Unified Subspace Analysis for Face Recognition

Xiaogang Wang and Xiaoou Tang

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

Abstract

We propose a face difference model that decomposes face difference into three components, intrinsic difference, transformation difference, and noise. Using the face difference model and a detailed subspace analysis on the three components we develop a unified framework for subspace analysis. Using this framework we discover the inherent relationship among different subspace methods and their unique contributions to the extraction of discriminating information from the face difference. This eventually leads to the construction of a 3D parameter space that uses three subspace dimensions as axis. Within this parameter space, we develop a unified subspace analysis method that achieves better recognition performance than the standard subspace methods on over 2000 face images from the FERET database.

1. Introduction

Many face recognition techniques have been developed over the past few decades [6]. Among the existing face recognition techniques, subspace methods are widely used to reduce the high dimensionality of the raw face image. Eigenface method (PCA) [5] is a first breakthrough for the subspace techniques. It uses the Karhunen-Loeve Transform (KLT) to produce a most expressive subspace for face representation and recognition. LDA or Fisher Face [1], is an example of the most discriminating subspace methods. Linear discriminant analysis is adopted to seek a set of features best separating face classes. The Bayesian algorithm using probabilistic subspace is proposed in [3]. It casts the face recognition problem as classifying intrapersonal and extrapersonal variations.

In this work, we develop a unified subspace analysis method based on a new framework for the three subspace face recognition methods: PCA, LDA and Bayesian algorithms. As discussed earlier, they represent three major approaches for subspace based face recognition. PCA has become an evaluation benchmark for face recognition. Both LDA and Bayesian algorithms achieved superior performance in FERET competition [4]. A unified framework on the three methods will greatly help to understand the family of subspace methods.

We first propose a face difference model decomposing face difference into three components, intrinsic difference $\tilde{I}$, transformation difference $\tilde{T}$, and noise $\tilde{N}$. A unified framework is then constructed using this face difference model and a detailed subspace analysis on the three components. Using this framework we discover the inherent relationship among different subspace methods and their unique contributions to the extraction of discriminating information from the face difference. This eventually leads to the construction of a 3D parameter space that uses the three subspace dimensions as axis. Within this parameter space, we develop a unified subspace analysis method that achieves better recognition performance than the standard subspace methods.

2. Review of subspace methods

We formulate the face recognition problem as following. A 2D face image is viewed as a vector in the image space. A set of sample face images $\{\vec{x}_i\}$ can be represented by an N by M matrix $X = [\vec{x}_1, \ldots \vec{x}_M]$, where M is the number of samples and N is the number of pixels in the images. Each face image $\vec{x}_i$ belongs to one of the L individual classes $\{X_1, \ldots X_L\}$, and $\ell(\vec{x}_i)$ is the class label for $\vec{x}_i$. When a test image $\vec{T}$ is the input, the face recognition task is to find its class in the database. Based on this formulation, a short review for the PCA, LDA, and Bayes approaches is given in this section.

2.1 PCA

In the PCA method, a set of eigenfaces are typically computed from the eigenvectors of sample covariance matrix C,

$$C = \sum_{i=1}^{M} (\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})^T , \quad (1)$$

where $$\vec{m} = \frac{1}{M} \sum_{i=1}^{M} \vec{x}_i . \quad (2)$$

The eigenspace U is spanned by K eigenfaces with the largest eigenvalues, $U = [\vec{u}_1, \ldots \vec{u}_K]$. In the recognition process, the prototype $\vec{P}$ for each face class and the testing image $\vec{T}$ are projected onto the eigenspace to get the weight vectors,

$$\vec{w}_p = U^T (\vec{P} - \vec{m}). \quad (3)$$

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE

$$\vec{w}_T = U^T\left(\vec{T} - \vec{m}\right). \quad (4)$$

The face class is found to minimize the distance

$$\varepsilon = \| \vec{w}_T - \vec{w}_p \| . \quad (5)$$

2.2 LDA

LDA finds the subspace best discriminating different face classes. It is carried out by maximizing the between-class scatter matrix S_b and minimizing the within-class scatter matrix S_w in the projective subspace. S_b and S_w are defined as

$$S_w = \sum_{i=1}^{L} \sum_{\vec{x}_k \in X_i} (\vec{x}_k - \vec{m}_i)(\vec{x}_k - \vec{m}_i)^T , \quad (6)$$

$$S_b = \sum_{i=1}^{L} n_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T , \quad (7)$$

where $\vec{m}_i$ is the mean face for the class X_i, and n_i is the number of samples in class X_i.

The subspace for LDA is spanned by a set of vectors $W = [\vec{w}_1, \ldots, \vec{w}_{L-1}]$, satisfying

$$W = \arg\max \left| \frac{W^T S_b W}{W^T S_w W} \right| , \quad (8)$$

where W can therefore be constructed by the eigenvectors of $S_w^{-1} S_b$.

