Short Papers

Asymmetric Principal Component and Discriminant Analyses for Pattern Classification

Xudong Jiang, *Senior Member*, *IEEE*

Abstract—This paper studies the roles of the principal component and discriminant analyses in the pattern classification and explores their problems with the asymmetric classes and/or the unbalanced training data. An asymmetric principal component analysis (APCA) is proposed to remove the unreliable dimensions more effectively than the conventional PCA. Targeted at the two-class problem, an asymmetric discriminant analysis in the APCA subspace is proposed to regularize the eigenvalue that is, in general, a biased estimate of the variance in the corresponding dimension. These efforts facilitate a reliable and discriminative feature extraction for the asymmetric classes and/or the unbalanced training data. The proposed approach is validated in the experiments by comparing it with the related methods. It consistently achieves the highest classification accuracy among all tested methods in the experiments.

Index Terms—Dimension reduction, feature extraction, principal component analysis, discriminant analysis, classification, face detection.

1 Introduction

PRINCIPAL component analysis (PCA) and discriminant analysis (DA) are two fundamental tools of dimension reduction and feature extraction. They are widely applied in computer vision and pattern recognition. Both methods apply eigenvector decomposition on the covariance matrices to decorrelate features and hence to extract the uncorrelated features that are the most significant in some senses. However, the objectives of the two methods are different: one is to maximize the data reconstruction capability of the features and the other is to maximize the discriminatory power of the features. Although some approaches applying PCA for dimension reduction in the areas of face recognition and object detection [1], [2], many researchers turn to DA for feature extraction [3], [4], [5], [6] due to the predominant view that the discrimination of features is the most important for classification. Various variants of the discriminant analysis are summarized in [7], [8].

As the objective of PCA is the best pattern reconstruction that may not be optimal for classification, most approaches apply PCA only aimed at solving the singularity problem of the within-class scatter matrix for the subsequent DA. In fact, the role of PCA can be far beyond solving the singularity problem of the scatter matrix if it is properly applied. This work analyzes the role of PCA in the classification and addresses the problem of applying PCA on the asymmetric classes and/or the unbalanced training data. Classification of two asymmetric classes is a common problem in various verification and object detection tasks, such as biometric person verification, face detection, and other visual object detection. In such tasks, the "positive" class is a single type of object while the "negative" class is "the rest of the world" that contains all the other objects. For such classification tasks, it is extremely difficult to collect a training set that well represents the negative class while a representative training set for the positive class is relatively easy to obtain. As a result, the training data are often unbalanced for the two classes. An asymmetric principal component analysis is proposed in this work to alleviate this problem based on the analysis of the role of PCA in the classification.

Although PCA can improve the classification performance, it is not effective to extract a compact feature set for an efficient (fast) classification. To extract a small number of features, we need discriminant analysis to maximize the discriminatory power of the extracted features [4]. For a two-class problem, only a single feature can be extracted by linear discriminant analysis (LDA), which is far from sufficient for a reasonable classification. One solution is to apply the covariance discriminant analysis (CDA), e.g., the method in [9], [10]. However, the unbalanced training data between the two classes adversely affects the effectiveness of the discriminant evaluation as it is based on the comparison of the class-conditional covariance matrices of the two classes. This work explores this problem and proposes an asymmetric discriminant analysis method that integrates LDA and CDA in a single discriminant evaluation, and regularizes the two covariance matrices. It alleviates problems cause by the imbalance between the training data sets of the two classes. Accordingly, the covariance matrices are also regularized in the classification process.

It is worth noting that some other approaches, such as cascade classification structure and AdaBoost-based algorithms, also tackle the problem of the unbalanced training data. Some of these approaches are successfully applied in the face detection [11], [12]. This paper only addresses the effects of this problem on the PCA and DA, and explores ways to alleviate this problem within the scope of the principal component and discriminant analyses. We also limit our discussion to two-class problem. Some multiclass problems can be converted into two-class problems, as shown in [7], [13].

2 Asymmetric Principal Component Analysis

Given q n-dimensional column vectors for training where the positive class ω_o has q_o samples and the negative class ω_c has q_c samples, $q = q_o + q_c$, compute the class-conditional mean vectors M_o, M_c, and covariance matrices $\mathbf{\Sigma}_o$, $\mathbf{\Sigma}_c$. The covariance matrix of the class mean is computed as

$$\mathbf{\Sigma}_m = \frac{1}{q}[q_o(M_o - M)(M_o - M)^T) + q_c(M_c - M)(M_c - M)^T],$$

where M is the mean over all training samples. It is not difficult to get the covariance matrix of the total training data by

$$\mathbf{\Sigma}_t = \frac{1}{q}(q_o\mathbf{\Sigma}_o + q_c\mathbf{\Sigma}_c) + \mathbf{\Sigma}_m. \quad (1)$$

If the a priori probabilities of the two classes are estimated by $p_o = q_o/q$ and $p_c = q_c/q$, the covariance matrix of the total training data can be expressed as

$$\mathbf{\Sigma}_t = p_o\mathbf{\Sigma}_o + p_c\mathbf{\Sigma}_c + \mathbf{\Sigma}_m. \quad (2)$$

In the literature, $\mathbf{\Sigma}_t$ is often called total scatter matrix, $\mathbf{\Sigma}_m$ is called between-class scatter matrix, and $\mathbf{\Sigma}_w = p_o\mathbf{\Sigma}_o + p_c\mathbf{\Sigma}_c$ is often called within-class scatter matrix.

• *The author is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Link, Singapore 639798. E-mail: exdjiang@ntu.edu.sg.*

Manuscript received 14 Apr. 2008; revised 28 Aug. 2008; accepted 9 Oct. 2008; published online 16 Oct. 2008.
Recommended for acceptance by S. Li.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-04-0215.
Digital Object Identifier no. 10.1109/TPAMI.2008.258.

PCA applies eigen-decomposition on Σ_t, i.e., $\Sigma_t = \Phi\Lambda\Phi^T$, and keeps the m eigenvectors Φ_m, $\Phi_m \in \mathbb{R}^{n\times m}$, corresponding to the m largest eigenvalues. An n-dimensional pattern vector X is transformed to an m-dimensional feature vector $\hat{X}$ by $\hat{X} = \Phi_m^T X$, $m < n$. Since PCA is optimal for the pattern reconstruction but not necessarily optimal for classification, people turn to the DA where PCA is applied only to solve the singularity problem of Σ_w before applying DA.

