Subspace Distance Analysis with Application to Adaptive Bayesian Algorithm for Face Recognition

Liwei Wang, Xiao Wang, Jufu Feng
Center for Information Sciences
School of Electronics Engineering and Computer Sciences, Peking University
Beijing, 100871, China
Tel: 86-10-62754861; Fax: 86-10-62755569
{wanglw, wangxiao, fjf}@cis.pku.edu.cn

Abstract
We propose subspace distance measures to analyze the similarity between intrapersonal face subspaces, which characterize the variations between face images of the same individual. We call the conventional intrapersonal subspace Average Intrapersonal Subspace (AIS) because the image differences often come from a large number of persons. An intrapersonal subspace is referred to as Specific Intrapersonal Subspace (SIS) if the image differences are from just one person. We demonstrate that SIS varies significantly from person to person, and most SISs are not similar to AIS. Based on these observations, we introduce the maximum a posteriori (MAP) adaptation to the problem of SIS estimation, and apply it to the Bayesian face recognition algorithm. Experimental results show that the adaptive Bayesian algorithm outperforms the non-adaptive Bayesian algorithm as well as Eigenface and Fisherface methods if a small number of adaptation images are available.

Keywords: Face recognition, Intrapersonal subspace, Bayesian face recognition, Subspace distance, Maximum a Posteriori Adaptation

1 Introduction
Interest and research activities in subspace methods for human face recognition have increased significantly over the last decade. A pioneer work is the Karhunen-Loeve transform of faces [1], which led to the Principal Component Analysis (PCA) [2] “Eigenface” technique [3]. In fact, the state-of-the-art in face recognition, to a large extent, characterized by a family of subspace methods: The Eigenface, Fisherface [4] and Bayesian algorithms [5] have by now become common performance benchmarks in the field. These methods assume that the face images lie in a linear manifold in the pixel space. A remarkable theoretical result supporting this idea was proposed by Basri and Jacobs [6]. They showed that a low dimensional linear subspace could capture the set of images of an object produced by a variety of lighting conditions. Hence, images of a face with fixed pose but under varying illuminations do constitute a low dimensional linear manifold in pixel space.

Among the existing subspace techniques, the Bayesian algorithm achieves high recognition accuracy (for performance benchmarks, please see [7]). Bayesian algorithm adopts a probabilistic measure of similarity based on a Bayesian (MAP) analysis of face differences. This similarity measure characterizes what kind of image variation is typical for the same person and what is
for different persons. Such information is crucial for face recognition because intrapersonal variations are sometimes larger (in terms of Euclidean distance) than extrapersonal variations due to, for instance the different lighting conditions and facial expressions. The Bayesian algorithm classifies the face difference $\Delta = I_1 - I_2$ as intrapersonal variation $\Omega_I$ for the same individual and extrapersonal variation $\Omega_E$ for different individuals. The MAP similarity between two images is defined as the intrapersonal a posteriori probability

$$S(I_1, I_2) = p(\Omega_I | \Delta) = \frac{p(\Delta | \Omega_I)p(\Omega_I)}{p(\Delta | \Omega_I)p(\Omega_I) + p(\Delta | \Omega_E)p(\Omega_E)} \quad (1)$$

A main contribution of the Bayesian algorithm is that it develops an efficient method to estimate the likelihood density $p(\Delta | \Omega_I)$ (and also $p(\Delta | \Omega_E)$) in high-dimensional space. It decomposes the pixel space into intrapersonal principal subspace $F$ and its orthogonal complementary space $\bar{F}$, where $F$ is spanned by the principal components of all intrapersonal image differences, i.e. $\{\Delta | \Delta \in \Omega_I\}$. The likelihood can be estimated as

$$\hat{p}(\Delta | \Omega_I) = \left[ \exp\left( -\frac{1}{2} \sum_{i=1}^{m} \frac{y_i^2}{\lambda_i} \right) \right] \cdot \left[ \exp\left( -\frac{\varepsilon^2(\Delta)}{2\rho} \right) \right] \cdot \left[ \prod_{i=1}^{m} \lambda_i^{1/2} \right] \cdot \left[ \frac{(2\pi)^{m/2} \prod_{i=1}^{m} \lambda_i^{1/2}}{(2\pi)^{(n-m)/2}} \right] \quad (2)$$