Computing the eigenvectors of $S_w^{-1} S_b$ is equivalent to simultaneous diagonalization of S_w and S_b [2]. First S_w is whitened by

$$\Theta^{-1/2} \Phi^T S_w \Phi \Theta^{-1/2} = I \quad (9)$$

where Φ, Θ are the eigenvector matrix and eigenvalue matrix of S_w. Second, apply PCA on class centers using the transformed data. Projecting the class centers onto $\Theta^{-1/2}\Phi^T$, S_b is transformed to K_b as,

$$K_b = \Theta^{-1/2} \Phi^T S_b \Phi \Theta^{-1/2} . \quad (10)$$

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

$$W = \Phi \Theta^{-1/2} \Psi . \quad (11)$$

As shown in [2], W is the eigenvector matrix of $S_w^{-1} S_b$.

The face class is chosen to minimize the linear discriminant function,

$$\| d(\vec{T}) \| = \| W^T (\vec{T} - \vec{P}) \| . \quad (12)$$

To avoid degeneration of S_w, most LDA methods usually first reduce the data dimensionality by PCA, then apply discriminant analysis in the reduced PCA space.

2.3 Bayesian algorithm

The Bayesian algorithm classifies the face intensity difference Δ as intrapersonal variation (Ω_I) for the same individual and extrapersonal variation (Ω_E) for different individuals [3]. The MAP similarity between two images is defined as the intrapersonal a posterior probability

$$S(I_1, I_2) = P(\Omega_I \mid \Delta) = \frac{P(\Delta \mid \Omega_I) P(\Omega_I)}{P(\Delta \mid \Omega_I) P(\Omega_I) + P(\Delta \mid \Omega_E) P(\Omega_E)} . \quad (13)$$

To estimate $P(\Delta \mid \Omega_I)$, PCA on the set $\{\Delta \mid \Delta \in \Omega_I\}$ decomposes the image difference space into intrapersonal principal subspace F and its orthogonal complementary space $\bar{F}$. The likehood can be estimated as,

$$\hat{P}(\Delta \mid \Omega_I) = \left[\frac{\exp\left(-\frac{1}{2} d_F(\Delta)\right)}{(2\pi)^{K/2} \prod_{i=1}^{K} \lambda_i^{1/2}} \right] \left[\frac{\exp\left(-\varepsilon^2(\Delta)/2\rho\right)}{(2\pi\rho)^{(N-K)/2}} \right] . \quad (14)$$

In Eq. (14), $d_F(\Delta)$ is a Mahalanobis distance in F, referred as "distance-in-feature-space" (DIFS),

$$d_F(\Delta) = \sum_{i=1}^{K} \frac{y_i^2}{\lambda_i} , \quad (15)$$

where y_i is the principal component and λ_i is the eigenvalue. $\varepsilon^2(\Delta)$ is defined as "distance-from-feature-space" (DFFS), equivalent to PCA residual error in $\bar{F}$. ρ is the average eigenvalue in $\bar{F}$. $P(\Delta \mid \Omega_E)$ can be estimated in a similar way. The principal subspace computed from the set $\{\Delta \mid \Delta \in \Omega_E\}$ is called extrapersonal eigenspace.

An alternative maximum likehood (ML) measure is defined as

$$S'(\Delta) = P(\Delta \mid \Omega_I) . \quad (16)$$

It has been shown to be simpler but almost as effective as the MAP measure in Eq. (13) [3].

3. A unified framework

In this section, we construct a unified framework revealing the intrinsic connections of the three methods. Let us first look at the matching criterions and focus on the difference $\Delta = \vec{T} - \vec{P}$ between the testing image $\vec{T}$ and the prototype $\vec{P}$. The matching criterion for PCA in Eq. (5) can be rewritten as

$$\varepsilon_{PCA} = \| U^T(\vec{T} - \vec{m}) - U^T(\vec{P} - \vec{m}) \| = \| U^T(\Delta) \| \quad (17)$$

For LDA, according to Eq. (12), the linear discriminant function can also be expressed in terms of Δ,

$$\varepsilon_{LDA} = \| W^T \Delta \| . \quad (18)$$

Finally, for the Bayesian algorithm using ML measure, the similarity measure of Eq. (14) can be evaluated as a distance measure,

$$\varepsilon_{Bayes} = d_F(\Delta) + \varepsilon^2(\Delta)/\rho \quad (19)$$

From Eq. (17), (18), and (19), we see that the recognition process of the three methods can be shown by a simple framework in Fig. 1. When a testing face image $\vec{T}$ is input, it is first subtracted of each class prototype $\vec{P}$. The difference Δ is projected onto an image subspace to

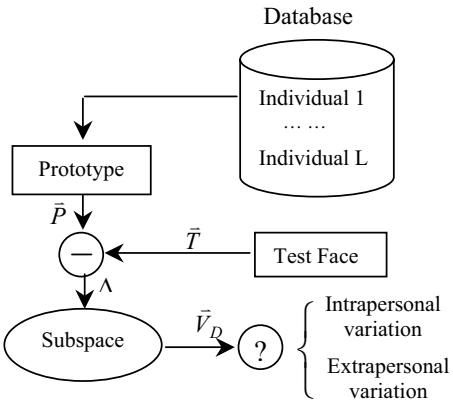


Figure 1. Diagram of the unified framework for face recognition.

extract the feature vector and evaluated to be intrapersonal variation or extrapersonal variation.