In fact, the role of PCA in the classification is far beyond the low-dimensional data representation or solving the singularity problem of Σ_w. For a quantitative analysis of the role of PCA in the classification, we model the class-conditional distributions by multivariate Gaussian density functions. The Bayes optimal decision rule detects a positive sample X if

$$(X - M_c)^T \Sigma_c^{-1}(X - M_c) - (X - M_o)^T \Sigma_o^{-1}(X - M_o) > b, \qquad (3)$$

where $b = \ln(|\Sigma_o|/|\Sigma_c|) + 2(\ln p_c - \ln p_o)$. After applying eigen-decomposition, the Bayes decision rule (3) is simplified as

$$\sum_{k=1}^{n} \frac{g_k^2}{\lambda_k^c} - \sum_{k=1}^{n} \frac{h_k^2}{\lambda_k^o} > b, \qquad (4)$$

where g_k is the projection of $(X - M_c)$ on the eigenvector Φ_k^c corresponding to the eigenvalue λ_k^c of Σ_c and h_k is the projection of $(X - M_o)$ on the eigenvector Φ_k^o corresponding to the eigenvalue λ_k^o of Σ_o.

However, a lot of approaches [13], [14], [15], [16], [17], [18] for visual object detection and recognition tasks modify the above optimal decision rule into

$$\sum_{k=1}^{m} \frac{g_k^2}{\lambda_k^c} + \sum_{k=m+1}^{n} \frac{g_k^2}{\rho_c} - \sum_{k=1}^{m} \frac{h_k^2}{\lambda_k^o} - \sum_{k=m+1}^{n} \frac{h_k^2}{\rho_o} > b, \qquad (5)$$

which replaces the $n - m$ smallest eigenvalues of both classes by two constants ρ_c and ρ_o, respectively, and often $m \ll n$. It is worth exploring why the decision rule (5) may outperform (3) or (4).

Eigenvalue λ_k^c or λ_k^o is the variance of the positive or negative training samples projected on the eigenvector Φ_k^c or Φ_k^o. It is an estimate of the class true (ensemble) variance based on the available training data. If the eigenvalues deviate from the ensemble variances, the decision rule (3) or (4) overfits the training samples, and hence, leads to a poor generalization on the novel testing data. This problem will become very severe if some eigenvalues largely deviate from the ensemble variances.

Fig. 1 plots an eigenspectrum (λ_k^o sorted in descending order) obtained from 2,000 face images of size 20×20 and the variances v_k^o of other 2,000 face images projected on the eigenvectors Φ_k^o. Fig. 1 also plots eigenvalues sorted in ascending order λ_{n-k+1}^c obtained from 2,000 nonface images of size 20×20 and the variances v_{n-k+1}^c of other 2,000 nonface images projected on the eigenvectors Φ_{n-k+1}^c. All images are taken from the ECU face detection database [19]. Fig. 1 shows large deviations of the small eigenvalues from the variances of the novel images projected on the eigenvectors. Other sets of face and nonface images produce results similar to Fig. 1.

This problem was well addressed in [8]. Although, in general, the largest sample-based eigenvalues are biased upward and the smallest ones are biased downward, the bias is most pronounced when the population eigenvalues tend toward equality, and it is correspondingly less severe when their values are highly disparate [20]. In most applications, eigenspectrum often first decays very rapidly and then stabilizes. Therefore, the smallest eigenvalues are biased much more severely than the largest ones. This is evidenced by Fig. 1.

Thus, similar to (5) that replaces the smallest eigenvalues by a constant, removing the subspace spanned by the eigenvectors of

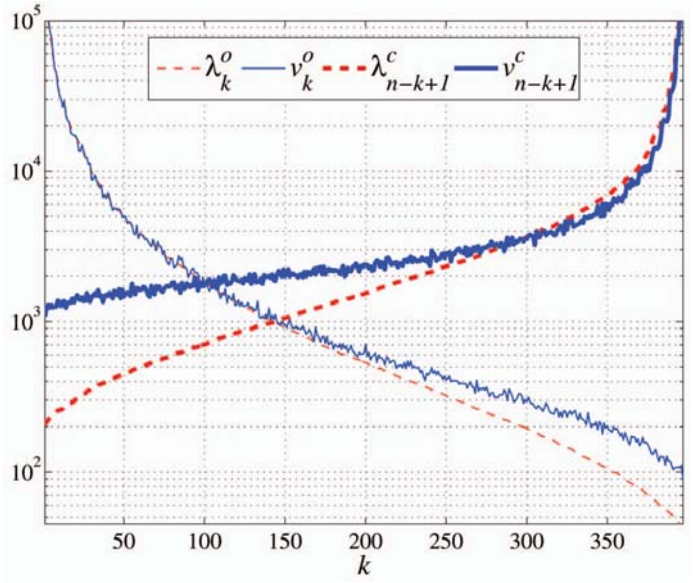

Fig. 1. Eigen-spectra $\lambda_k^o/\lambda_{n-k+1}^c$ computed from 2,000 face/nonface images and variances v_k^o/v_{n-k+1}^c of other 2,000 face/nonface images projected on the eigenvectors Φ_k^o/Φ_k^c.

Σ_o and Σ_c corresponding to the smallest eigenvalues improves the generalization of the classifier, i.e., reduces the classification error on the novel testing data. However, the principal components of Σ_m should not be removed as they contain the discriminative information. Therefore, it is clear that PCA on $\Sigma_t = p_o\Sigma_o + p_c\Sigma_c + \Sigma_m$ plays an important role in the classification. It alleviates the overfitting problem or improves the generalization capability by removing the subspace spanned by eigenvectors of Σ_o and Σ_c corresponding to the small eigenvalues while keeping the principal components of Σ_m. A good example of applying PCA to improve the face identification accuracy can be found in [21].

However, in (1) and (2), Σ_o and Σ_c are weighted by q_o/q and q_c/q or by p_o and p_c. These weights are required for PCA to achieve the least-mean-square reconstruction error. For the classification purpose, however, the objective is to remove the dimensions in which the sample-based class-conditional variances are unreliable. The reliability of a covariance matrix is not dependent on the class prior probability. The Bayes optimal decision rule (3) minimizes the sum of the two errors weighted by p_o and p_c so that the threshold b depends on p_o and p_c. In practice, the threshold b is often determined by some factors other than p_o and p_c to achieve a desired positive or negative error rate. More training samples of a class may result in a more reliable covariance matrix if they are properly collected. However, it is not the more but the less reliable covariance matrix that should be heavily weighted so that more dimensions characterized by the small variances of this class can be removed. It is thus clear that PCA on the total data scatter matrix Σ_t (1) or (2) does not effectively remove the unreliable dimensions because Σ_t is not constructed from the classification point of view.