where $y_i$ is the principal component of $\Delta$ projecting to the $i$th intrapersonal eigenvector, and $\lambda_i$ is the corresponding eigenvalue. $\varepsilon^2(\Delta)$ is the PCA residual error in $\bar{F}$. $\rho$ is the average eigenvalue in $\bar{F}$. $p(\Delta | \Omega_E)$ can be estimated in a similar way. The intrapersonal subspace is of particular importance. First, it plays a dominant role in the Bayesian algorithm. Experimental results have shown that maximizing the intrapersonal likelihood $p(\Delta | \Omega_I)$ alone (called ML measure) is almost as effective as the MAP measure [5]. Second, intrapersonal subspace is closely related to the LDA subspace [8], and the Bayesian algorithm can be viewed as a nonlinear generalization of the Fisherface method.

Although the intrapersonal subspace represents the variations between images from the same individual, it is more appropriate to call it average intrapersonal subspace (AIS). Because the intrapersonal differences $\Delta \in \Omega_I$ often come from a large number of persons. In this paper, we would also like to consider specific intrapersonal subspace (SIS). That is, we extract principal components from intrapersonal differences $\Delta$ that come from one person. Thus AIS is, in a sense, the average of all SISs.

Two questions naturally arise: Are two SISs similar to each other? And are SISs similar to the AIS? To study these similarity problems, we define distance measures for subspaces. The subspace distance is the first contribution of this paper. In fact, the proposed distance is quite general and can be used in a variety of applications (see Section 1.1 for details). Based on the subspace distance, we perform a series of experiments. It turns out that SIS varies significantly from person to person and the average model AIS could not well represent all SISs. These results imply that AIS is a rather coarse model of the intrapersonal image variation.

The above intrapersonal subspace analyses immediately find applications in face recognition algorithms. Mismatch between the average model and the person specific model is a common phenomenon in many pattern recognition tasks. A widely used technique to handle this problem is adaptation. Adaptation combines a priori knowledge in the average model and, usually a small amount of person specific adaptation data to estimate an accurate person dependent model. Adaptation has been quite successful in speech recognition [9]. It significantly improves the recognition rate for outlier speakers not well represented in the average model. Adaptation is often posed in a Bayesian (MAP) estimation framework [10], which requires a prior distribution...
of the person specific model as well as the conditional distribution of the adaptation data.

In this paper, we introduce the MAP adaptation to the Bayesian face recognition algorithm. We adaptively estimate SIS of each person in the gallery, and use SISs instead of the AIS to compute the probabilistic similarity (1). As stated above, we need to define a prior distribution of the SIS and a conditional distribution of the adaptation data. The difficulty lies in two aspects: One is the prior distribution. Note that the random object here is subspace, different from the standard statistical scenario in which the random object is variables. The other is the implementation of the MAP adaptation of SIS. The second contribution of this paper is that the presented subspace distance naturally induces a prior distribution, and we show that the MAP estimation of SIS is an eigenvalue problem. Experimental results demonstrate that the adaptive Bayesian algorithm outperforms the non-adaptive Bayesian algorithm as well as the Eigenface and Fisherface methods if a small amount of adaptation images are available.

1.1 Related Work

Recently, there are growing interest in design and analysis of similarity/distance measures over subspaces [11, 12, 13]. However, the researches are focused on a slightly different context the video based recognition. For instance, face recognition in video sequences; discrimination of irregular motion trajectory of an individual in a video and so on. In these tasks, a video sequence contains a set of images each represented by a vector. These vectors then span a subspace in the image space. Therefore, subspace similarity/distance is a natural measure to compare two video sequences.