The two central components of this framework are the image difference Δ and the subspace onto which Δ is projected. We model the difference Δ by three components: intrinsic difference ($\widetilde{I}$) that discriminates different individuals; transformation difference ($\widetilde{T}$), arising from all kinds of transformations, such as expressions, illuminations, and view changes; noise ($\widetilde{N}$), randomly distributed in the face images.

The intrapersonal variation Ω_I is composed of $\widetilde{T}$ and $\widetilde{N}$, since it comes from the same individual. For Ω_E, $\widetilde{I}$, $\widetilde{T}$ and $\widetilde{N}$, are coupled together. Therefore, we have,

$$\Omega_I = \widetilde{T} + \widetilde{N}, \tag{20}$$

$$\Omega_E = \widetilde{I} + \widetilde{T} + \widetilde{N}. \tag{21}$$

$\widetilde{T}$ and $\widetilde{N}$ are the two components deteriorating the recognition performance. Normally, $\widetilde{N}$ is of small energy. The main difficulty for face recognition comes from $\widetilde{T}$. Under a large transformation, $\widetilde{T}$ can potentially be greater than $\widetilde{I}$. A successful approach should reduce the energy of $\widetilde{T}$ and $\widetilde{N}$ as much as possible without sacrificing much of $\widetilde{I}$. We now analyze the behavior of the three subspaces for PCA, LDA and Bayes in order to discover how they process the three components.

3.1 Eigenspace for PCA

Eigenfaces are computed from the ensemble covariance matrix C. We can show that C can also be computed from the set, $\{(\vec{x}_i - \vec{x}_j)\}$, containing all the differences between any pair of face images in the training set.

Theorem 1. The eigenspace of PCA characterizes the difference between any two face images, which may belong to the same individual or different individuals.

Proof. We only need to show that the covariance matrix C for $\{\vec{x}_i\}$ can also be computed as

$$C = \frac{1}{2M}\sum_{i=1}^{M}\sum_{j=1}^{M}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T .$$

From Eq. (1) we have

$$C = \sum_{i=1}^{M}(\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})^T .$$

Replace $\vec{m}$ with Eq. (2),

$$C = \sum_{i=1}^{M}\left(\vec{x}_i - \frac{\vec{x}_1 + \ldots + \vec{x}_M}{M}\right)\left(\vec{x}_i - \frac{\vec{x}_1 + \ldots + \vec{x}_M}{M}\right)^T$$

$$= \frac{1}{M^2}\sum_{i=1}^{M}\left[\sum_{j=1}^{M}\sum_{k=1}^{M}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_k)^T\right]. \tag{22}$$

Rewrite C using different subscripts (exchange i and j),

$$C = \frac{1}{M^2}\sum_{j=1}^{M}\left[\sum_{i=1}^{M}\sum_{k=1}^{M}(\vec{x}_j - \vec{x}_i)(\vec{x}_j - \vec{x}_k)^T\right].$$

Change the order of summation,

$$C = \frac{1}{M^2}\sum_{i=1}^{M}\left[\sum_{j=1}^{M}\sum_{k=1}^{M}(\vec{x}_j - \vec{x}_i)(\vec{x}_j - \vec{x}_k)^T\right] \tag{23}$$

Average (22) and (23),

$$C = \frac{1}{2}\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T$$

$$= \frac{1}{2M}\sum_{i=1}^{M}\sum_{j=1}^{M}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T . \tag{24}$$

Removing the scale *1/2M* will not affect the eigenvectors of C, thus

$$C = \sum_{i=1}^{M}\sum_{j=1}^{M}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T \tag{25}$$

Therefore, C is also the covariance matrix for the face difference set $\{(\vec{x}_i - \vec{x}_j)\}$.

3.2 Intrapersonal and Extrapersonal Subspaces

In the Bayesian algorithm, the eigenvectors of intrapersonal subspace are computed from the image difference set $\{(\vec{x}_i - \vec{x}_i) \mid \ell(\vec{x}_i) = \ell(\vec{x}_j)\}$, for which the covariance matrix is

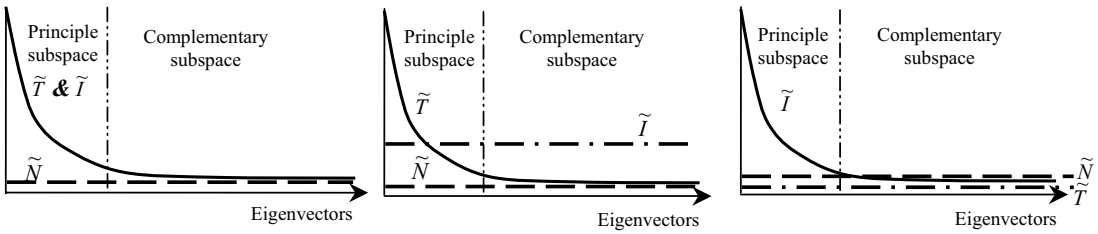
$$C_I = \sum_{\ell(\vec{x}_i)=\ell(\vec{x}_j)}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T . \tag{26}$$