To tackle this problem, we propose to construct an asymmetric pooled covariance matrix by

$$\Sigma_\alpha = \alpha_o\Sigma_o + \alpha_c\Sigma_c + \Sigma_m, \qquad (6)$$

where α_o and α_c are determined by the reliability of the covariance matrices Σ_o and Σ_c, $\alpha_o + \alpha_c = 1$. Different from (1) and (2), α_o and α_c are unrelated to the class a priori probabilities. The objective of the proposed asymmetric pooled covariance matrix Σ_α is to facilitate an effective removal of the unreliable dimensions. Thus, larger value of α_o or α_c should be assigned to the less reliable covariance matrix so that more dimensions characterized by the

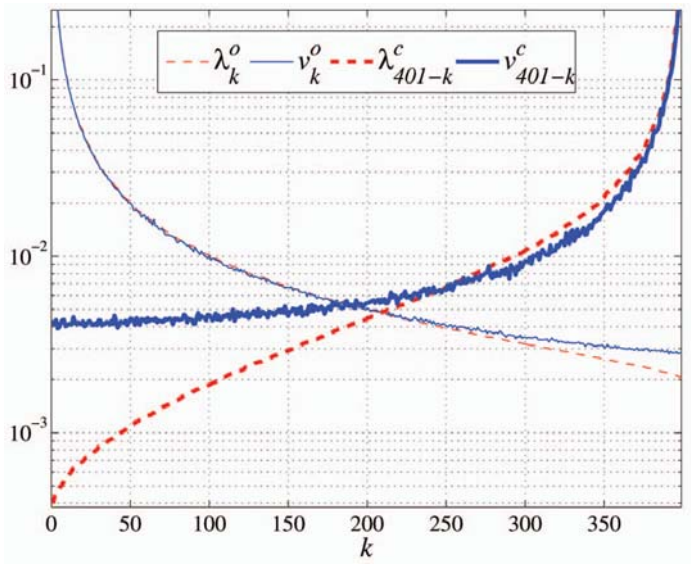

Fig. 2. Eigen-spectra $\lambda_k^o/\lambda_{401-k}^c$ of covariance matrices computed from $8,000/800$ random vectors and the ensemble variances v_k^o/v_{401-k}^c projected on the eigenvectors Φ_k^o/Φ_{401-k}^c.

small variances of the less reliable class can be removed by eigen-decomposition of $\mathbf{\Sigma}_\alpha$.

In the applications of verification and object detection, the positive and negative classes are highly asymmetric because the positive class represents only one particular object while the negative class represents the whole "rest of the world" that contains all other objects. Thus, it is much more difficult to collect a representative training set for the negative class than for the positive class. As a result, the reliability of the two class-conditional covariance matrices greatly differs from each other. Fig. 1 clearly shows much larger biases of small eigenvalues of the nonface class than those of the face class. Even if a huge negative database is collected, samples could be heavily biased to a small subset of the "whole world." Therefore, in general, we need to assign a larger weight α_c to the negative class than to the positive class. However, the optimal value of the weight is application dependent that varies from one training database to another.

If there is no prior knowledge about the class characteristics and the data collection procedure, less training samples in general result in a less reliable covariance matrix. Thus, we suggest constructing the asymmetric pooled covariance matrix in the form of

$$\mathbf{\Sigma}_\alpha = \frac{1}{q}(q_c\mathbf{\Sigma}_o + q_o\mathbf{\Sigma}_c) + \mathbf{\Sigma}_m. \tag{7}$$

In sharp contrast to the scatter matrix $\mathbf{\Sigma}_t$ (1) that weights the covariance matrices proportionally to the number of training samples, the proposed $\mathbf{\Sigma}_\alpha$ (7) pools them with weights *inversely* proportional to the number of training samples.

Fig. 2 plots the eigenspectrum λ_k^o of a covariance matrix computed from 8,000 400-dimensional samples generated from 400 independent Gaussian random variables. The ensemble variance of the i^{th} random variable is $1/i$. The ensemble variances projected on the eigenvectors Φ_k^o are plotted in Fig. 2 (denoted by v_k^o) to show the eigenvalue bias. Fig. 2 also plots another set of eigenvalues sorted in ascending order λ_{401-k}^c and the projected ensemble variances v_{401-k}^c from the same random vector but using 800 samples only. Fig. 2 clearly shows that smaller number of training samples results in larger biases of the eigenvalues. However, PCA on $\mathbf{\Sigma}_t$ (1) removes dimensions specified mainly by the small λ_k^o as the weight of $\mathbf{\Sigma}_o$ is 10 times larger than that of $\mathbf{\Sigma}_c$. This is undesirable for classification. Obviously, PCA on the proposed (7) that puts heavier weight on $\mathbf{\Sigma}_c$ rather than $\mathbf{\Sigma}_o$ removes the unreliable dimensions more effectively.

The proposed asymmetric principal component analysis (APCA) applies eigen-decomposition on $\mathbf{\Sigma}_\alpha$ (6), i.e.,

$$\alpha_o\mathbf{\Sigma}_o + \alpha_c\mathbf{\Sigma}_c + \mathbf{\Sigma}_m = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T, \tag{8}$$

and extracts the m eigenvectors $\hat{\mathbf{\Phi}}$ from $\mathbf{\Phi}$ corresponding to the m largest eigenvalues in $\mathbf{\Lambda}$. Its purpose is neither to have a low-dimensional data representation with least-mean-square reconstruction error, nor to extract a compact feature set for fast classification. The objective of the proposed APCA is to remove the unreliable dimensions to alleviate the overfitting problem and hence to achieve better classification generalization.