Almost all subspace similarities proposed in this context are based on the principal angles. A recursive definition of the principal angles due to Hotelling [14] is as follows: The principal angles $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k \leq (\pi/2)$ between two subspaces $U$ and $V$ are uniquely defined as

$$
cos(\theta_k) = \max_{u \in U} \max_{v \in V} u^T v.
$$

subject to

$$
u^T u = v^T v = 1,
$$

$$
u^T u_i = v^T v_i = 0, \quad i = 1, 2, \cdots, k - 1.
$$

Yamaguchi et al. [13] were the first to use principal angles as a measure for matching two video sequences. However, only the smallest principal angle is considered in their similarity. Wolf and Shashua [12] adopted $\prod_{k=1}^{k} \cos^2(\theta_i)$ as a subspace similarity measure, and make use of the kernel trick to generalize it for nonlinear subspaces. They elegantly proved that this similarity is a positive definite kernel and therefore can be plugged into the Support Vector Machines (SVMs). One limitation in their work is that they require the two subspaces must have the same dimensionality.

In spite of their success in video based recognition, these similarities are less useful for our problem. On one hand they are restricted for subspaces of the same dimensionality, whereas the SIS and AIS are computed by PCA, which usually results in different dimensionalities. On the other hand, these similarities do not induce prior distribution of SIS suitable for the MAP adaptation. This will be clear in later sections.

2 Subspace Distance

We define distance measures between two linear subspaces. They help us to study the similarity of two SISs and that between an SIS and the AIS. The subspace distance will play
an important role in the prior distribution of SIS and make the MAP adaptation an eigenvalue problem. However, the discussion in this section is general and the subspace distance may have other applications such as the video based recognition. We also consider the distance between two nonlinear surfaces. In particular, we use the kernel trick, for which the surfaces are the pre-images of linear subspaces in high-dimensional feature space.

2.1 Linear Subspaces

Although a subspace can be seen as a set of points, common distance measures of sets are not appropriate for subspaces. For example, the minimum distance between two point sets \( A \) and \( B \) in Euclidean space, defined as

\[
d(A, B) = \min_{a \in A, b \in B} \|a - b\|
\]

is always zero when \( A \) and \( B \) are subspaces, since the origin belongs to all subspaces. Another commonly used distance, the Hausdorff distance

\[
(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|
\]

is infinity if \( A \neq B \), due to the unboundedness of linear subspace. Intuitively, the distance/similarity of two subspaces should reflect the difference between their directions. That is, if two subspaces nearly coincide, they should have a small distance; if they are almost perpendicular to each other, they have a large distance. These intuitive ideas will be incorporated in our definition of subspace distance.

To begin with, consider two linear subspaces \( U \) and \( V \) in \( \mathbb{R}^d \). We first assume that \( U \) and \( V \) have the same dimensionality, say \( m \) \((m \leq d)\). Let \( u_1, u_2, \cdots, u_m \) be an orthonormal basis of \( U \). Let \( d(u_i, V) \) denote the so-called \( L_2 \)-Hausdorff distance from the end point of vector \( u_i \) to subspace \( V \). That is

\[
d(u_i, V) = \min_{v \in V} \|u_i - v\|. \tag{3}
\]

where \( \|\cdot\| \) is the Euclidean norm. We then define the subspace distance.

**Definition 1** The subspace distance \( d(U, V) \) for two \( m \)-dimensional subspaces \( U \) and \( V \) is defined as

\[
d(U, V) = \sqrt{\sum_{i=1}^{m} d^2(u_i, V)}. \tag{4}
\]

The simplest example is that both \( U \) and \( V \) are one-dimensional subspaces, i.e. vectors. In this case, \( d(U, V) = \sin \theta \), where \( \theta \) is the angle between the two vectors.

It is necessary to check some desired properties of this subspace distance. First of all, in Definition 1, we employ a particular orthonormal basis \( u_1, u_2, \cdots, u_m \) of \( U \). However the distance is invariant to the orthonormal basis.

