Appears in the *IEEE Signal Processing Society 1998 International Conference on Image Processing*, **ICIP'98**, October 4-7, 1998, Chicago, Illinois, USA

A Unified Bayesian Framework for Face Recognition

Chengjun Liu and Harry Wechsler
Department of Computer Science, George Mason University,
4400 University Drive, Fairfax, VA 22030-4444, USA
{cliu, wechsler}@cs.gmu.edu

Abstract

This paper introduces a Bayesian framework for face recognition which unifies popular methods such as the eigenfaces and Fisherfaces and can generate two novel probabilistic reasoning models (PRM) with enhanced performance. The Bayesian framework first applies Principal Component Analysis (PCA) for dimensionality reduction with the resulting image representation enjoying noise reduction and enhanced generalization abilities for classification tasks. Following data compression, the Bayes classifier, which yields the minimum error when the underlying probability density functions (pdf) are known, carries out the recognition in the reduced PCA subspace using the Maximum A Posteriori (MAP) rule, which is the optimal criterion for classification because it measures class separability. The PRM models are described within this unified Bayesian framework and shown to yield better performance against both the eigenfaces and Fisherfaces methods.

1. Introduction

Face recognition is a standard pattern recognition problem whose actual performance depends not only on the features to be chosen, but also on the classifier being used [3], [13], [2]. Feature selection in pattern recognition involves the derivation of salient features from the raw input data in order to reduce the amount of data used for classification and simultaneously provide enhanced discriminatory power. The selection of an appropriate set of features often exploits the design criteria such as (A) redundancy minimization and decorrelation, (B) minimization of the reconstruction error, (C) maximization of information transmission (infomax) [9], and (D) sparseness of the neural code [10]. The Bayes classifier, on the other hand, yields the minimum error when the underlying probability density functions (pdf) are known. This error, known as the Bayes error, is the optimal measure for feature effectiveness when classification is of concern, since it is a measure of class separability.

In this paper we describe a unified Bayesian framework by combining a feature selection technique (principal component analysis, PCA [7]) and the Bayes classifier, and show that this framework unifies some popular face recognition methods and generates novel models with enhanced performance. Two such models, called probabilistic reasoning models (PRM-1 and PRM-2), first estimate the conditional pdf using the within class scatter, and then apply the Maximum *A Posteriori* (MAP) decision rule as the classification criterion. The within class scatter and MAP rule optimize the class separability in the sense of the Bayes error and should improve on PCA and Fisher Linear Discriminant (FLD) based methods, which utilize criteria not related to this error (Fukunaga [5]).

2. Related Research

Sirovich and Kirby [8] are among the first researchers who applied PCA for representing face images. They showed that any particular face can be economically represented along the eigenpictures coordinate space, and that any face can be approximately reconstructed by using just a small collection of eigenpictures and the corresponding projections ('coefficients') along each eigenpicture. Turk and Pentland [15] applied PCA further for recognizing faces and developed a well known face recognition method, known as **eigenfaces**, where the eigenfaces correspond to the eigenvectors associated with the dominant eigenvalues of the face covariance matrix. The eigenfaces define a feature space, or "face space", which drastically reduces the dimensionality of the original space, and face detection and identification are carried out in the reduced space.

The advantage of applying PCA directly for face recognition (eigenfaces) comes from its generalization ability [11]. PCA yields projection axes based on the variations from all the training samples, hence these axes are fairly robust for representing both training and testing images (not seen during training). The disadvantage of the direct PCA approach is that it does not distinguish the different roles

of the variations, and treats all of them equally. This will lead to poor performance when the distributions of the face classes are not separated by the mean-difference but separated by the covariance-difference [5].

To improve upon the performance of direct PCA approach, Belhumire, Hespanha, and Kriegman [1] developed an method called **Fisherfaces** by applying first PCA for dimensionality reduction and then Fisher's Linear Discriminant (FLD) [4] for discriminant analysis. Using a similar approach, Swets and Weng [14] have pointed out that the eigenfaces derived using PCA are only the most expressive features (MEF). The MEF are unrelated to actual face recognition, and in order to derive the most discriminating features (**MDF**), one needs a subsequent FLD projection. One can show that the MDF space is, however, superior to the MEF space for face recognition, only when the training images are representative of the range of face (class) variations; otherwise, the performance difference between the MEF and MDF is not significant.