3 Asymmetric Discriminant Analysis

The objective of APCA is to alleviate overfitting problem. For a fast classification, DA is necessary to extract a compact feature set from the reliable APCA subspace. The Bhattacharyya distance measures the separability between two classes. For Gaussian distribution, its analytical form in the APCA subspace is given by

$$\begin{aligned} D = &\frac{1}{8}(\hat{M}_o - \hat{M}_c)^T\left(\frac{\hat{\mathbf{\Sigma}}_o + \hat{\mathbf{\Sigma}}_c}{2}\right)^{-1}(\hat{M}_o - \hat{M}_c) \\ &+ \frac{1}{2}\ln\frac{|(\hat{\mathbf{\Sigma}}_o + \hat{\mathbf{\Sigma}}_c)/2|}{\sqrt{|\hat{\mathbf{\Sigma}}_o||\hat{\mathbf{\Sigma}}_c|}}, \end{aligned} \tag{9}$$

where $\hat{M}_o = \hat{\mathbf{\Phi}}^T M_o$, $\hat{M}_c = \hat{\mathbf{\Phi}}^T M_c$, $\hat{\mathbf{\Sigma}}_o = \hat{\mathbf{\Phi}}^T\mathbf{\Sigma}_o\hat{\mathbf{\Phi}}$, and $\hat{\mathbf{\Sigma}}_c = \hat{\mathbf{\Phi}}^T\mathbf{\Sigma}_c\hat{\mathbf{\Phi}}$. Let $\hat{\mathbf{\Sigma}}_m = \hat{\mathbf{\Phi}}^T\mathbf{\Sigma}_m\hat{\mathbf{\Phi}}$. The first term of the Bhattacharyya distance is maximized by LDA

$$\hat{\mathbf{\Sigma}}_m\mathbf{\Phi} = (\hat{\mathbf{\Sigma}}_o + \hat{\mathbf{\Sigma}}_c)\mathbf{\Phi}\mathbf{\Lambda}. \tag{10}$$

It is proven in [22] that the second term of the Bhattacharyya distance is maximized in the subspace spanned by the generalized eigenvectors corresponding to the largest $\lambda_k + 1/\lambda_k$, where λ_k is the generalized eigenvalue of matrix pair either $(\hat{\mathbf{\Sigma}}_o, \hat{\mathbf{\Sigma}}_c)$ or $(\hat{\mathbf{\Sigma}}_c, \hat{\mathbf{\Sigma}}_o)$. We call it CDA. Obviously, λ_k is the ratio between the two class-conditional variances projected on the eigenvector, v_k^o/v_k^c or v_k^c/v_k^o. Thus, the largest $\lambda_k + 1/\lambda_k$ maximizes the sum of the two ratios, $v_k^o/v_k^c + v_k^c/v_k^o$.

There are three problems associated with this optimization process. First, the generalized eigenvectors of $(\hat{\mathbf{\Sigma}}_o, \hat{\mathbf{\Sigma}}_c)$ are different from those of $(\hat{\mathbf{\Sigma}}_c, \hat{\mathbf{\Sigma}}_o)$. Which pair should we choose? The second problem is that $\hat{\mathbf{\Sigma}}_o$ and $\hat{\mathbf{\Sigma}}_c$ are still biased although the problem is alleviated after removing dimensions of highly biased small eigenvalues by APCA. This adversely affects the eigenvector selection. The last problem is the separate maximization of the two terms.

To tackle the first problem, we propose to solve the generalized eigenvalue problem $\hat{\mathbf{\Sigma}}_o\mathbf{\Phi} = (\hat{\mathbf{\Sigma}}_o + \hat{\mathbf{\Sigma}}_c)\mathbf{\Phi}\mathbf{\Lambda}$ or $\hat{\mathbf{\Sigma}}_c\mathbf{\Phi} = (\hat{\mathbf{\Sigma}}_o + \hat{\mathbf{\Sigma}}_c)\mathbf{\Phi}\mathbf{\Lambda}$ instead of $\hat{\mathbf{\Sigma}}_o\mathbf{\Phi} = \hat{\mathbf{\Sigma}}_c\mathbf{\Phi}\mathbf{\Lambda}$ or $\hat{\mathbf{\Sigma}}_c\mathbf{\Phi} = \hat{\mathbf{\Sigma}}_o\mathbf{\Phi}\mathbf{\Lambda}$. All of them find dimensions that maximize or minimize the ratio between the projected variances of the two classes. However, it is better to have the eigenvectors of the former two because they are orthogonal with respect to the pooled covariance matrix while the eigenvectors of the latter two are orthogonal with respect to one of the two covariance matrices, respectively. It is easy to prove that there is no difference of applying $\hat{\mathbf{\Sigma}}_o\mathbf{\Phi} = (\hat{\mathbf{\Sigma}}_o + \hat{\mathbf{\Sigma}}_c)\mathbf{\Phi}\mathbf{\Lambda}$ from applying $\hat{\mathbf{\Sigma}}_c\mathbf{\Phi} = (\hat{\mathbf{\Sigma}}_o + \hat{\mathbf{\Sigma}}_c)\mathbf{\Phi}\mathbf{\Lambda}$. Let λ_k denote the generalized eigenvalues of one of them. The sum of the ratios between the two class-conditional variances projected on the eigenvector is $v_k^o/v_k^c + v_k^c/v_k^o = \lambda_k/(1-\lambda_k) + (1-\lambda_k)/\lambda_k$. It is not difficult to prove that this is maximized by $\max_k\{\max(\lambda_k, 1-\lambda_k)\}$.

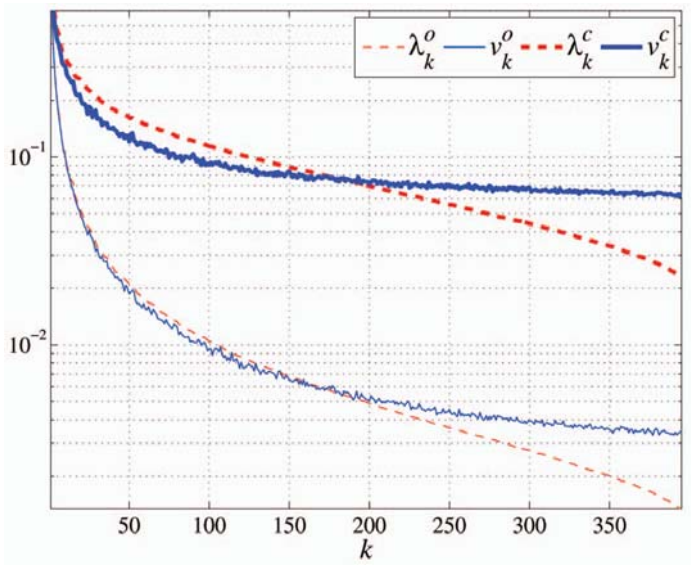

Fig. 3. Eigen-spectra λ_k^o, λ_k^c with different decaying rates and the true variances v_k^o, v_k^c projected on the eigenvectors Φ_k^o, Φ_k^c.