**Theorem 1** The subspace distance defined above is invariant to the choice of orthonormal basis.

**Proof** Let \( u_1, u_2, \cdots, u_m \) and \( \tilde{u}_1, \tilde{u}_2, \cdots, \tilde{u}_m \) be two arbitrary orthonormal bases of \( U \). It suffices to show that

\[
\sum_{i=1}^{m} d^2(u_i, V) = \sum_{i=1}^{m} d^2(\tilde{u}_i, V).
\]
Let \( v_1, v_2, \cdots, v_m \) be an orthonormal basis of \( V \). Then
\[
d^2(u_i, V)
\]
can be written as
\[
d^2(u_i, V) = ||u_i||^2 - \sum_{j=1}^{m} (u_i^Tv_j)^2
\]
and
\[
\sum_{i=1}^{m} d^2(u_i, V) = m - \sum_{i=1}^{m} \sum_{j=1}^{m} (u_i^Tv_j)^2.
\] (5)

Thus, to complete the proof, we need only to show that
\[
\sum_{i=1}^{m} \sum_{j=1}^{m} (u_i^Tv_j)^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} (\tilde{u}_i^Tv_j)^2.
\]

In fact, we can show that, for every \( j, j = 1, 2, \cdots, m \),
\[
\sum_{i=1}^{m} (u_i^Tv_j)^2 = \sum_{i=1}^{m} (\tilde{u}_i^Tv_j)^2.
\]

To see that the above equality holds, applying Parseval’s theorem to the vector \( P_Uv_j \), where \( P_Uv_j \) is the projection of \( v_j \) onto subspace \( U \). And this completes the proof.

In addition, by Eqs. (4) and (5), it is easy to see that the subspace distance has the following properties.

1. Nonnegativity: \( d(U, V) \geq 0 \), and \( d(U, V) = 0 \) if and only if \( U = V \);
2. Symmetry: \( d(U, V) = d(V, U) \);
3. Upper boundedness: \( d(U, V) \leq \sqrt{m} \), and \( d(U, V) = \sqrt{m} \) if and only if \( U \perp V \). That is, each vector in \( U \) is perpendicular to \( V \) and vice versa.

1. is, each vector in \( U \) is perpendicular to \( V \) and vice versa. This distance measure agrees with the afore-mentioned intuitive idea that it reflects the difference between the directions of the subspaces.

The definition of subspace distance can be generalized to the case in which \( U \) and \( V \) have different dimensions. Let \( U \) and \( V \) be \( m \) and \( n \)-dimensional subspaces, respectively. Also let \( u_1, u_2, \cdots, u_m \) and \( v_1, v_2, \cdots, v_n \) be orthonormal bases of \( U \) and \( V \) respectively.

**Definition 2** The directional distance from \( m \)-dimensional subspace \( U \) to \( n \)-dimensional subspace \( V \) is defined as
\[ d(U, V) = \sqrt{\sum_{i=1}^{m} d^2(u_i, V)} \]

\[ = \sqrt{m - \sum_{i=1}^{m} \sum_{j=1}^{n} (u_i^T v_j)^2}. \]  

Clearly, the directional distance is not symmetry. We then define the symmetric distance.

**Definition 3** The symmetric distance between \( m \)-dimensional subspace \( U \) and \( n \)-dimensional subspace \( V \) is defined as

\[ d(U, V) = \max(d(U, V), d(V, U)) \]

\[ = \sqrt{\max(m, n) - \sum_{i=1}^{m} \sum_{j=1}^{n} (u_i^T v_j)^2}. \]

It is easy to check that this distance is also invariant to the choice of orthonormal basis, and \( d(U, V) \leq \sqrt{\max(m, n)} \).

The symmetric subspace distance will be employed as a measure to study the similarity between two SISs and that between an SIS and the AIS. Two subspaces are said to be similar if and only if

\[ \frac{d(U, V)}{\sqrt{\max(m, n)}} \leq \frac{1}{2}. \]

The experimental results are given in Section 4, where we show that SIS varies from person to person, and most SISs are not similar to AIS.

### 2.2 Nonlinear Generalization

In spite of the theoretical result that lighting variational subspace is almost linear, there are cases, e.g. pose and expression, that the variation induces a nonlinear subspace. In this section, we design and analyze distance measures that are nonlinear generalizations of the results in the previous section. In particular, we are interested in linear subspaces in a high (possibly infinite) dimensional feature space, which are nonlinearly related to the input variables. In fact, such subspaces, known as the kernel PCA [15] subspaces, have been widely applied in face recognition systems as generalizations of the linear subspace methods, and often achieve better performance [16].