The advantage of the indirect methods (combining PCA and FLD) is that they distinguish the different roles of within and between class scatter by applying discriminant analysis, e.g. FLD, and they usually produce non-orthogonal projection axes. The disadvantage of these methods comes from their poor generalization to new data, because they overfit to the training data. As the FLD procedure involves the simultaneous diagonalization of the two within and between class scatter matrices, it is equivalent to two-step operations: first 'whitening' the within class scatter matrix — applying an appropriate transformation that will make the within class scatter matrix equal to unity, and second applying PCA on the transformed between class scatter matrix [5]. The purpose of the 'whitening' step here is to normalize the within class scatter to unity, while the second step would then maximize the between class scatter. The robustness of the FLD procedure thus depends on whether or not the within class scatter can capture enough variations for a specific class. When the training samples do not include most of the variations due to lighting, facial expression, pose, and/or duplicate images as those encountered during testing, the 'whitening' step is likely to fit misleading variations. As a result the normalized within class scatter would best fit the training samples but it would generalize poorly when exposed to new data.

The unified Bayesian framework detailed in the following section generalizes some popular statistical face recognition methods and can generate two novel probabilistic reasoning models (PRM) with enhanced performance. Experimental results using more than one thousand images from the FERET database show that the PRM models improve the recognition performance as compared to direct PCA approach (eigenfaces) and enhance the generalization capability of the indirect methods (Fisherfaces/MDF).

3. A Unified Bayesian Framework

Integrating principal component analysis technique and Bayes classifier leads to a unified Bayesian framework which incorporates some of the popular face recognition methods and generates novel models with enhanced performance. First, PCA, as applied for dimensionality reduction, condenses the original image space into a compact one with the merits of suppressing noises and enhancing generalization as well as battling "the curse of dimensionality". Then, the Bayes classifier, which yields the minimum error when the underlying probability density functions are known, carries out the recognition in the reduced PCA subspace using the MAP rule, which is the optimal measure for classification because it measures class separability.

3.1. Principal Component Analysis

The rationale behind applying PCA first for dimensionality reduction instead of exploiting the Bayes classifier and the MAP rule directly on the original data is two-fold. On the one hand, the high-dimensionality of the original face space makes the pdf estimation very difficult, if not impossible, due to the fact that high-dimensional space is mostly empty. This problem of sparsity limits the success of direct Bayesian analysis in the original space, since the amount of training data needed to get reasonably low variance estimators becomes ridiculously high [6]. On the other hand, it has been confirmed by many researchers that the PCA representation has the feature of object constancy in the sense that it suppresses input noise [8], [11].

PCA generates a set of orthonormal basis vectors, known as principal components, that maximize the scatter of all the projected samples. Let $X = [X_1, X_2, \ldots, X_n]$ be the sample set of the original images. After normalizing the images to unity norm and subtracting the grand mean a new image set $Y = [Y_1, Y_2, \ldots, Y_n]$ is derived. Each Y_i represents a normalized image with dimensionality N, $Y_i = (y_{i_1}, y_{i_2}, \ldots, y_{i_N})^t$, $(i = 1, 2, \ldots, n)$. The covariance matrix of the normalized image set is defined as

$$\Sigma_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^t = \frac{1}{n} Y Y^t \tag{1}$$

and the eigenvector and eigenvalue matrices Φ, Λ are computed as

$$\Sigma_Y \Phi = \Phi \Lambda \tag{2}$$

Note that YY^t is an $N \times N$ matrix while Y^tY is an $n \times n$ matrix. If the sample size n is much smaller than the dimensionality N, then the following method saves some computation [15]

$$(Y^tY)\Psi = \Psi \Lambda_1 \tag{3}$$

$$\Im = Y\Psi \tag{4}$$

where $\Lambda_1 = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, and $\Im = [\Phi_1, \Phi_2, \ldots, \Phi_n]$. If one assumes that the eigenvalues are sorted in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, then the first m leading eigenvectors define a matrix P

$$P = [\Phi_1, \Phi_2, \ldots, \Phi_m] \tag{5}$$

The new feature set Z with lower dimensionality m ($m \ll N$) is then computed as

$$Z = P^t Y \tag{6}$$

3.2. Bayes Classifier

For pattern recognition, the Bayes classifier is the best classifier, the Bayes error the best criterion to evaluate feature sets, and *a posteriori* probability functions are thus optimal features [5]. Let $\omega_1, \omega_2, \ldots, \omega_L$ denote the object classes, and Z an image in the reduced PCA subspace. The *a posteriori* probability function of ω_i given Z is defined as

$$P(\omega_i|Z) = \frac{p(Z|\omega_i)P(\omega_i)}{p(Z)} \tag{7}$$

where $P(\omega_i)$ is *a priori* probability, $p(Z|\omega_i)$ the conditional probability density function of ω_i, and $p(Z)$ is the mixture density. The Maximum *A Posteriori* decision rule for the Bayes classifier is

$$p(Z|\omega_i)P(\omega_i) = \max_j \{p(Z|\omega_j)P(\omega_j)\}, \quad Z \in \omega_i \tag{8}$$