The second problem is difficult to solve. The proposed APCA removes unreliable dimensions caused by small eigenvalues. The remaining large eigenvalues are in general biased upward. Larger amount of bias will be produced for the class that is worse represented by their training data. This is evidenced by Figs. 1 and 2. Furthermore, the bias is more pronounced when eigenvalues tend toward equality and less severe when their values are highly disparate. This was pointed out in [20] and is further evidenced in Fig. 3. Two thousand samples for each of the two classes are generated by a 400-dimensional random vector that obeys multivariate Gaussian distribution. The ensemble variances of the positive and negative classes decay, respectively, by $1/i$ and $\sqrt{1/i}$, where i is the index of dimension. Fig. 3 shows the eigenspectra of the two class-conditional covariance matrices and the ensemble variances projected on the corresponding eigenvectors. It clearly shows that the different upward bias amounts between the positive and negative classes.

In general, the negative class occupies a larger subspace, and hence, has flatter eigenspectrum than the positive class. It is also more difficult to collect a representative training set for the negative class than for the positive class. As a result, eigenvalues of negative class are biased higher in the APCA principal subspace than those of the positive class, which adversely affects the discriminant evaluation. To tackle this problem, we propose to apply CDA to the regularized covariance matrices in the APCA subspace as

$$\hat{\mathbf{\Sigma}}_o \mathbf{\Phi} = (\hat{\mathbf{\Sigma}}_o + \beta \hat{\mathbf{\Sigma}}_c) \mathbf{\Phi} \mathbf{\Lambda}. \tag{11}$$

The parameter β, $0.5 < \beta \leq 1$, can be determined by some prior knowledge about the asymmetry of the two classes. In general, we have $\beta < 1$ because the negative class occupies larger subspace, and it is more difficult to have a representative training data set than the positive class. If the discriminant evaluation (11) is applied in a small APCA principal subspace, e.g., $m \ll n/2$, all eigenvalues of the negative class will most likely have higher upward biases than those of the positive class. In this case, $\beta < 1$ should be applied. If (11) is applied in a large principal subspace, e.g., $m \gg n/2$, there are both upward- and downward-biased eigenvalues. In this case, we should let $\beta = 1$. The optimal value of β is application dependent that varies from one training task to another.

The third problem can be solved by integrating the two optimization processes (10) and (11). It is now possible because both solve the eigen-decomposition normalized by the same pooled covariance matrix. The proposed asymmetric discriminant analysis (ADA) is to solve the following generalized eigen-decomposition problem:

$$(\hat{\mathbf{\Sigma}}_o + \gamma \hat{\mathbf{\Sigma}}_m) \mathbf{\Phi} = (\hat{\mathbf{\Sigma}}_o + \beta \hat{\mathbf{\Sigma}}_c) \mathbf{\Phi} \mathbf{\Lambda} \tag{12}$$

in the APCA subspace, where γ is a constant that weights the discriminatory information about class mean against that about covariance. If a sufficiently large value is assigned to γ, the eigenvector corresponding to the largest eigenvalue approaches to the solution of LDA (10) and the remaining eigenvectors approach to the solution of (11).

The proposed asymmetric principal and discriminant analysis (APCDA) algorithm applies ADA (12) in the APCA subspace and extracts d eigenvectors $\tilde{\mathbf{\Phi}}$ from $\mathbf{\Phi}$ corresponding to the d largest $\max(\lambda_k, 1 - \lambda_k)$, where λ_k are the generalized eigenvalues of (12). Its purpose is to extract d discriminating features from the reliable APCA subspace. The ADA part of the APCDA algorithm differs from other approaches in three aspects: it extracts discriminant features about class mean and covariance in one eigen-decomposition; the extracted features are orthogonal with respect to the pooled covariance matrix; and the class-conditional covariance matrix is regularized.

For Gaussian distribution, the Bayes optimal classifier is simplified to a minimum Mahalanobis distance classifier. Similar to the ADA algorithm, we propose to regularize the covariance matrix in the feature space. The minimum Mahalanobis distance classifier is thus modified as

$$(\tilde{X} - \tilde{M}_c)^T (\beta \tilde{\mathbf{\Sigma}}_c)^{-1} (\tilde{X} - \tilde{M}_c) - (\tilde{X} - \tilde{M}_o)^T \tilde{\mathbf{\Sigma}}_o^{-1} (\tilde{X} - \tilde{M}_o) > b, \tag{13}$$

which is called minimum asymmetric Mahalanobis distance classifier. All variables with tilde sign in (13) are the corresponding variables in (3) projected in the APCA or APCDA feature space. An n-dimensional pattern vector X is transformed to a d-dimensional feature vector $\tilde{X}$ by $\tilde{X} = \mathbf{U}^T X$, where $\mathbf{U} = \hat{\mathbf{\Phi}} \tilde{\mathbf{\Phi}}$, $\hat{\mathbf{\Phi}} \in \mathbb{R}^{n \times m}$, $\tilde{\mathbf{\Phi}} \in \mathbb{R}^{m \times d}$, and, hence, $\mathbf{U} \in \mathbb{R}^{n \times d}$. The constant b is a decision threshold determined by the compromise between the positive and negative error rates in a specific application.

PCA and LDA only use the second-order statistics of the training data. As the Gaussian distribution is widely applied as a probability distribution fully specified by the second-order statistics, the quantitative analysis in this work is based on the Gaussian assumption. Nevertheless, it can also be applied for some other distributions. For example, for classification, Mahalanobis distance is only optimal under Gaussian assumption and LDA is only optimal for Gaussian distribution with the same covariance matrices of all classes. However, they are widely employed in various applications. This work may not be suitable for some classifiers where the data variance is not applied, such as the nearest neighbor classifiers. For multimodal distribution, we can decompose it into several single-modal distributions using Gaussian Mixture Model or clustering technique, and then apply the proposed approach on each single-modal distribution.

