality. A improved method is that PCA is first used to reduce the dimension of data to $N - 1$ (the rank of the total scatter matrix) through removing the null space of the total scatter matrix. Then, SNMMC is performed in the transformed space. Yang *et al.* [14] shows that no discriminant information is lost in this transformed space.

Another drawback of SNMMC is that it's time-consuming in the training procedure due to the stepwise dimensionality reduction process.

However, once the final transform matrix $\widehat{W}$ is found, it's unnecessary to perform the stepwise dimensionality reduction process to unknown test samples. Thus, SNMMC is as efficient as the traditional LDA methods in the test phase.

4. Experiments

In this section, we apply our method to face recognition and compare it with the existing variant LDA methods and the other state-of-art face recognition approaches, such as PCA [12], PCA+LDA [1], NLDA [3], MMC [7] and Bayesian [9] approaches. All the experiments are repeated 5 times independently and the average results are calculated. The classifier is nearest neighbor classifier. The stepwise dimensionality reduction process of SNMMC is performed by 20 steps with the same interval in our experiments.

4.1. Datasets

To evaluate the robustness of SNMMC, we perform the experiments on three datasets from the popular ATT face database [11] and FERET face database [10]. The descriptions of the three datasets are below:

ATT Dataset This dataset is the ATT face database (formerly 'The ORL Database of Faces'), which contains 400 images (112×92) of 40 persons, 10 images per person. Each image is linearly stretched to the full range of pixel values of [0,255]. The set of the 10 images for each person is randomly partitioned into a training subset of 5 images and a test set of the other 5. The training set is then used to learn basis components, and the test set for evaluate.

FERET Dataset 1 This dataset is a subset of the FERET database with 194 subjects only. Each subject has 3 images: (a) one taken under controlled lighting condition with a neutral expression; (b) one taken under the same lighting condition as above but with different facial expressions (mostly smiling); and (c) one taken under different lighting condition and mostly with a neutral expression. All images are pre-processed using zero-mean-unit-variance operation and manually registered using the eye positions. All the images are normalized by the eye locations and are cropped to the size of 75×65. A mask template is used to remove the background and the hair. Histogram equalization is applied to the face images for photometric normalization. Two images for each person is randomly selected for training and the rest one is used for test.

FERET Dataset 2 This dataset is a different subset of the FERET database. All the 1195 people from the FERET Fa/Fb data set are used in the experiment. There are two face images for each person. This dataset has no overlap between the training set and the galley/probe set according to the FERET protocol [10]. 500 people are randomly selected for training, and the remaining 695 people are used for testing. For each testing people, one face image is in the gallery and the other is for probe. All images are pre-processed by using the same method in FERET Dataset 1.

4.2. Experimental Results

Figure 3 shows the rank-1 recognition rates with the different number of features on the three different datasets. It is shown that SNMMC outperforms the other methods. The other methods have relative poor performances at the same dimensionality of features. Moreover, SNMMC does not suffer from the overfitting.

When dataset contains the changes of lighting condition (such as FERET Dataset 1), SNMMC also has obviously better performance than the others.

Different from ATT dataset and FERET dataset 1, where the class labels involved in training and testing are the same, the FERER dataset 2 has no overlap between the training set and the galley/probe set according to the FERET protocol [10]. The ability of generalization from known subjects in the training set to unknown subjects in the gallery/probe set is needed for each method. Thus, the result on FERET dataset 2 is more convincing to evaluate the robust of each method. We can see that SNMMC also gives the best performance than the other methods on FERET dataset 2.

Figure 4 shows cumulative recognition rates on the three different datasets, which shows the SNMMC has outstanding performance in cumulative recognition rates of face recognition.

A major character, displayed by the experimental results, is that SNMMC always has a stable and high recognition rates on the three different datasets, while the other methods have unstable performances. SNNMC is robust to variations of pose, illumination and expression.

5. Conclusion

In this paper, we proposed a new feature extraction method, stepwise nonparametric margin maximum criterion(SNMMC), which finds the most discriminant directions without assuming the class densities belong to any particular parametric family. It does not depend on the nonsingularity of the within-class scatter matrix either. SNMMC is very efficient, accurate and robust for face recognition with the variations of pose, illumination and expression. In the further works, we will extend SNMMC to non-linear discriminant analysis with the kernel method.