The face image Z is classified to ω_i of whom the *A Posteriori* probability given Z is the largest among all the classes.

Usually there are not enough samples to estimate the conditional probability density function for each class (within class density). A compromise, therefore, is to make an assumption of a particular density form, and convert the general density estimation question into a parametric one. The within class densities are usually modeled as normal distributions

$$\begin{aligned} p(Z|\omega_i) &= \frac{1}{(2\pi)^{m/2}|\Sigma_i|^{1/2}} \\ &\times exp\left\{-\frac{1}{2}(Z - M_i)^t \Sigma_i^{-1}(Z - M_i)\right\} \end{aligned} \tag{9}$$

where M_i (see Eq. 10) and Σ_i are the mean and covariance matrix of class ω_i, respectively.

4. Face Recognition Methods

Estimating the covariance matrix Σ_i in Eq. 9 with respect to each class is still difficult due to the limited number of samples for each class. Note that while the mixture covariance matrix is diagonal following PCA, the within class covariance matrices are not necessarily diagonal. Further assumptions lead to different face recognition methodologies.

4.1. Eigenfaces

Assume the within class covariance matrices to be unit matrices: $\Sigma_I = \Sigma_i = I_m$, and under this assumption the conditional pdf (Eq. 9) relaxes to $p(Z|\omega_i) = \frac{1}{(2\pi)^{m/2}} exp\{-\frac{1}{2}(Z - M_i)^t(Z - M_i)\}$. As a result, the MAP rule (Eq. 8) leads to a distance classifier which corresponds to the eigenfaces method by Turk and Pentland [15].

4.2. Fisherfaces

After PCA (Eq. 6), apply FLD analysis to derive a new feature set $W = P_{FLD}^t Z$. In this FLD subspace, assume all the within class covariance matrices to be unit matrices: $\Sigma_I = \Sigma_i = I_m$, and under this assumption the conditional pdf (Eq. 9) reduces to $p(W|\omega_i) = \frac{1}{(2\pi)^{m/2}} exp\{-\frac{1}{2}(W - M_i')^t(W - M_i')\}$. Again the MAP rule (Eq. 8) leads to a distance classifier which corresponds to Fisherfaces method by Belhumeur, Hespanha, and Kriegman [1].

5. Probabilistic Reasoning Models (PRM)

Under the unified Bayesian framework, we derived two new probabilistic reasoning models, PRM-1 and PRM-2, which utilize the within class scatters to derive averaged estimations of within class covariance matrices. For the PRM-1 model, in the reduced PCA subspace, we assume all the within class covariance matrices are identical, diagonal, and each diagonal element is estimated by the sample variance in the one dimensional PCA subspace. As a result, the conditional pdf specifies the MAP rule as a quadratic classifier characterized by Mahalanobis distance. For PRM-2, we first compute the averaged within class covariance matrix based on all the within class scatters in the reduced PCA subspace, then diagonalize this covariance matrix, and use the ordered diagonal elements as estimations of the within class covariance matrices which are assumed to be diagonal. As a result, PRM-2 derives another quadratic classifier.

In particular, let $\omega_1, \omega_2, \ldots, \omega_L$ and $N_1, N_2, \ldots, N_L$ denote the classes and number of images within each class, respectively. Let $M_1, M_2, \ldots, M_L$ be the means of the classes in the reduced PCA subspace $span[\Phi_1, \Phi_2, \ldots, \Phi_m]$. We then have

$$M_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Z_j^{(i)}, \qquad i = 1, 2, \ldots, L \tag{10}$$

where $Z_j^{(i)}, j = 1, 2, \ldots, N_i$, represents the sample images from class ω_i.