4 Experiments

The proposed APCDA algorithm is related to the statistical feature extraction methods PCA, LDA, CDA, and the Bayes discriminating feature (BDF) [14], [17], [18]. We shall carry out three groups of experiments to validate the analyses in the previous sections and the feasibility of the proposed methods. The minimum Mahalanobis distance classifier is applied. We first compare APCA with PCA and then compare APCDA with PCA+LDA+CDA (called PLCDA) using synthetic data. Finally, we will compare the APCA and APCDA algorithms with PCA, PLCDA, and BDF using a real

TABLE 1
Mean μ and Standard Deviation σ of MTER for Different m

m	300	280	260	240	220	200	180	160
PCA μ	20.1	20.0	19.9	19.7	19.4	19.1	18.7	18.4
	20.0	19.9	19.8	19.6	19.2	18.8	18.4	18.0
APCA μ	14.9	12.0	9.94	8.85	8.52	8.68	9.34	10.3
	14.4	11.3	9.04	7.78	7.34	7.37	7.97	9.12
PCA σ	.007	.024	.054	.074	.131	.126	.164	.186
	.010	.026	.037	.050	.098	.159	.168	.165
APCA σ	.248	.273	.259	.204	.157	.145	.153	.166
	.201	.107	.154	.113	.113	.125	.114	.115

image database. The parameters $\gamma = 10$, $\alpha_c = 0.8$, and $\alpha_o = 0.2$ are kept unchanged in all experiments to test the proposed approach with significantly asymmetric parameters, $\alpha_c/\alpha_o = 4$. For the synthetic data, we set $\beta = 0.95$ because ADA is applied in a large subspace $m > n/2$. For the real image database, we set $\beta = 0.75$ because ADA is applied in a small subspace $m \ll n/2$.

4.1 Results of APCA on Synthetic Data

The positive data of class ω_o are drawn from a 400-dimensional random vector Ω_o that has Gaussian distribution with zero mean and covariance matrix of $diag\{1, 1/2^{0.5}, \ldots, 1/400^{0.5}\}$. The negative data of class ω_c are drawn from another Gaussian random vector Ω_c that has covariance matrix of $1/50^{0.25} diag\{1, 1/2^{0.25}, \ldots, 1/400^{0.25}\}$. This leads to a same variance of the two classes at the 50th dimension. The means of Ω_c are $1/50^{0.25}$ in the 50th dimension and zero in the other dimensions. With the assumption of $p_o = 0.8$ and $p_c = 0.2$, 2,000 and 500 training samples are generated from random vectors Ω_o and Ω_c, respectively. Similarly, 20,000 and 5,000 novel samples for ω_o and ω_c are, respectively, generated for testing. The minimum total error rate (MTER) over different decision thresholds b is used to measure the classification performance. Note that the Bayes rule minimizes the total error rate. Ten runs of experiments with independent training and testing sets are performed, and for each run, $m = 400, 380, 360, \ldots, 100$ are tested. With the decrease of m, the MTERs of both PCA and APCA monotonically decrease to their minimums and then increase. Table 1 records the mean and the standard deviation of the MTER over the 10 runs. The above experiments are repeated with the uniform distribution of Ω_o and Ω_c. Results are recorded in the corresponding second rows in Table 1.

Table 1 shows that the proposed APCA consistently outperforms PCA for all numbers of features m. The Student's t-test is used to compare the means of MTER of PCA and APCA. The statistic significance or confidence level is higher than 99.95 percent for all results in Table 1. Thus, the probability of the APCA that does not outperform PCA is less than 0.0005 based on the Student's t-test. To further show the consistency of the APCA in the accuracy improvement, Fig. 4 plots the ROC curves for $m = 200$. It shows that APCA consistently outperforms PCA for all different decision thresholds b.

In this training task, as the positive class is much better represented by 2,000 training samples than the negative class that is represented by only 500 samples, eigenvalues of the data total scatter matrix are not good indicators of the reliability of the corresponding dimensions. Thus, it is not a surprise that the proposed APCA significantly and consistently outperforms PCA, although PCA minimizes the data reconstruction error but APCA does not.

4.2 Results of APCDA on Synthetic Data

The positive data of class ω_o are drawn from a 200-dimensional Gaussian random vector Ω_o with zero mean and covariance matrix of $diag\{1, 1/2, \ldots, 1/200\}$. The negative data of class ω_c are drawn from another Gaussian random vector Ω_c that has covariance matrix of $1/20^{0.5}\ diag\{1, 1/2^{0.5}, \ldots, 1/200^{0.5}\}$. The means of Ω_c are $1/20^{0.5}$ in the 20th dimension and zero in the other dimensions. A total of 210 samples and 10,000 novel samples for each class are generated, respectively, for training and testing. Data dimensionality is reduced by PCA or APCA, and PLCDA and APCDA are applied to extract d discriminative features with $m = 120$. Fig. 5 shows the mean of the MTER over 10 runs of experiments with independent training and testing sets.

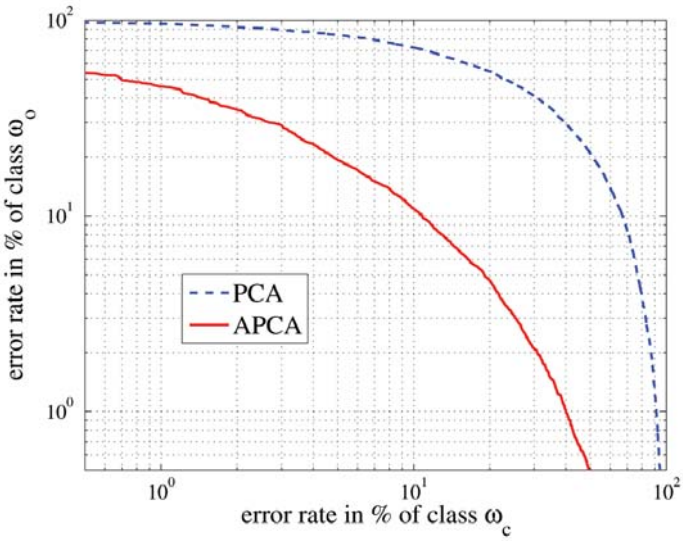

Fig. 4. ROC curves of PCA and APCA with $m = 200$.