The eigenvectors of extrapersonal subspace are derived from the difference set $\{(\vec{x}_i - \vec{x}_i) \mid \ell(\vec{x}_i) \neq \ell(\vec{x}_j)\}$, with covariance matrix

$$C_E = \sum_{\ell(\vec{x}_i)\neq\ell(\vec{x}_j)}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T . \tag{27}$$

Comparing C_I and C_E with C, we derive the following theorem,

Theorem 2. The intrapersonal and extrapersonal subspaces are the two components of the PCA eigenspace, and the extrapersonal eigenfaces are similar



(a) PCA subspace (b) Intrapersonal subspace (c) LDA subspace

Figure 2. Energy distribution of the three components $\tilde{I}$, $\tilde{T}$, and $\tilde{N}$ on eigenvectors in the PCA subspace (a), the intrapersonal subspace (b) and the LDA subspace(c).

to the PCA eigenfaces.

Proof. From Eq. (25), (26) and (27) we have

$$C = C_I + C_E . \quad (28)$$

C is composed of C_I and C_E. Therefore the intrapersonal and extrapersonal subspaces are the two components of the PCA eigenspace. Since the sample number for C_E is far greater than that of C_I, the energy of C_E dominates the computation of C. So the extrapersonal eigenfaces are similar to the standard eigenfaces.

In Ω_E, $\tilde{T}$ and $\tilde{I}$ are coupled. Therefore, as discussed later the extrapersonal subspace, which is similar to the PCA eigenspace, cannot contribute much to separating $\tilde{T}$ and $\tilde{I}$. This shows that the improvement of the Bayesian algorithm over the PCA mostly benefits from the intrapersonal subspace. It demonstrates that why the ML measure using the intrapersonal subspace alone is almost as effective as the MAP measure using two subspaces [3].

3.3 Subspace for LDA

The subspace for LDA is derived from the within-class scatter matrix and the between-class scatter matrix. We also can study the LDA subspace using image difference.

Theorem 3. The within-class scatter matrix is identical to C_I, the covariance matrix of the intrapersonal subspace, which characterizes the face variation for the same individuals. Using the mean face image to describe each individual class, the between-class scatter matrix characterizes the variation between mean face images.

Proof. For simplicity, we assume that each class has the same sample number *n*. Similar to the proof of Theorem 1, we have,

$$S_W = \sum_{i=1}^{L} \sum_{\vec{x}_k \in X_i} (\vec{x}_k - \vec{m}_i)(\vec{x}_k - \vec{m}_i)^T = \frac{1}{2n} \sum_{i=1}^{L} \sum_{\vec{x}_{k_1}, \vec{x}_{k_2} \in X_i} (\vec{x}_{k_1} - \vec{x}_{k_2})(\vec{x}_{k_1} - \vec{x}_{k_2})^T \quad (29)$$

Therefore, $S_W = C_I$.

$$S_b = \sum_{i=1}^{L} n(\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T = \frac{n}{2L} \sum_{i=1}^{L} \sum_{j=1}^{L} (\vec{m}_i - \vec{m}_j)(\vec{m}_i - \vec{m}_j)^T \quad (30)$$

This shows that S_b is the covariance matrix of the face difference set $\{(\vec{m}_i - \vec{m}_j)\}$.

3.4 Comparison of the three subspaces

We now investigate how these subspaces separate discriminating information $\tilde{I}$ from the deteriorating factors $\tilde{T}$ and $\tilde{N}$.

As shown in Fig. 2 (a), in the PCA subspace, both $\tilde{T}$ and $\tilde{I}$, as structured signals embedded in the original face image, concentrate on the small number of principal eigenvectors. By selecting the principal components, most noise encoded on the large number of trailing eigenvectors is removed from $\tilde{T}$ and $\tilde{I}$. Because of the presence of $\tilde{T}$, the PCA subspace is not ideal for face recognition.

For the Bayesian algorithm, the intrapersonal subspace plays a critical role. Since intrapersonal variation only contains $\tilde{T}$ and $\tilde{N}$, PCA on intrapersonal variation arranges the axes according to the energy distribution of $\tilde{T}$, as shown in Fig. 2 (b). When we project a face difference Δ onto the intrapersonal subspace, most energy of the $\tilde{T}$ component will concentrate on the first few largest eigenvectors, while the $\tilde{I}$ and $\tilde{N}$ components are randomly distributed over all of the eigenvectors. This is because $\tilde{I}$ and $\tilde{N}$ are somewhat independent of $\tilde{T}$, which forms the principal vectors of the intrapersonal subspace. In Eq. (19), the Mahalanobis distance in F weights the feature vectors by the inverse of eigenvalues. It effectively reduces the $\tilde{T}$ component since the principal components with large eigenvalues are significantly diminished. $\varepsilon^2(\Delta)$ is also a distinctive component for recognition, since it throws away most of the component $\tilde{T}$ on the largest eigenvectors, while keeps the majority of $\tilde{I}$.

