Why Recognition in a Statistics-based Face Recognition System Should be based on the Pure Face Portion: a Probabilistic Decision-based Proof*

Li-Fen Chen§ Hong-Yuan Mark Liao‡† Ja-Chen Lin§

Chin-Chuan Han‡

§Department of Computer and Information Science,
National Chiao Tung University, Taiwan
‡Institute of Information Science, Academia Sinica, Taiwan
TEL:+886-2-27883799 ext.1811 FAX:+886-2-27824814
E-mail: liao@iis.sinica.edu.tw

Abstract

It is evident that the process of face recognition, by definition, should be based on the content of a face. The problem is: what is a "face"? Recently, a state-of-the-art statistics-based face recognition system, the PCA plus LDA approach, has been proposed [1]. However, the authors used "face" images that included hair, shoulders, face and background. Our intuition tells us that only a recognition process based on a "pure" face portion can be called face recognition. The mixture of irrelevant data may result in an incorrect set of decision boundaries. In this paper, we propose a statistics-based technique to quantitatively prove our assertion. For the purpose of evaluating how the

*This work was supported by the National Science Council under grant no. NSC87-2213-E-001-025.
†To whom correspondence should be sent.

different portions of a face image will influence the recognition results, a hypothesis testing model is proposed. We then implement the above mentioned face recognition system and use the proposed hypothesis testing model to evaluate the system. Experimental results show that the influence of the "real" face portion is much less than that of the nonface portion. This outcome confirms quantitatively that recognition in a statistics-based face recognition system should be based solely on the "pure" face portion.

Keywords: *statistics-based face recognition, face-only database, hypothesis testing*

1 Introduction

Face recognition has been a very popular research topic in recent years [2–6]. It covers a wide variety of application domains, including security systems, personal identification, image and film processing, and human-computer interaction. A complete face recognition system should include two stages. The first stage is detecting the location and size of a "face", which is difficult and complicated because of the unknown position, orientation and scaling of faces in an arbitrary image [7–14]. The second stage of a face recognition system involves recognizing the target faces obtained in the first stage. In order to design a good face recognition system, the features chosen for recognition play a crucial role. In the literature [15–21], two main approaches to feature extraction have been extensively used. The first one is based on extracting structural facial features that are local structures of face images, for example, the shapes of the eyes, nose, and mouth. The structure-based approaches are not affected by irrelevant data, such as hair or background, because they deal with local data instead of global data. On the other hand, the statistics-based approaches extract features from the whole image. Since the global data of an image are used to determine the set of decision boundaries, data which are irrelevant to facial portions should be disregarded. Otherwise, these irrelevant portions may contribute to the decision boundary determination process and later mislead the recognition results. From the psychological viewpoint, Hay and Young [22] pointed out that the internal facial features, such as the eyes, nose, and mouth, are very important for human beings to see and to recognize familiar faces. However, it was also pointed out in [23] that, in statistics-based systems, if face images in the database cover the face, hair, shoulder,

and background, the "facial" portion will not play a key role during the execution of 'face' recognition. In [24], Bruce et al. compared two of the most successful systems, one proposed by Pentland et al. based on PCA [15, 25] and the other proposed by von der Malsburg [26, 27] based on graph matching. They indicated that the PCA system gave higher correlations to the rating obtained with hair than did the graph matching system.

In recent years, many researchers have noticed this problem and tried to exclude those irrelevant "nonface" portions while performing face recognition. In [2], Belhumeur et al. eliminated the nonface portion of face images with dark backgrounds. Similarly, Goudail et al. [28] constructed face databases under constrained conditions, such as asking people to wear dark jackets and to sit in front of a dark background. In [15], Turk and Pentland multiplied the input face image by a two-dimensional Gaussian window centered on the face to diminish the effect caused by the nonface portion. For the same purpose, Sung et al. [9] tried to eliminate the near-boundary pixels of a normalized face image by using a fixed-size mask. Moghaddam and Pentland [12] and Lin et al. [17] both used probabilistics-based face detectors to extract facial features or cut out the middle portion of a face image for correct recognition. In [23], Liao et al. proposed a face-only database as the basis for face recognition. All the above mentioned works tried to use the most "correct" information for the face recognition task. Besides, the works in Ref. [2, 23] also conducted some related experiments to show that if the database contains "full face" images, changing the background or hair style may decrease recognition rate significantly. However, they only tried to explain the phenomena observed from their experiments, but a quantitative measure was not introduced to support their assertion. In a statistics-based face recognition system, global information (pixel level) is used to determine the set of decision boundaries and to perform recognition. Therefore, a mixture of irrelevant data may result in an incorrect set of decision boundaries. The question is: can we measure, quantitatively, the influence of the irrelevant data on the face recognition result? In this paper, we shall use a statistics-based technique to perform this task.