5.1. PRM-1

The PRM-1 model assumes the within class covariance matrices are identical and diagonal

$$\Sigma_I = \Sigma_i = diag\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2\} \tag{11}$$

Each component σ_i^2 can be estimated by sample variance in the one dimensional PCA subspace

$$\sigma_i^2 = \frac{1}{L}\sum_{k=1}^{L}\left\{\frac{1}{N_k - 1}\sum_{j=1}^{N_k}\left(z_{ji}^{(k)} - m_{ki}\right)^2\right\} \tag{12}$$

where $z_{ji}^{(k)}$ is the i-th element of the sample $Z_j^{(k)}$, m_{ki} the i-th element of M_k, and L the number of classes.

From Eqs. 11 and 9, it follows

$$p(Z|\omega_i) = \frac{1}{(2\pi)^{m/2}\prod_{j=1}^{m}\sigma_j} exp\left\{-\frac{1}{2}\sum_{j=1}^{m}\frac{(z_j - m_{ij})^2}{\sigma_j^2}\right\} \tag{13}$$

Thus the MAP rule (Eq. 8) specifies a quadratic classifier characterized by Mahalanobis distance (note that the *priors* are set to be equal).

$$\sum_{j=1}^{m}\frac{(z_j - m_{ij})^2}{\sigma_j^2} = \min_k\left\{\sum_{j=1}^{m}\frac{(z_j - m_{kj})^2}{\sigma_j^2}\right\} \Longrightarrow Z \in \omega_i \tag{14}$$

5.2. PRM-2

The PRM-2 model estimates the within class scatter matrix Σ_w in the reduced PCA subspace as

$$\Sigma_w = \frac{1}{L}\sum_{k=1}^{L}\left\{\frac{1}{N_k}\sum_{j=1}^{N_k}\left(Z_j^{(k)} - M_k\right)\left(Z_j^{(k)} - M_k\right)^t\right\} \tag{15}$$

To avoid the explicit calculation of Σ_w and to improve numerical accuracy, we calculate the singular value decomposition (SVD) of matrix Z, where $\Sigma_w = ZZ^t$.

$$Z = USV^t \tag{16}$$

where U and V are unitary matrices, S is a diagonal one

$$S = diag\{s_1, s_2, \ldots, s_m\} \tag{17}$$

with non-negative singular values as diagonal elements. Order the squared diagonal elements as

$$\left(s_{(1)}^2, s_{(2)}^2, \ldots, s_{(m)}^2\right) = order\left\{s_1^2, s_2^2, \ldots, s_m^2\right\} \tag{18}$$

Finally, the within class covariance matrix is derived as

$$\Sigma_I = diag\left\{s_{(1)}^2, s_{(2)}^2, \ldots, s_{(m)}^2\right\} \tag{19}$$

and the corresponding pdf is specified as

$$p(Z|\omega_i) = \frac{1}{(2\pi)^{m/2}\prod_{j=1}^{m}s_{(j)}} exp\left\{-\frac{1}{2}\sum_{j=1}^{m}\frac{(z_j - m_{ij})^2}{s_{(j)}^2}\right\} \tag{20}$$

Thus the MAP rule defines another quadratic classifier (note that the *priors* are set to be equal).

$$\sum_{j=1}^{m}\frac{(z_j - m_{ij})^2}{s_{(j)}^2} = \min_k\left\{\sum_{j=1}^{m}\frac{(z_j - m_{kj})^2}{s_{(j)}^2}\right\} \Longrightarrow Z \in \omega_i \tag{21}$$

6. Experimental Results

The experimental data consists of 1,107 facial images corresponding to 369 subjects and comes from the US Army FERET database [12]. 600 out of the 1,107 images correspond to 200 subjects with each subject having three images — two of them are the first and the second shot, and the third shot is taken under low illumination. For the remaining 169 subjects there are also three images for each subject, but two out of the three images are duplicates taken at a different time. Two images of each subject are used for training with the remaining image for testing. The images are cropped to the size of 64 x 96, and the eye coordinates are manually detected.

Fig. 1 shows the comparative performances of Eigenfaces, Fisherfaces, PRM-1 and PRM-2 methods for the top 1 recognition rates. Top 1 recognition rate means the accuracy rate for the top response being correct, while top 3 recognition rate represents the accuracy rate for the correct response being included among the first three ranked choices. The top 3 recognition rates for the testing performance of the methods are plotted in Fig. 2. For the top 1 recognition rates, the PRM models (PRM-1 and PRM-2) increase the recognition rate by 5% when compared to eigenfaces and Fisherfaces methods; the peak recognition rate for both PRMs is about 96% using 44 features. PRM-1 and PRM-2 models also increase the top 3 recognition rate by 3% when compared to eigenfaces and Fisherfaces methods with a peak recognition rate of about 99%.