Acknowledgments

The research described in this paper has been supported by NSF of China (69935010).

References

[1] P.N. Belhumeur, J. Hespanda, and D. Kiregeman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] M. Bressan and J. Vitrià. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 24:2743C2749, 2003.

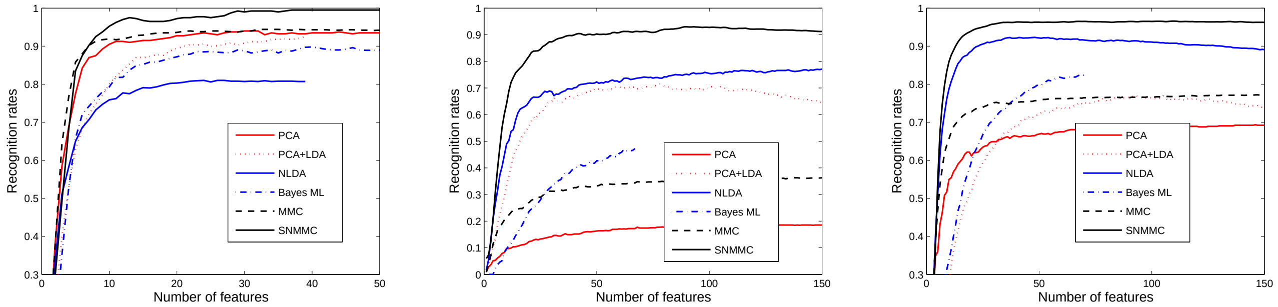

Figure 3: Rank-1 recognition rates with the different number of features on the three different datasets. (Left: ATT dataset; Middle: FERET dataset 1; Right: FERET dataset 2)

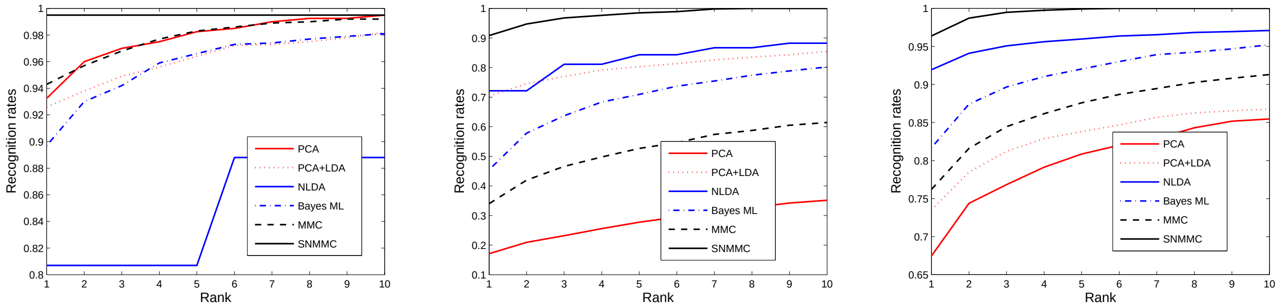

Figure 4: Cumulative recognition rates on the three different datasets. Left:ATT dataset(the number of features is 39; Middle:FERET dataset 1 (the number of features is 60); Right: FERET dataset 2 (the number of features is 60)

[3] 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. *Pattern Recognition*, 33(10):1713–1726, 2000.

[4] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Boston, 2nd edition, 1990.

[5] K. Fukunaga and J. Mantock. Nonparametric discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:671C678, 1983.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[7] H.F. Li, T. Jiang, and K.S. Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Proc. of Neural Information Processing Systems*, 2003.

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

[9] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33:1771–1782, 2000.

[10] P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[11] Ferdinando Samaria and Andy Harter. Parameterisation of a stochastic model for human face identification. In *Proc. of 2nd IEEE Workshop on Applications of Computer Vision*, 1994.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[13] X.G. Wang and X.O. Tang. Dual-space linear discriminant analysis for face recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[14] J. Yang and J.Y. Yang. Why can LDA be performed in PCA transformed space? *Pattern Recognition*, 36:563–566, 2003.

[15] H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.

[16] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.