The Bayesian algorithm successfully separates $\tilde{T}$ from $\tilde{I}$. However, $\tilde{I}$ and $\tilde{N}$ are still coupled on the small

eigenvectors. Even though $\widetilde{N}$ is usually of small energy, when it is normalized by the small eigenvalues as shown in Eq. (15) and (19), the effect of $\widetilde{N}$ could be significantly enlarged in the probabilistic measure.

Finally, we look at the LDA subspace. The LDA procedure can be divided into three steps. First, PCA is used to reduce the data dimension. Same as discussed earlier, noise $\widetilde{N}$ is significantly reduced in this step. In the second step, to whiten the within-class scatter matrix we first compute its eigenvector matrix Φ and eigenvalue matrix Θ. From Theorem 3, we know that Φ spans the intrapersonal subspace, therefore Θ essentially represents the energy distribution of $\widetilde{T}$. The whitening process projects data onto intrapersonal subspace Φ and normalizes them by $\Theta^{-1/2}$. Therefore this step reduces $\widetilde{T}$ in a manner similar to the Bayes analysis.

In the third step of LDA, PCA is again applied on the whitened class centers. Through averaging to compute the class centers, the noise $\widetilde{N}$ is further reduced in this step. This is useful since $\widetilde{N}$ may have been enlarged in the second step whitening process. Since both $\widetilde{T}$ and $\widetilde{N}$ have been reduced up to this point, the main energy in the class centers is the intrinsic difference $\widetilde{I}$. However, as shown in Fig. 2 (b), $\widetilde{I}$ is obtained by discarding principal component $\widetilde{T}$ in the intrapersonal subspace, so $\widetilde{I}$ may spread over the entire axis after the whitening. PCA on the class centers therefore serves two purposes. First, it further reduces the noise as PCA usually does. Second, it concentrates the energy of $\widetilde{I}$ on to a small number of principal components, as shown in Fig. 2 (c).

The subspace analysis results of the three methods on the face difference model are summarized in Table 1. We can clearly see the unique contribution of each subspace to the processing of the face difference model.

Table 1. Behavior of the subspaces on characterizing the face difference

Algorithm	Subspace	Decompose Face Image Difference	
		Principle Space	Complementary Space
PCA	Eigenspace	$\widetilde{I}+\widetilde{T}$	$\widetilde{N}$
LDA	Subspace for LDA	$\widetilde{I}$	$\widetilde{T}+\widetilde{N}$
Bayes	Intrapersonal subspace	$\widetilde{T}$	$\widetilde{I}+\widetilde{N}$
	Extrapersonal subspace	$\widetilde{I}+\widetilde{T}$	$\widetilde{N}$

4. Unified subspace analysis

There are two major difficulties for subspace based face recognition: small number of samples for each class and large number of classes. First, if there are too few samples for each class, the training set used to derive the intrapersonal subspace may not contain all the transformations in the testing set. So $\widetilde{T}$ cannot be effectively estimated and reduced. Second, for a large class number, it is difficult to effectively extract all the intrinsic difference $\widetilde{I}$ to cover the differences between every two individuals.

In order to alleviate these two problems, using the above new framework, we propose a unified subspace analysis method for face recognition as follows:

(1) Project face vectors to PCA subspace and adjust the PCA dimension (*dp*) to reduce most noise.

(2) Apply Bayesian analysis in the reduced PCA subspace and adjust the dimension (*di*) of intrapersonal subspace. Since human faces share similar intrapersonal variation, the transformation $\widetilde{T}$ for a testing individual can be estimated from faces of others. Therefore, our intrapersonal subspace is computed from an enlarged intrapersonal difference set that contains individuals both inside and outside of the gallery, so that the intrapersonal subspace is robust to all the transformations in the testing set.

(3) For the *L* individuals in the gallery, compute their training data class centers. Project all the class centers onto the intrapersonal subspace, and then normalize the projections by intrapersonal eigenvalues to compute the whitened feature vectors.

(4) Apply PCA on the whitened feature vector centers to compute a discriminant feature vector of dimension dl_1.

(5) For a probe face, retrieve the top *N* individuals from the gallery using the dl_1 discriminant features.

(6) Using only the top *N* class centers, re-compute dl_2 discriminant features, i.e. repeat step-4 for the *N* classes.

(7) Re-rank the top *N* individuals using the dl_2 new features.