A crucial observation of our previous results is that all the presented subspace distances can be expressed solely in terms of inner products (see Eqs. (4), (5), (6) and (7)). This makes our definitions of subspace distance immediately generalize to the nonlinear case using the kernel trick.

Let \( x_1, x_2, \ldots, x_r \) and \( y_1, y_2, \ldots, y_s \) be two sets of vectors in the input space \( \mathbb{R}^N \). We then consider their principal subspaces \( U \) and \( V \) in a very high dimensional feature space \( H \), which is related to \( \mathbb{R}^N \) by a nonlinear mapping \( \Psi: \mathbb{R}^N \rightarrow H \). The mapping \( \Psi \) is made implicit by the use of a Mercer kernel [17] \( K \):

\[ K(x_i, x_j) = \Psi(x_i) \ast \Psi(x_j), \]

in which the kernel function \( K \) evaluates the inner product in the feature space.
The feature principal subspace $U$ of the feature vectors $\Psi(x_1), \Psi(x_2), \cdots, \Psi(x_r)$ can be computed efficiently by the kernel Principal Component Analysis (KPCA) algorithm [15]. In [15], it is shown that each eigenvector $u_i$ of the covariance matrix of $\Psi(x_1), \Psi(x_2), \cdots, \Psi(x_r)$ is a linear combination of the $r$ feature vectors

$$u_i = \sum_{k=1}^{r} \gamma_{ik} \Psi(x_k).$$

$i = 1, 2, \cdots, r$. The KPCA algorithm takes $x_1, x_2 \cdots, x_r$ as inputs and returns all the coefficients $\gamma_{ik}$, $i, k = 1, 2, \cdots, r$. The leading eigenvectors $u_i$s constitute an orthonormal basis of $U$. In like manner, let

$$v_i = \sum_{k=1}^{s} \omega_{ik} \Psi(y_k)$$

$i = 1, 2, \cdots, s$, be the eigenvectors of $\Psi(y_1), \Psi(y_2), \cdots, \Psi(y_s)$. Then the leading $v_i$s form an orthonormal basis of $V$. We can then compute the symmetric distance between the two feature subspaces $U$ and $V$ as

$$d(U, V) = \sqrt{\max(m, n) - \sum_{i=1}^{m} \sum_{j=1}^{n} (u_i^T v_j)^2}$$

$$= \sqrt{\max(m, n) - \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \sum_{k=1}^{r} \sum_{k'=1}^{s} \gamma_{ik} \omega_{jk'} \Psi(x_k) \ast \Psi(y_{k'}) \right)^2}$$

$$= \sqrt{\max(m, n) - \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \sum_{k=1}^{r} \sum_{k'=1}^{s} \gamma_{ik} \omega_{jk'} K(x_k, y_{k'}) \right)^2}.$$ (9)

where $m, n$ are dimensions of $U$ and $V$ respectively.

In Section 4, we will employ two typical Mercer kernels (Gaussian and polynomial) to measure the similarity between the feature variation subspaces. The experimental results show that there are little similarity between SIS and the AIS.

3 Application: MAP Adaptation of SIS

The dissimilarity between most SISs and AIS implies that the average model AIS does not well represent each persons variation subspace. Thus one may expect an improved performance if we can accurately estimate SISs for all individuals in the gallery, and use SISs instead of a single AIS in the Bayesian face recognition algorithm. However, such an accurate estimation often needs a large number of face images of each person, which are not available in most applications. Adaptation is a powerful tool to solve this problem. It promises to present a good estimate of the person specific model but requires only a small amount of data.

