Intrapersonal Subspace Analysis with Application to Adaptive Bayesian Face Recognition

Liwei Wang, Xiao Wang, Jufu Feng
Center for Information Sciences, Peking University, Peking, China, 100871
Tel: 86-10-62756823; Fax: 86-10-62755569
{wanglw, wangxiao, fjf}@cis.pku.edu.cn

Abstract

We propose a subspace distance measure 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. We call an intrapersonal subspace specific intrapersonal subspace (SIS) if the image differences are from just one person. We demonstrate that SIS varies 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 only a small number of adaptation images are available.

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

1. Introduction

Subspace analysis has attracted much attention in the area of face recognition over the last decade. Eigenface, Fisherface and Bayesian algorithm [1] 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 [2]. 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 a good performance. It develops a probabilistic measure of similarity based on a Bayesian (MAP) analysis of face differences. 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 $p(\Delta | \Omega_I)$ in high-dimensional space. It decomposes the pixel space into intrapersonal principal subspace $F$ and its orthogonal complementary space $\tilde{F}$, 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
\[
\hat{p}(\Delta \mid \Omega_y) = \frac{\exp\left(-\frac{1}{2} \sum_{i=1}^{m} \frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{m/2} \prod_{i=1}^{m} \lambda_i^{1/2}} \exp\left(-\frac{\varepsilon^2(\Delta)}{2\rho}\right)
\]

where \( y_i \), \( i = 1, 2, \cdots, m \) are principal components and \( \varepsilon^2(\Delta) \) is the residual. \( p(\Delta \mid \Omega_y) \) can be estimated in a similar way. The intrapersonal subspace plays a dominant role in Bayesian algorithm. It has been shown that maximizing the intrapersonal likelihood \( p(\Delta \mid \Omega_y) \) alone (called ML measure) is almost as effective as the MAP measure.

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 \( \Delta \in \Omega_y \) 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 a distance measure for linear subspaces. Experimental results show 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 average model and 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 specific model. Adaptation has been quite successful in speech recognition [3]. 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 [4]. It 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). We define a prior distribution of the SIS and a conditional distribution of the adaptation data given the SIS. It can be shown 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 only a small amount of adaptation images are available.

2. Intrapersonal Subspace Analysis

We will define a distance measure between two linear subspaces. It helps us to study the similarity of two SISs and that between an SIS and the AIS. The reason that we define a distance measure instead of using a similarity measure (e.g. the principal angles [5, p.349]) is that, this distance later plays an important role in the prior distribution of SIS in the MAP adaptation algorithm (see Section 3).

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 \( d(A, B) = \min_{a \in A} \min_{b \in B} \|a - b\| \), is always zero when \( A \) and \( B \) are subspaces, since origin belongs to
all subspaces. Another commonly used distance, the Hausdorff distance: $d(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$ 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 $u_1, u_2, \cdots, u_m$ and $v_1, v_2, \cdots, v_m$ be standard orthogonal bases of $U$ and $V$ respectively. 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

$$d(u_i, V) = \min_{v \in V} \|u_i - v\|.$$  \hfill (3)

We then define the subspace distance $d(U, V)$ for $m$-dimensional subspaces $U$ and $V$ as

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

Since $v_1, v_2, \cdots, v_m$ is a standard orthogonal basis of $V$, it is easy to see that

$$d(U, V) = \sqrt{\sum_{i=1}^{m} \left( \sum_{j=1}^{m} (u_i^T v_j)^2 \right)} = \sqrt{m - \sum_{i=1}^{m} \sum_{j=1}^{m} (u_i^T v_j)^2}. \hfill (5)$$

We need to check some desired properties of this subspace distance. First of all, in the definition, we use a particular standard orthogonal basis $u_1, u_2, \cdots, u_m$, but the distance is invariant to the standard orthogonal basis.

**Theorem 1**

The subspace distance defined above is invariant to the choice of standard orthogonal basis.

**Proof.**

Let $u_1, u_2, \cdots, u_m$ and $\tilde{u}_1, \tilde{u}_2, \cdots, \tilde{u}_m$ be two standard orthogonal bases of $U$. Let $v_1, v_2, \cdots, v_m$ be a standard orthogonal basis of $V$. To prove the theorem, it suffices to show that

$$\sqrt{m - \sum_{i=1}^{m} \sum_{j=1}^{m} (u_i^T v_j)^2} = \sqrt{m - \sum_{i=1}^{m} \sum_{j=1}^{m} (\tilde{u}_i^T v_j)^2}. \hfill (6)$$

or equivalently

$$\sum_{i=1}^{m} \sum_{j=1}^{m} (u_i^T v_j)^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} (\tilde{u}_i^T v_j)^2. \hfill (7)$$

In fact, we can show that, for every $j$, $j = 1, 2, \cdots, m$,

$$\sum_{i=1}^{m} (u_i^T v_j)^2 = \sum_{i=1}^{m} (\tilde{u}_i^T v_j)^2. \hfill (8)$$

To see that the above equality holds, applying Parseval’s theorem to the vector $P_{\perp} v_j$, where $P_{\perp} v_j$ is the projection of $v_j$ onto subspace $U$. And this completes the proof. \hfill \square

Two additional properties of the distance are immediate from (4) and (5): the nonnegativity $d(U, V) \geq 0$, and the symmetry $d(U, V) = d(V, U)$. We also address the minimum and maximum of the distance. By (4) and (5), it is easy to see that $0 \leq d(U, V) \leq \sqrt{m}$; $d(U, V) = 0$ iff $U = V$; and $d(U, V) = \sqrt{m}$ iff $U \perp V$, that is, each vector in $U$ is perpendicular to $V$ and vice versa. The 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, \( u_1, u_2, \ldots, u_m \) and \( v_1, v_2, \ldots, v_n \) be standard orthogonal bases of \( U \) and \( V \) respectively. Define the directional distance from \( U \) to \( V \) as

\[
\bar{d}(U, V) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (u_i^T v_j)^2}.
\]

(9)

Clearly, the directional distance is not symmetry. We define the symmetric distance between \( U \) and \( V \) as

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

(10)

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

We use this subspace distance measure to study the similarity between two SISs and that between an SIS and the AIS. Two subspaces are said to be similar if \( d(U, V) / \sqrt{\max(m, n)} \leq 1/2 \).