This algorithm has three major improvements over traditional subspace methods. First, it provides a new parameter space to improve recognition performance. It controls $\widetilde{I}$, $\widetilde{T}$ and $\widetilde{N}$ components in the image difference by adjusting the dimensionality of the three subspaces, the PCA subspace (*dp*), intrapersonal subspace (*di*), and discriminant subspace (*dl*). The interaction of the three parameters greatly affects the system performance. Using each of the three subspace dimensions as a parameter axis, the algorithm provides a three-dimensional parameter space, as shown in Fig. 4.

The original PCA, LDA, and Bayes methods only occupy some local lines or areas in the 3D parameter space. PCA changes parameters in the *dp* direction on line *AD*. DIFS and DFFS of the Bayesian algorithm change on the line *DEF* in the *di* direction. Fisher Face [1] corresponds to point *B* (*dp=di=M-L*, *dl=L-1*) in the graph. All these methods change parameters only in the local regions. However, for our new algorithm, optimal parameters may be searched in the full 3D space. We can clearly see the advantage of this in the experiments.

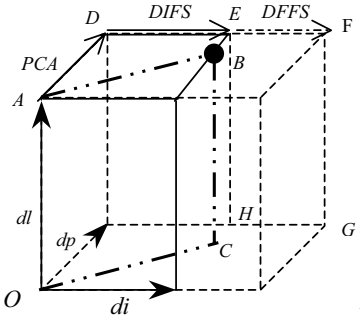


Figure 4. 3D parameter space.

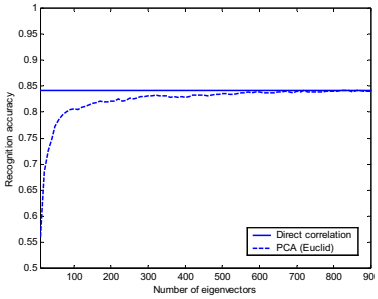


Figure 5. Recognition accuracy of the PCA method on the FERET database.

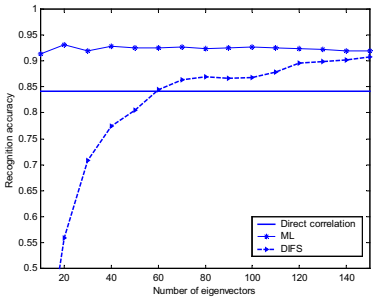


Figure 6. Recognition accuracy of the Bayesian algorithm on the FERET database.

The second improvement of the algorithm is the adoption of different training data at different steps of the training process according to the special requirement of the step. In traditional method, the same training data is used throughout the algorithm. The conflict requirements of each step limit the optimization ability of the algorithm. For example, in LDA, S_w and S_b come from the same training data. If only the individuals in the gallery are selected for training, the samples for each class may be too few to estimate the transformation difference $\tilde{T}$, since sometimes there is only one sample for each individual in the gallery. However, if we add to the training set with many more individuals outside the gallery, S_b may be too distracted to extract optimal features targeting the discrimination of the individuals in the gallery.

In order to accomodate this conflicting requirement, we use different training set for different steps. For the intrapersonal subspace estimation (step 2) we use an enlarged intrapersonal difference set that contains individuals both inside and outside of the gallery to effectively estimate $\tilde{T}$. Then for the discriminant analysis step (step-3,4), we only use the class centers of the individuals in the gallery, so that the features extracted are specifically tuned for the individuals in the gallery.

The third improvement of the algorithm is the design of a two-step approach to solve the large class number problem. In the first step, we first retrieve the top N individuals most similar to the probe face. This is a significant reduction of class number. In the second step, we re-compute the discriminate features using only the top N class centers. Unlike the features used in step-1, which are computed using the whole gallery, these new features are more closely related to the probe face since they are computed from faces that are very similar to the probe, thus should be more effective in discriminating this group of similar faces. In addition, we only need to classify N individuals instead of L using the new features. It is much easier to seek for the intrinsic difference for N classes than L classes. Re-computing the discriminant features only needs to solve an N by N matrix. The cost is minimal since N is very small ($N << L$).

5. Experiment

In this section, we conduct experiments on face images of 1195 people selected from the FERET face database with two images for each person. Images of 495 people are used for training and the remaining 700 people are used for testing. So there are 990 face images in the training set, 700 face image in the gallery, and 700 face images for probe. All the images are normalized by the eye locations. A mask template is used to remove the background and the hair. Histogram equalization is applied to the face images for photometric normalization.

5.1 PCA Experiment

The recognition accuracy of the PCA method using different eigenspace dimension (*dp*) is shown in Fig. 5. The accuracy of direct correlation is 84.1%. We use the direct correlation as a benchmark since it is essentially a direct use of image difference without subspace analysis. When *dp* is small, the PCA result is worse than direct correlation. As *dp* increases, it steadily approaches the benchmark. The results show that PCA is no better than direct correlation in terms of accuracy. Even though PCA can effectively reduce subspace dimension through removing noise $\tilde{N}$, it cannot decouple $\tilde{I}$ and $\tilde{T}$ to improve recognition accuracy.