Although the two classes have the same number of training samples, Fig. 5 shows that APCA outperforms PCA for all numbers of features as the "larger" class has flatter eigenspectrum, and hence, larger eigenvalue bias. However, if the dimension is overreduced, APCA may not outperform PAC, as shown in Fig. 5. Note that although the covariance matrices are of full rank, the discriminant analyses LDA+CDA and ADA perform very badly if applying them directly on the original 200-dimensional space. However, after applying PCA or APCA to remove the unreliable dimensions, the discriminant analysis LDA+CDA or ADA outperforms PCA or APCA. Fig. 5 shows that the overdimension reduction by PCA or APCA results in sharp increase of the classification error while the discriminant

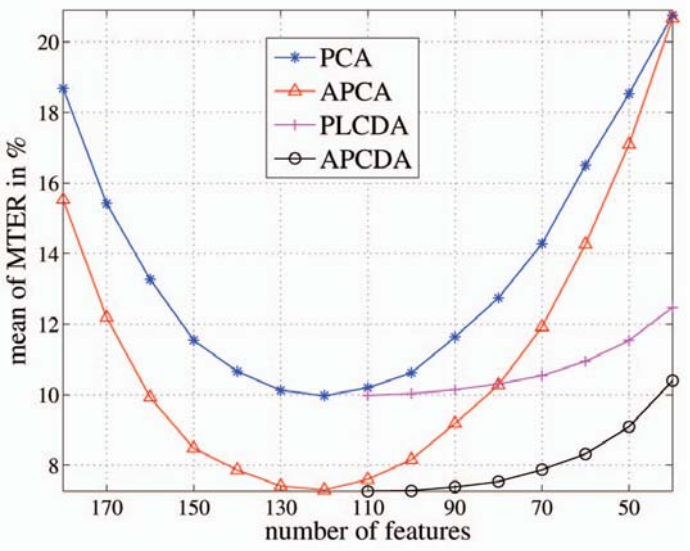

Fig. 5. Mean of the MTER against the number of features.

TABLE 2
Mean μ and Standard Deviation σ of MTER for Different m or d

m or d	110	100	90	80	70	60	50
PCA μ	10.2	10.6	11.6	12.7	14.3	16.5	18.5
APCA μ	7.60	8.16	9.18	10.3	11.9	14.3	17.1
PLCDA μ	9.99	10.0	10.1	10.3	10.5	10.9	11.5
APCDA μ	7.26	7.27	7.38	7.53	7.87	8.32	9.09
PCA σ	.575	.444	.371	.546	.409	.534	.539
APCA σ	.362	.411	.546	.578	.521	.555	.687
PLCDA σ	.605	.611	.557	.596	.584	.603	.726
APCDA σ	.452	.468	.468	.462	.464	.464	.469

methods LDA+CDA and ADA are much less sensitive to the number of features. Among all tested methods, the best classification results are achieved by the proposed APCDA (APCA+ADA) consistently for all number of features. Table 2 numerically records the mean and the standard deviation of the MTER over the 10 runs of experiments. The statistic significance comparing APCDA and PLCDA is higher than 99.95 percent for all results in Table 1 based on the Student's t-test. Thus, the probability of the APCDA that does not outperform PLCDA is less than 0.0005.

4.3 Results on Real Image Data of Face Detection

The database is taken from the ECU face and skin detection database created by Edith Cowan University [19]. It contains 9,339 face images with various lighting, pose, and expression variations, and 8,951 nonface images randomly extracted from natural images and normalized into the size of 20×20. Fig. 6 shows 20 face and nonface images in the database.

Four sets of experiments are carried out, each of which has distinct testing set. In the database partition i, $i = 1, 2, 3, 4$, the ith 25 percent of face and nonface images in the database are picked out for testing and the remaining 75 percent of images serve as the training data. Thus, all 18,290 images have served as testing samples once and only once in the four sets of experiments. We first apply PCA and APCA to reduce the dimension from 400 to $m = 390, 380, \ldots, 30, 20$. For all the four database partitions, the equal error rates (EERs) of both PCA and APCA monotonically decrease to their minimums at the dimensionality around 90 and then increase. Thus, PLCDA and APCDA are applied to extract d features with m fixed at 90. Fig. 7 plots the average EER over the four database partitions against the number of features m (for PCA and APCA) or d (for PLCDA and APCDA). It shows that APCA outperforms PCA consistently, but has problem if the dimensionality is overreduced. The proposed APCDA achieves the best detection results consistently for all number of features tested.

To have a full picture of the face detection performances at various decision thresholds, Fig. 8 plots the average ROC curves over the four database partitions with $d = 50$. The number of the principal components of the BDF approach is $M = 10$, as suggested in [18]. (Indeed, BDF has much better results with $M = 10$ than that with $M = 50$.) From Fig. 8, we see that BDF performs worse than PCA at most operating thresholds. This suggests that, for this detection task, removing the unreliable dimensions is a better way than keeping them and scaling them by the average eigenvalues over these dimensions. It is a surprise that the discriminant approach PLCDA is not always better than PCA alone. For some operating points, APCA does not outperform PCA because the dimension is overreduced (from 400 to 50). Although the relative detection performances of the five approaches vary at different operating thresholds, the proposed APCDA approach consistently outperforms the other methods at all operating thresholds.

Fig. 6. Sample images taken from the ECU database.

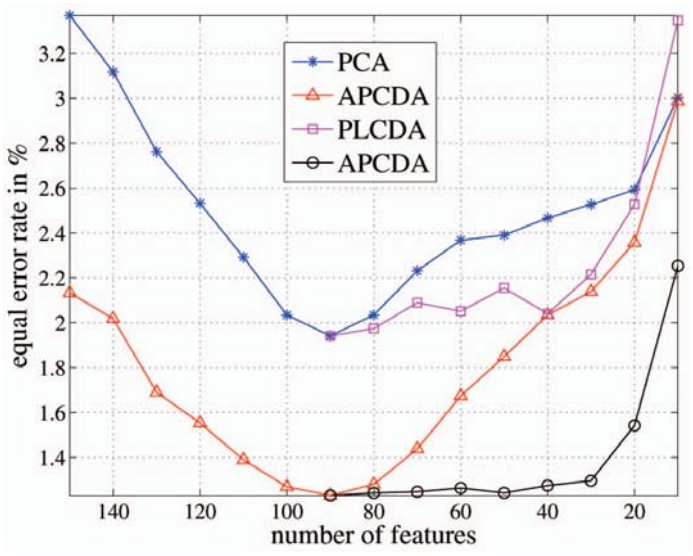

Fig. 7. Average equal error rate over the four database partitions against the number of features.

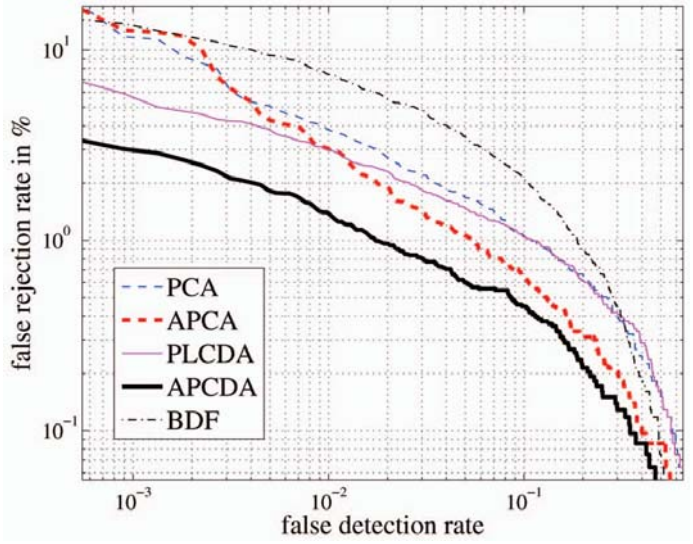

Fig. 8. Average ROC curves over the four database partitions with $m = 90$, $d = 50$ for PLCDA and APCDA, $m = 50$ for PCA and APCA, and $M = 10$ for BDF.