Adaptation is usually accomplished using a maximum a posteriori (MAP) approach. MAP adaptation involves the use of prior knowledge about the model parameter distribution. Hence, if we know what the parameters of the person-specific model likely to be (before observing any adaptation data) using the prior knowledge, we might well be able to make good use of the limited adaptation data to obtain a decent MAP estimate. In most MAP adaptation problems, the prior used are the average models parameters, which is often referred to as an informative prior.
Now consider the SIS adaptation. We will estimate the SIS of an individual $X$. Let $V$ be the SIS (to be estimated) of $X$. Also let $\Delta_1, \Delta_2, \ldots, \Delta_k$ be the adaptation data, where $\Delta_1, \Delta_2, \ldots, \Delta_k$ are face differences of $X$ obtained from $X$’s images in the gallery. For the purpose of MAP adaptation, we shall define a prior distribution of $V$ which we denote by $p(V)$, and the conditional distribution of the face differences $p(\Delta_1, \Delta_2, \ldots, \Delta_k|V)$. The MAP estimation of $V$ is then

$$\hat{V}_{MAP} = \arg \max_V p(V|\Delta_1, \Delta_2, \ldots, \Delta_k) = \arg \max_V p(\Delta_1, \Delta_2, \ldots, \Delta_k|V) \cdot p(V).$$

(10)

**Definition 4** The prior distribution of an SIS $V$ is defined as

$$p(V) = \frac{1}{Z} \exp\{-d^2(V,V_0)\}.$$  

(11)

where $d(\cdot, \cdot)$ is the symmetric subspace distance proposed in Section 2, $V_0$ is the average subspace AIS, and $Z$ is a normalizing factor.

The meaning of this definition is clear: Person-specific subspace is distributed around the average subspace, and here AIS plays the role of an informative prior.

Before giving the definition of the conditional distribution of image differences $p(\Delta_1, \Delta_2, \ldots, \Delta_k|V)$, we consider an extreme case of the MAP adaptation: There is no prior knowledge of $V$. In this situation, MAP estimation (10) reduces to ML estimation

$$\hat{V}_{ML} = \arg \max_V p(\Delta_1, \Delta_2, \ldots, \Delta_k|V).$$

(12)

That is, given face differences $\Delta_1, \Delta_2, \ldots, \Delta_k$, compute the intrapersonal subspace without any prior knowledge. Note that PCA deals with exactly such a problem, and the Bayesian face recognition algorithm uses PCA to compute the AIS. For this reason, we would like to define the conditional distribution $p(\Delta_1, \Delta_2, \ldots, \Delta_k|V)$ in accordance with PCA. It is well known that PCA presents a subspace, which is the least square approximation of the data. That is, given $\Delta_1, \Delta_2, \ldots, \Delta_k$, the principal subspace with a fixed dimension is given by

$$\hat{V}_{PCA} = \arg \min_V \sum_{i=1}^k d^2(\Delta_i, V).$$

(13)

where

$$d(\Delta_i, V)$$

has been given in (3). Accordingly, we make the following definition.

**Definition 5** The conditional distribution is defined as

$$p(\Delta_1, \Delta_2, \ldots, \Delta_k|V) = \frac{1}{Z'} \exp\{-\sum_{i=1}^k d^2(\Delta_i, V)\}.$$  

(14)

where $Z'$ is also a normalizing factor.
Hence, if no prior knowledge is available, the ML estimation is equivalent to PCA. Combining Eqs. (10), (11) and (14), the MAP adaptation of SIS $V$, given the AIS $V_0$, is

$$
\hat{V}_{MAP} = \arg \max_V \frac{\exp(-d^2(V,V_0))}{Z} \cdot \frac{\exp(-\sum_{i=1}^{k} d^2(\Delta_i, V))}{Z'}
$$

(15)

From the computational perspective, MAP is essentially an optimization problem. Hence the existence of an efficient algorithm is arguably the most important issue for MAP adaptation. The presented subspace distance makes the adaptation of SIS very easy to implement. In fact, we can show that it is an eigenvalue problem.

**Theorem 2** Let $w_1, w_2, \ldots, w_n$ be an orthonormal basis of the $n$-dimensional AIS $V_0$. Also let $\Delta_1, \Delta_2, \ldots, \Delta_k$ be the adaptation data. Then the MAP estimation of an $m$-dimensional SIS $V$ is an eigenvalue problem. More precisely, the first $m$ eigenvectors (corresponding to the largest $m$ eigenvalues) of the following matrix $M$

$$
M = \sum_{i=1}^{k} \Delta_i \Delta_i^T + \sum_{i=1}^{n} w_i w_i^T
$$

(16)

constitute an orthonormal basis of $V$.