7. Conclusions

By integrating the principal component analysis and the Bayes classifier, we have derived a unified Bayesian framework, which includes many of the popular face recognition

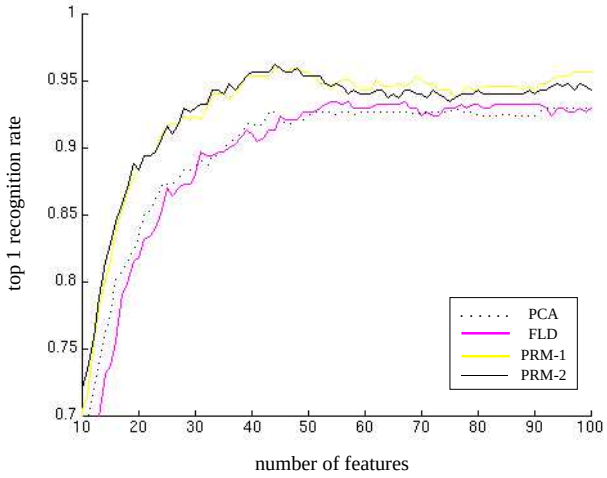

Figure 1. Comparative Testing Performances (top 1 recognition rate) for the PCA (eigenfaces), FLD (Fisherfaces method), and PRM Approach (PRM-1, PRM-2).

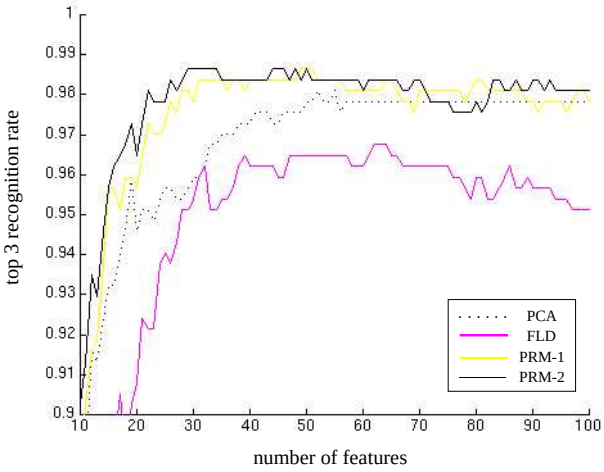

Figure 2. Comparative Testing Performances (top 3 recognition rate) for the PCA (eigenfaces), FLD (Fisherfaces method), and PRM Approach (PRM-1, PRM-2).

methods and generates novel models with enhanced performance. Two novel probabilistic reasoning models (PRM-1 and PRM-2), operating in the reduced PCA subspace, optimize the class separability in the sense of the Bayes error. Experimental results show that the PRM achieve better performance on face recognition when compared against both PCA and Fisherfaces methods.

Acknowledgments: This work was partially supported by the DoD Counterdrug Technology Development Program, with the U.S. Army Research Laboratory as Technical Agent, under contract DAAL01-97-K-0118.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fi sherfaces: Recognition using class specifi c linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] R. Brunelli and T. Poggio. Face recognition: Features vs. templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(10):1042–1053, 1993.

[3] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proc. IEEE*, 83(5):705–740, 1995.

[4] R. Fisher. The use of multiple measures in taxonomic problems. *Ann. Eugenics*, 7:179–188, 1936.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1991.

[6] N. Intrator and L. Cooper. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17, 1992.

[7] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, 1986.

[8] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[9] R. Linsker. Self-organization in a perceptual network. *Computer*, 21:105–117, 1988.

[10] B. Olshausen and D. Field. Emergence of simple-cell receptive fi eld properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[11] P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7:477–500, 1996.

[12] P. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET september 1996 database and evaluation procedure. In *Proc. First Int'l Conf. on Audio and Video-based Biometric Person Authentication*, pages 12–14, Switzerland, 1997.

[13] A. Samal and P. Iyengar. Automatic recognition and analysis of human faces and facial expression: A survey. *Pattern Recognition*, 25(1):65–77, 1992.

[14] D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. on PAMI*, 18(8):831–836, 1996.

[15] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 13(1):71–86, 1991.