TABLE 3
False Face Rejection Rate in Percent at Small Number of False Detections (#FD) out of 8,951 Nonface Images

#FD	0	8	16	24	32	40	48	56
BDF	17.2	13.8	12.0	11.0	10.3	9.67	9.35	8.97
PCA	26.5	11.8	9.31	8.01	5.63	5.19	4.86	4.56
APCA	24.5	12.7	11.8	7.26	5.77	4.57	4.18	4.04
PLCDA	9.75	5.91	4.78	4.33	4.21	3.94	3.63	3.48
APCDA	4.64	3.02	2.66	2.31	2.07	1.96	1.78	1.77

Table 3 numerically records the false face rejection rates (in percent) at some small numbers of false detections out of the 8,951 nonface images with $d = 50$ and $M = 10$. It shows that the proposed APCDA method delivers significant lower false rejection rate than other methods.

5 Conclusion

This paper analyzes the role of PCA in the classification, which is far beyond a simple dimension reduction for a compact data representation with the least-mean-square reconstruction error. The crucial role of PCA in the classification is to remove the unreliable dimensions caused by insufficient or unrepresentative training data. This work demonstrates that applying PCA on the data total scatter matrix does not effectively remove the unreliable dimensions if one class is represented by its training data much better or much worse than the other class. The proposed APCA alleviates this problem by asymmetrically weighting the class conditional covariance matrices. In many real-world applications such as verification and object detection, the positive class represents only a single object, but the negative class is a much larger category composed of all objects of "the rest of the world." In general, the collected negative samples represent "the rest of the world" much worse than the positive samples. For such an asymmetric two-class problem, the proposed APCA is more effective than PCA in removing the unreliable dimensions.

APCA solves overfitting problems and hence leads to better generalization for the novel test data, but may not necessarily produce a compact feature set for fast classification. It is the discriminant method that plays an important role in extracting a compact feature set. For a two-class problem, LDA and CDA can be applied in the reliable APCA subspace. However, the amount of eigenvalue bias of one class may differ from that of the other class in the APCA subspace if the training data represent one class better or worse than the other class. This adversely affects the covariance discriminant evaluation and the classification. The proposed ADA integrates LDA and CDA in a single discriminant evaluation and regularizes the covariance matrix. It alleviates the problem of the biased eigenvalues in the APCA subspace, and hence, extracts the discriminatory features more effectively. Extensive experiments on the synthetic data and real image database demonstrate that the proposed APCDA approach consistently outperforms PCA, PCA+LDA+CDA, and BDF methods, which verifies the feasibility and effectiveness of the proposed methods.

Acknowledgments

This work was supported by Singapore A*STAR SERC research project grant no. 062 130 0056.

References

[1] M. Kirby and L. Sirovich, "Application of Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 1, pp. 103-108, Jan. 1990.
[2] H. Schneiderman and T. Kanade, "A Statistical Model for 3D Object Detection Applied to Faces and Cars," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2005.
[3] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.
[4] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, Aug. 1996.
[5] W. Liu, Y. Wang, S.Z. Li, and T.N. Tan, "Null Space Approach of Fisher Discriminant Analysis for Face Recognition," *Proc. European Conf. Computer Vision Workshop Biometric Authentication,* pp. 32-44, May 2004.
[6] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, and S.Z. Li, "Ensemble-Based Discriminant Learning with Boosting for Face Recognition," *IEEE Trans. Neural Networks,* vol. 17, no. 1, pp. 166-178, Jan. 2006.
[7] S. Zhang and T. Sim, "Discriminant Subspace Analysis: A Fukunaga-Koontz Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 10, pp. 1732-1745, 2007.
[8] X.D. Jiang, B. Mandal, and A. Kot, "Eigenfeature Regularization and Extraction in Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 3, pp. 383-394, Mar. 2008.
[9] P.F. Hsieh and D.A. Landgrebe, "Linear Feature Extraction for Multiclass Problems," *Proc. IEEE Int'l Geoscience and Remote Sensing Symp.,* vol. 4, pp. 2050-2052, 1998.
[10] P.F. Hsieh, D.S. Wang, and C.W. Hsu, "A Linear Feature Extraction for Multiclass Classification Problems Based on Class Mean and Covariance Discriminant Information," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 2, pp. 223-235, Feb. 2006.
[11] P. Viola and M. Jones, "Robust Real-time Object Detection," *Int'l J. Computer Vision,* vol. 57, no. 2, pp. 137-154, 2001.
[12] S.Z. Li and Z.Q. Zhang, "FloatBoost Learning and Statistical Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1112-1123, Sept. 2004.
[13] B. Moghaddam, "Principal Manifolds and Probabilistic Subspace for Visual Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 6, pp. 780-788, June 2002.
[14] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.
[15] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian Face Recognition," *Pattern Recognition,* vol. 33, no. 11, pp. 1771-1782, Nov. 2000.
[16] X.D. Jiang, B. Mandal, and A. Kot, "Enhanced Maximum Likelihood Face Recognition," *Electronics Letters,* vol. 42, no. 19, pp. 1089-1090, Sept. 2006.
[17] K.K. Sung and T. Poggio, "Example-based Learning for View-based Human Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 1, pp. 39-51, Jan. 1998.
[18] C. Liu, "A Bayesian Discriminating Features Method for Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 6, pp. 725-740, June 2003.
[19] S.L. Phung, A. Bouzerdoum, and D. Chai, "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, pp. 148-154, Jan. 2005.
[20] J.H. Friedman, "Regularized Discriminant Analysis," *J. Am. Statistical Assoc.,* vol. 84, no. 405, pp. 165-175, Mar. 1989.
[21] X. Wang and X. Tang, "A Unified Framework for Subspace Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1222-1228, Sept. 2004.
[22] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Academic Press, 1990.