5.2 Bayesian Experiment

Experimental results for the Bayesian algorithm are reported in Fig. 6. It has achieved around 10% improvement over direct correlation, and is stable even for a small feature number. When only 20 features are selected, the accuracy of PCA is less than 70%, while the ML measure achieves 93% accuracy. When only a small number of eigenvectors are selected, the principal subspace does not have enough information on $\tilde{I}$, so the accuracy of DIFS is low (below 60% for 20 eigenvectors). However, the lost information can be compensated from DFFS in the complementary subspace.

Table 2. Recognition accuracy of Bayesian analysis in the reduced PCA space.

	Euclid	*dp*	DIFS (*di*)									
			10	20	50	100	150	200	250	300	400	490
PCA	0.773	50	0.277	0.609	0.937	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	0.807	100	0.271	0.581	0.854	0.954	N/A	N/A	N/A	N/A	N/A	N/A
	0.817	150	0.276	0.573	0.814	0.909	0.960	N/A	N/A	N/A	N/A	N/A
	0.821	200	0.276	0.580	0.813	0.893	0.923	0.953	N/A	N/A	N/A	N/A
	0.831	300	0.271	0.567	0.806	0.879	0.937	0.937	0.944	0.930	N/A	N/A
	0.836	500	0.266	0.563	0.804	0.871	0.907	0.916	0.927	0.931	0.930	0.670
	0.840	700	0.267	0.560	0.803	0.869	0.907	0.920	0.926	0.931	0.927	0.911
	0.840	900	0.266	0.560	0.804	0.869	0.907	0.917	0.926	0.930	0.926	0.909
Bayes on raw data			0.267	0.559	0.804	0.869	0.907	0.919	0.930	0.930	0.926	0.906

So the accuracy of ML is high by combining the two components together.

5.3 Bayesian analysis in reduced PCA space

After comparing the PCA and Bayesian methods individually, we now use a set of experiments to investigate how the two subspace dimensions in our 3D parameter space may interact with each other. We first apply PCA on the raw face vector to reduce the dimensionality and remove the noise. Then the Bayesian analysis is implemented in the reduced PCA space. This corresponds to the *dp*-*di* plane in the 3D space in Fig. 4.

Results are reported in Table 2. The vertical direction is the dimensionality of PCA space (*dp*) and the horizontal direction is the dimensionality of intrapersonal space (*di*). The *dp*-*di* accuracy surface is also plotted in Fig. 7. There are two benchmark curves in the 3D space of Fig. 7. One is traditional PCA accuracy curve as reported in the second column in Table 2. This can be used to evaluate the improvement of Bayesian analysis. The second curve is the DIFS curve of the standard Bayesian algorithm based on raw face vectors. It is reported in the bottom row of Table 2. We will compare it with DIFS curves in different PCA spaces. The maximum for *di* is $\min\{d_p, 495\}$.

The shape of *dp*-*di* accuracy surface clearly reflects the effect of noise. When *dp* is small, there is little noise in the PCA subspace. So the recognition accuracy monotonically increases with *di* as more discriminating information $\tilde{I}$ is added, and finally reaches the highest point at the full dimensionality of the intrapersonal subspace. However, as *dp* increases, noise begins to appear in the PCA subspace. The curve starts to decrease after reaching a peak point before *di* reaches the full dimensionality. The decrease in accuracy at the end of the curve is because noise on the small eigenvectors is magnified by the inverse of the small eigenvalues.

This effect of noise is especially severe when both *dp* and *di* are around 495, i.e. the largest possible *di*. In this region, the accuracy becomes as low as 67%. Because of the large *dp*, noise has become a fairly significant problem. When *di* becomes the same size as *dp*, all the energy in the PCA subspace, including noise, are selected for the Bayesian analysis. Noise concentrated on the last few very small eigenvectors will be drastically magnified because of the very small eigenvalues.

We plot the highest accuracy of each accuracy curve of different *dp* in Fig. 8. The maximum point with 96% accuracy could be found at (*dp*=150, *di*=150).

5.4 Extract discriminant features from intrapersonal subspace

We now investigate the effect of the third dimension *dl* in the 3D parameter space. For ease of comparison, we choose three representative points on the *dp*-*di* surface, and report the accuracy along the dimension of *dl* as shown in Fig. 9. The curves first increase to a maximum point and then drop with further increase of *dl*. For traditional LDA, the *dl* dimension is usually chosen as $L-1$, which corresponds to the last point of the curve with *di* = 494. The result is clearly much lower than the highest accuracy in the Fig. 9. As discussed in Section 3, this dimension mainly serves to compact $\tilde{I}$ and remove more noise $\tilde{N}$, so the dimensionality should be reasonably small instead of being fixed by *L*. The best results on the curses are indeed better than using the first two dimensions only.

As shown by these experiments, although we have not explored the entire 3D parameter space, better results are already found comparing to the standard subspace methods. A careful investigation of the whole parameter space should lead to further improvement.