In order to conduct the experiments, two different face databases were adopted. One was a training database built under constrained environments. The other was a synthesized face database which contained a set of synthesized face images. Every synthesized face image consisted of two parts: one was the middle face portion that includes the eyes, nose, and

mouth of a face image. The other portion was the complement of the middle face, called the "nonface" portion, of another face image. We will show in details how to construct these two face databases in the following sections. Based on these two databases, the distances between the distribution of the original training images and that of the synthesized images could be calculated. For the purpose of evaluating how the different portions of a face image influence the recognition result, a hypothesis testing model was employed. We then implemented a state-of-the-art face recognition system and used the proposed hypothesis testing model to evaluate the system. Experimental results obtained from the system show that the influence of the middle face portion on the recognition process is much less than that of the nonface portion. This outcome is important because it proves, quantitatively or statistically, that recognition in statistics-based face recognition systems should be based on pure-face databases.

The organization of this paper is as follows. In Section 2, a state-of-the-art face recognition system which will be examined in this paper is introduced. Descriptions of the proposed hypothesis testing model and experimental results are given in Sections 3 and 4, respectively. Conclusions are drawn in Section 5.

2 State-of-the-art: PCA plus LDA Face Recognition

In this section, a state-of-the-art face recognition system, which was implemented and used in the experiments, will be introduced. Swets and Weng [1] first proposed principal component analysis(PCA) plus linear discriminant analysis(LDA) for face recognition. They applied the PCA technique to reduce the dimensionality of the original image. In their work, the top 15 principal axes were selected and used to derive a 15-dimensional feature vector for every sample. These transformed samples were then used as bases to execute LDA. In other words, their approach can be decomposed into two processes, the PCA process followed by the LDA process. All the details can be found in [1]. They reported a peak recognition rate of more than 90%. Recently, Belhumeur et al. [2] and Zhao et al. [3] have proposed systems which use similar methodology and the former one is named "Fisherfaces". The methodology adopted in the above mentioned approaches is efficient and correct. However, for a statistics-based face recognition system like that in [1], we would like to point out that the database used in their

system is incorrect. According to [1], the face images used in their system contained face, hair, shoulders, and background, not solely face. We wonder whether inclusion of irrelevant "facial" portions, such as hair, shoulders, and background, will generate incorrect decision boundaries for recognition. Therefore, in this paper, we shall answer this question based on results obtained using statistical methods. Since the method proposed in [1] combined the PCA and LDA techniques to decide on the projection axes for the recognition purpose, we shall briefly introduce the PCA and LDA approaches, respectively, in the following paragraphs.

Principal component analysis (PCA) finds a set of the most expressive projection vectors such that the projected samples retain the most information about the original samples. The most expressive vectors derived from a PCA process are the eigenvectors corresponding to the leading largest eigenvalues of the total scatter matrix, $S_t = \sum_{i=1}^{N}(\boldsymbol{s}_i - \boldsymbol{m})(\boldsymbol{s}_i - \boldsymbol{m})^t$, [29], in the form

$$\boldsymbol{t}_i = W^t(\boldsymbol{s}_i - \boldsymbol{m}), \tag{1}$$

where $\boldsymbol{s}_i$ is the ith original sample, $\boldsymbol{m} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{s}_i$ is the total mean vector, and $\boldsymbol{t}_i$ is the projected sample of $\boldsymbol{s}_i$ through W, which is the set of projection column vectors. The corresponding computational algorithm of a PCA process can be found in [1].