**Proof** By Eq. (15) and the definition of symmetric subspace distance, the MAP estimation of an $m$-dimensional SIS $V$ is to solve the following optimization problem:

$$
\max_{v_1, v_2, \ldots, v_m} \sum_{i=1}^{m} \sum_{j=1}^{n} (v_i^T w_j)^2 + \sum_{i=1}^{m} \sum_{j=1}^{k} (v_i^T \Delta_j)^2.
$$

(17)

subject to the orthogonal condition:

$$
v_i^T v_j = \begin{cases} 
1 & i = j, i, j = 1, 2, \ldots, m. \\
0 & i \neq j.
\end{cases}
$$

The optimal $v_1, v_2, \ldots, v_m$ constitute an orthonormal basis of $V$. Translate the above formulae into the matrix expressions, the optimization problem is of the form

$$
\max_{v_1, v_2, \ldots, v_m} \sum_{i=1}^{m} v_i^T M v_i
$$

(18)

s.t. $v_i^T v_j = \begin{cases} 
1 & i = j, i, j = 1, 2, \ldots, m. \\
0 & i \neq j.
\end{cases}$

where the matrix $M$ has been given in Eq. (16). Using the standard argument for deriving the Karhunen-Loeve transformation (see for example [18, Ch. 9] and [2, Ch. 2]), it is readily seen that the maximum is achieved when $v_1, v_2, \ldots, v_m$ are the $m$ leading eigenvectors of $M$. The proof is then completed.

Using MAP adaptation, we may have an accurate estimate of the SIS and the corresponding eigenvalues of each individual in the gallery. These SISs will substitute for the AIS in the Bayesian face algorithm to improve the performance. We call this method adaptive Bayesian algorithm.
4 Experimental Result

The experiments in this section are divided into two sets. The primary goal of the first set of experiments is to study, with the presented subspace distances, the similarity between any two SISs and that between SIS and AIS. The objective of the second group of experiments is to compare the performance of the adaptive Bayesian face recognition algorithm to the original non-adaptive version as well as other typical methods.

4.1 Subspace Similarity

The experimental data used here are all from the Yale database. Each individual in this database contains images in 9 poses and various lighting conditions. In the preprocessing procedure, all images are normalized by the eye for scaling, translation, and rotation, so that the eye centers are in fixed positions. Then a mask template is used to remove the hair and the background.

To study the similarity between two SISs, we randomly select two individuals, and randomly picking out 30 images of them each. We then adopt the PCA (KPCA) algorithm to compute for each person the (feature) principal subspace that contains the 95% of the total energy. We next calculate \( d(U, V) / \sqrt{\max(m, n)} \) for the two SISs. This process is repeated for 500 times, so that we can plot the histograms of \( d(U, V) / \sqrt{\max(m, n)} \) for pairs of SISs.

The histograms are shown in Fig. 1, where in (a) we use the linear subspace distance, and in (b) use the feature subspace distance implicitly determined by the Gaussian kernel, and in (c) by a polynomial kernel. Recall that 1/2 is defined to be the upper bound of similarity. The results illustrate that almost all SISs are not similar to each other.

We next study the similarity between AIS and SIS. We compute the AIS using all the images in the database and then calculate the subspace distance between the AIS and the SISs obtained above. The histogram obtained with the three subspace distances are plotted in Fig. 2 (a), (b) and (c) respectively. There exists a large gap between the AIS and the SIS.

4.2 Performance Assessment of the Adaptive Bayesian Algorithm

We conduct two experiments to compare the performance of the adaptive Bayesian algorithm to the original non-adaptive version as well as the Eigenface and Fisherface methods.

In the first experiment, we test the behavior of the algorithm when the prior knowledge is far from ideal. To be concrete, the average model \( V_0 \) (see Section 3) and the models of the other three methods are trained using images from the FERET database, in which the intrapersonal variation are mostly caused by the changes in expression. The gallery and the probe set however, consist of images of Yale database, in which the variation are lighting condition and pose. Accordingly, the prior \( V_0 \) poorly represents the intrapersonal variation of the test images. For MAP adaptation, we exploit images in the gallery and \( V_0 \) to estimate the SIS for each individual. Then the Bayesian similarity is calculated using SIS instead of the AIS. We demonstrate how the performance of our algorithm depends on the number of adaptation data. To make the results reliable, we randomly construct gallery and repeat it ten times to obtain an average recognition rate. The result is plotted in Fig. 3 (a). The adaptive Bayesian algorithm outperforms all other methods when more than four adaptation images are available.