5.6 Unified subspace analysis

We now test the unified subspace analysis algorithm using the 495 individuals to compute intrapersonal subspace, and the 700 individuals in the gallery to compute the between class scatter matrix. With *dp*=*di*=*150*, $dl_1 = dl_2$ =*TopN-1*, the recognition accuracies using different number of discriminant features are even better. We also notice that it can achieve a relatively high accuracy using a very small number of discriminate

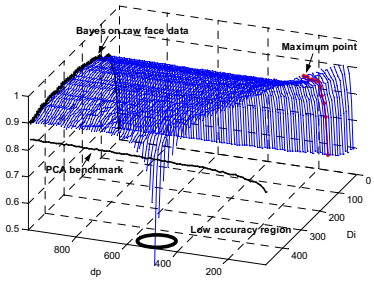


Figure 7. Accuracy curves for Bayesian analysis in PCA space.

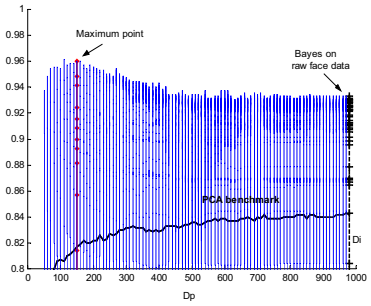


Figure 8. Highest accuracy of Bayes analysis in each PCA space.

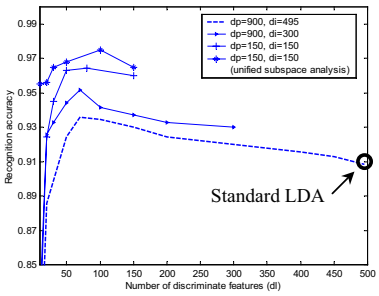


Figure 9. Accuracies using different number of discriminant features extracted from intrapersonal subspace.

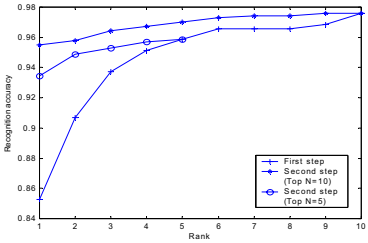


Figure 10. Recognition accuracy of the unified subspace analysis using 10 discriminant features.

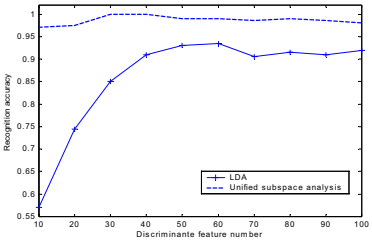


Figure 11. Compare recognition accuracy of the unified subspace analysis with LDA on the testing set containing 100 individuals.

features. As shown in Fig. 10, using 10 features the first round recognition can achieve only 83.1% accuracy for the first rank. Selecting top 10 classes, after re-computing the discriminant features, the second round of recognition improve the first rank accuracy to 95.5%. This shows that the new features are much more efficient in discriminating the top 10 classes than the features computed from 700 class centers.

To further demonstrate the effectiveness of the unified subspace analysis, we construct another data set using the FERET database. We use 100 people for testing. For each individual there are two face images in gallery, and another two taken in another session for probe. Even though the data size is much smaller, the recognition for images of different session is usually much more difficult than recognition of the same session data. Using the 200 images in the gallery as training data, the LDA method only achieves an accuracy of 93%, since there are not enough training samples to accurately estimate the intrapersonal subspace. Using 668 face images that include the 200 images in the gallery but not the images in the probe set to estimate the intrapersonal subspace, the unified subspace analysis method achieves 100% accuracy using only a small number of features.

6. Summary

Starting from a new face difference model, we develop a unified framework for subspace analysis. Using this framework we discover how each subspace method contributes to the extraction of discriminating information in the face difference. This eventually leads to the construction of a 3D parameter space that use three subspace dimensions as axis. Within this parameter space, we develop a unified subspace analysis method that achieves much better recognition performance than the standard subspace methods.

ACKNOWLEDGMENT

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

Reference

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

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

[3] B. Moghaddam, T. Jebara, and A. Pentland, “Bayesian Face Recognition”, *Pattern Recognition,* Vol. 33, pp. 1771-1782, 2000.

[4] P. J. Phillips, H. Moon, and S. A. Rozvi, “The FERET Evaluation Methodolody for Face Recognition Algorithms”, *IEEE Trans. PAMI*, Vol. 22, No. 10, pp. 1090-1104, Oct. 2000.

[5] M. Turk and A. Pentland, "Eigenfaces for Recognition", *J. of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86, 1991.

[6] W. Zhao, R. Chellappa, and P. Phillips. “Face Recognition: A Literature Survey”, *Technical Repot*, 2002.

[7] W. Zhao, R. Chellapa, and P. Philips, “Subspace Linear Discriminant Analysis for Face Recognition”, *Technical Report CAR-TR-914,* 1996.