In the normal LDA process, one determines the mapping

$$\boldsymbol{v}_m^k = A^t\boldsymbol{u}_m^k, \tag{2}$$

where $\boldsymbol{u}_m^k$ denotes the feature vector extracted from the mth face image of the kth class and $\boldsymbol{v}_m^k$ denotes the projective feature vector of $\boldsymbol{u}_m^k$ under transformation of the mapping matrix A. This mapping simultaneously maximizes the between-class scatter while minimizing the within-class scatter of all $\boldsymbol{v}_m^k$'s (where $k = 1, \ldots, K; m = 1, \ldots, M$) in the projective feature vector space. Here, in the PCA plus LDA approach, $\boldsymbol{u}_m^k$, the input of LDA, is the projective sample obtained from Eq. (1), the output of PCA. Let $\bar{\boldsymbol{v}}^k = \sum_{m=1}^{M}\boldsymbol{v}_m^k$ and $\bar{\boldsymbol{v}} = \sum_{k=1}^{K}\bar{\boldsymbol{v}}^k$. The within-class scatter in the projective feature space can be calculated as follows [30]:

$$S_w = \sum_{k=1}^{K}\sum_{m=1}^{M}(\boldsymbol{v}_m^k - \bar{\boldsymbol{v}}^k)(\boldsymbol{v}_m^k - \bar{\boldsymbol{v}}^k)^t. \tag{3}$$

The between-class scatter in the same space can be calculated as follows:

$$S_b = \sum_{k=1}^{K}(\bar{\boldsymbol{v}}^k - \bar{\boldsymbol{v}})(\bar{\boldsymbol{v}}^k - \bar{\boldsymbol{v}})^t. \tag{4}$$

The way to find the required mapping A is to maximize the following quantity:

$$tr(S_w^{-1}S_b). \tag{5}$$

An algorithm which solves the mapping matrix A can be found in [31]. However, the major drawback of applying LDA is that the within-class scatter matrix S_w in Eq. (5) may be singular when the number of samples is smaller than the dimensionality of samples. Some researchers have proposed different approaches to solve this problem [2, 31, 32]. In the PCA plus LDA approach [2], they first project samples into a reduced dimensional space through the PCA process such that S_w in the following LDA process is guaranteed to be nonsingular. A Euclidean distance classifier can be used to perform classification in the mapped space.

3 Hypothesis Testing Model

We have mentioned in the previous section that inclusion of irrelevant "facial" portions, such as hair, shoulders, and background, will mislead the face recognition process. In this section, we shall propose a statistics-based hypothesis testing model to prove our assertion. Before going further, we shall define some basic notations which will be used later.

Let $\boldsymbol{X}^k = \{\boldsymbol{x}_m^k, m = 1, \ldots, M \mid \boldsymbol{x}_m^k$ is the feature vector extracted from the mth face image of the kth person$\}$ denote the set of feature vectors of the M face images of class ω_k (person k), where $\boldsymbol{x}_m^k$ is a d-dimensional column vector and each class collects M different face images of a person. For simplicity, the M face images of every person are labelled and arranged in order. Each class is then represented by a likelihood function. Without loss of generality, assume that the class likelihood function, $p(\boldsymbol{x}|\omega_k)$, of class ω_k is a normal distribution [33]:

$$p(\boldsymbol{x}|\omega_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Lambda|^{\frac{1}{2}}} exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Lambda^{-1}(\boldsymbol{x}-\boldsymbol{\mu})), \tag{6}$$

where $\boldsymbol{x}$ is a d-dimensional column vector, and $\boldsymbol{\mu}$ and Λ are the mean vector and covariance matrix of $p(\boldsymbol{x}|\omega_k)$, respectively. Some researchers [12, 17, 34] have used this model to describe the face images of the same person (class) and adopted different criteria to estimate the parameters, $\boldsymbol{\mu}$ and Λ. Here, for simplicity, we use the sample mean, $\bar{\boldsymbol{x}}^k = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{x}_m^k$, and the sample covariance matrix, $\Lambda_k = \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{x}_m^k - \bar{\boldsymbol{x}}^k)(\boldsymbol{x}_m^k - \bar{\boldsymbol{x}}^k)^t$, to represent the estimates of $\boldsymbol{\mu}$ and Λ, respectively.