We next assess the performance of the adaptive Bayesian algorithm when the prior knowledge is relatively accurate. We randomly divide the individuals in Yale database in two sets. The prior \( V_0 \) is trained using images of the individuals in one set. Gallery and probes are constructed using images of the other set. Since all data come from the same database, \( V_0 \) is a good prior
subspace. The recognition rates are plotted in Fig. 3 (b). In this case, the adaptive Bayesian algorithm outperforms the other three methods when only two adaptation images are available.

5 Conclusion

We propose subspace distance measures to study the similarity of intrapersonal subspaces. Experimental results show that AIS does not well represent SISs. We then apply the MAP adaptation to the estimation of intrapersonal subspaces. Experimental results demonstrated that the adaptive Bayesian algorithm achieves a higher accuracy than the non-adaptive Bayesian algorithm as well as Eigenface and Fisherface if only a few images are available.

Besides the presented application in Bayesian face recognition algorithm, the subspace distance can also be used in other tasks such as the video based recognition, in which a video sequence is naturally represented with a subspace spanned by the frames images. However, whether this subspace distance satisfies the triangle inequality is still open, i.e.

\[ d(U, V) + d(V, W) \geq d(U, W). \]  \hspace{1cm} (19)

for all subspaces \( U, V, W \). We can show that it holds in two special cases. The first case is that the three subspaces are all of one dimension and lie in the same plane. That is, they are plane vectors. It is easily shown that Eq. (19) is exactly the triangle inequality in the plane geometry. The second case is \( U \perp V \), in which the right hand side of Eq. (19) achieves the maximum. By Eq. (7), it is easy to show

\[ d^2(U, V) + d^2(V, W) \geq d^2(U, W). \]

Therefore

\[ d(U, V) + d(V, W) \geq \sqrt{d^2(U, V) + d^2(V, W)} \]

\[ \geq d(U, W). \]

We believe that the subspace distance always satisfies the triangle inequality. And

\[ d(U, V) + d(V, W) = d(U, W) \]

if and only if \( U = V \) or \( V = W \).

Acknowledgement

This work is supported by National Key Basic Research Project of China (NKBPRC 2004C B318000).

References


**About the Author** LIWEI WANG received the B.S., M.S. degree in Electronic Engineering from Tsinghua University in 1999 and 2002 respectively and the Ph.D. degree in Mathematics from Peking University in 2005. He is currently an assistant professor in the Center for Information Sciences and National Laboratory on Machine Perception, School of Electronics Engineering and Computer Science, Peking University. His interests include pattern recognition and machine
learning.

About the Author XIAO WANG received the B.S. degree in Mathematics and the M.S. degree in School of Electronics Engineering and Computer Science from Peking University. Her research interest is human face recognition.

About the Author JUFU FENG received the B.S. degree in 1989 and Ph.D. degree in 1997 both in mathematics from Peking University, Beijing, P. R. China. Since 1992, he has been with the Center for Information Science and National Laboratory on Machine Perception at Peking University. His current research interests include image processing, pattern recognition and biometrics. He is currently a professor in the Center for Information Science and National Laboratory on Machine Perception, School of Electronics Engineering and Computer Science, Peking University.
Figure 1: Histograms of distances between SIS and SIS using (a) linear subspace distance, (b) Gaussian kernel subspace distance, and (c) Polynomial kernel subspace distance.
Figure 2: Histograms of distances between SIS and AIS using (a) linear subspace distance, (b) Gaussian kernel subspace distance, and (c) Polynomial kernel subspace distance.
Figure 3: Performance comparison of adaptive Bayesian algorithm with non-adaptive Bayesian algorithm, Eigenface and Fisherface Methods.