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.

3. Application: MAP Adaptation of SIS

The dissimilarity between most SISs and AIS implies that the average model AIS does not well represent each person’s subspace. Thus one may expect a performance improvement if we can estimate accurate 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.

The advantage of adaptation is due to the fact that it combines the prior knowledge of the person specific model (usually contained in the average model) and the adaptation data. Bayesian estimation provides a natural framework for adaptation, which is often called MAP adaptation. In the situation of SIS adaptation, let \( V \) be an SIS. MAP adaptation requires the definition of a prior distribution of \( V \) denoted by \( p(V) \), and the conditional distribution of the face differences \( p(\Delta_1, \Delta_2, \ldots, \Delta_n | V) \). The MAP estimate of \( V \) is

\[
\hat{V}_{\text{MAP}} = \arg \max \ p(V | \Delta_1, \Delta_2, \ldots, \Delta_n) = \arg \max \ p(\Delta_1, \Delta_2, \ldots, \Delta_n | V) \cdot p(V).
\]

(11)

In order to incorporate the \textit{a priori} knowledge contained in the average model (AIS) into the prior distribution \( p(V) \), we define \( p(V) \) as follows. Let \( V_0 \) be the AIS.

\[
p(V) \overset{\text{def}}{=} \frac{1}{Z} \exp(-d^2(V, V_0)).
\]

(12)

where \( Z \) is a normalizing constant, and \( d(V, V_0) \) is the subspace distance defined in Section 2. According to the discussion in the previous section, \( p(V) \) reflects the similarity between SIS \( V \) and AIS \( V_0 \).

Before giving the definition of the conditional distribution of image differences \( p(\Delta_1, \Delta_2, \ldots, \Delta_n | V) \), we consider an extreme case of the MAP adaptation: There is no prior knowledge of \( V \). In this situation, MAP estimation (11) reduces to ML estimation.
\[ \hat{V}_{ML} = \arg \max_{\nu} p(\Delta_1, \Delta_2, \cdots, \Delta_k \mid V). \] (13)

That is, given face differences \( \Delta_1, \Delta_2, \cdots, \Delta_k \), compute the 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 \( p(\Delta_1, \Delta_2, \cdots, \Delta_k \mid V) \) 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 \( \Delta_1, \Delta_2, \cdots, \Delta_k \), the principal subspace with a fixed dimension is given by
\[ \hat{V}_{PCA} = \arg \min_{\nu} \sum_{i=1}^{k} d^2(\Delta_i, V). \] (14)

where \( d(\Delta, V) \) was defined in (3). Accordingly, we define the conditional distribution as
\[ p(\Delta_1, \Delta_2, \cdots, \Delta_k \mid V) = \frac{1}{Z'} \exp(-\sum_{i=1}^{k} d^2(\Delta_i, V)). \] (15)

where \( Z' \) is also a normalizing constant. Hence, if no prior knowledge is available, the ML estimation is PCA.

Combining (12) and (15), the MAP adaptation of SIS \( V \), given the AIS \( V_0 \) is
\[ \hat{V}_{MAP} = \arg \max_{\nu} \exp\left(-d^2(V, V_0)\right) \frac{\exp(-\sum_{i=1}^{k} d^2(\Delta_i, V))}{Z'} \] 
\[ = \arg \min_{\nu} [d^2(V, V_0) + \sum_{i=1}^{k} d^2(\Delta_i, V)]. \] (16)

It can be shown that the Bayesian adaptation (16) is to solve the eigenvalue problem of matrix \( M \):
\[ M = \sum_{i=1}^{k} \Delta_i A_i^T + \sum_{i=1}^{n} w_i w_i^T. \] (17)

where \( w_i, i = 1, 2, \cdots, n \), are base vectors of \( V_0 \). If we fix the dimensionality of SIS to be \( m \), \( \hat{V}_{MAP} \) is the principal subspace spanned by the first \( m \) eigenvectors of \( M \).

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 Results

The goal of the first experiment is to study the similarity between any two SISs and between SIS and AIS. All face images are from Yale database. We compute an AIS and a number of SISs by PCA on image differences. We next calculate the normalized subspace distance \( d(U, V)\sqrt{\max(m, n)} \) for all SIS pairs and SIS-AIS pairs. The histograms of \( d(U, V)\sqrt{\max(m, n)} \) of the two kinds are depicted in Fig. 1. 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 and most SISs are not similar to AIS.

The purpose of the second experiment is to compare the performance of the adaptive Bayesian algorithm to non-adaptive Bayesian algorithm as well as the Eigenface, Fisherface methods. All the models including the AIS are trained by images from FERET database. Gallery and Probe set contain images all from Yale database. In adaptation, we use images in the gallery and the AIS to compute SISs for all persons. Then the Bayesian similarities are calculated using the SIS of each person 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. 2. The adaptive Bayesian algorithm outperforms all other methods when more than four adaptation images are available.
5. Conclusion

We propose a subspace distance measure 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. It is demonstrated that the adaptive Bayesian algorithm achieves a higher accuracy than the non-adaptive Bayesian algorithm as well as Eigenface and Fisherface when a few images are available.

Acknowledgement

This work is supported by the National Natural Science Foundation of China 60175004.

References