For each vector set $\boldsymbol{X}^k$ of class $\omega_k (k = 1, \ldots, K)$, an additional vector set, $\boldsymbol{Y}_k^l$ ($l = 1, \ldots, K$, $l \neq k$), is extracted and associated with it. The number of elements in $\boldsymbol{Y}_k^l$ (for a specific l) is equal to M, which is exactly the same as the number of elements in $\boldsymbol{X}^k$. The formation of the elements in $\boldsymbol{Y}_k^l$ is as follows. First, we manually point three landmarks on each face image to locate the positions of the eyes and the mouth. According to these landmarks, each face image can be adequately cropped to form an image block of the corresponding middle face. Two examples showing how to construct synthetic face images are shown in Figure 1. Figure 1(a) and (c) show two face images with landmarks. The landmarks on these images are manually located. According to the proportion of the distances between these landmarks, the corresponding middle face portion can be formed as shown in Figure 1(b) and (d), respectively. From Figure 1(b) and (d), it is easy to see that faces of different sizes will be adequately cropped. For constructing a synthetic face image, a normalization process is applied to deal with the issues such as scale, brightness, and boundary. Figure 1(e) shows the synthetic face image which is synthesized from the nonface portion of (a) and the middle face image of (c). Similarly, Figure 1(f) shows the synthetic face image which is synthesized from the nonface portion of (c) and the middle face image of (a). Hence, each element in $\boldsymbol{Y}_k^l$ is a d-dimensional feature vector extracted from a synthesized face image which combines the middle face portion of an element in ω_l and the nonface portion of its corresponding element in ω_k. We have mentioned that the M elements in $\boldsymbol{X}^k$ (extracted from ω_k, $k = 1, \ldots, K$) are arranged in order (from 1 to M). Therefore, the synthesized face image sets as well as the feature sets extracted from them are all arranged in order. The reason why we ordered these images is because here we want to make the synthesized images as real as possible. And it can be done when the images are obtained under constrained environments, such as controlled lighting condition, fixed view orientations, and neutral expression. In sum, for each vector set $\boldsymbol{X}^k$ of class ω_k ($k = 1, \ldots, K$), there are $(K-1)$ synthesized feature sets associated with it. In what follows, we shall provide some formal definitions of the synthesized sets.

Let $\boldsymbol{w}_q^p$ denote the pth face image of class ω_q ($p = 1, \ldots, M$). For $l = 1, \ldots, K$, $l \neq k$, we have the $(K-1)$ feature sets which are associated with $\boldsymbol{X}^k$, defined as follows:

$$\boldsymbol{Y}_k^l = \{\boldsymbol{y}_k^l(m), m = 1, \ldots, M | \ \boldsymbol{y}_k^l(m) \text{ is a } d\text{-dimensional feature vector extracted from a}$$

synthesized face image which combines the middle face portion of $\boldsymbol{w}_l^m$ and the nonface portion of $\boldsymbol{w}_k^m$}. (7)

Figure 2 is a graphical illustration showing how $\boldsymbol{Y}_k^l$ is extracted. One thing to be noted is that when we combine two different portions of images, some rescaling and normalization preprocessings are necessary in order to reduce boundary variations. Figure 3 is a typical example illustrating how the synthesized face image is combined with the middle face portion of an image in ω_l and the nonface portion of its corresponding image in ω_k.

Bichsel and Pentland [16] have shown, from the topological viewpoint, that when a face undergoes changes in its eye width, nose length, and hair style, it is still recognized as a human face. Therefore, it is reasonable to also represent the above mentioned feature vector set, $\boldsymbol{Y}_k^l$, as a normal distribution function. Now, since all the feature vector sets are represented by normal distributions, their distances can only be evaluated by using some specially defined metrics. In the literature [35–38], the Bhattacharyya distance is a well-known metric which is defined for measurement of the similarity (correlation) between two arbitrary statistical distributions. A lower distance between two distributions means higher correlation between them. For two arbitrary distributions $p(\boldsymbol{x}|\omega_1)$ and $p(\boldsymbol{x}|\omega_2)$ of classes ω_1 and ω_2, respectively, the general form of the Bhattacharyya distance is defined as

$$D(\omega_1, \omega_2) = -\ln \int \left(p(\boldsymbol{x}|\omega_1)p(\boldsymbol{x}|\omega_2)\right)^{1/2} d\boldsymbol{x}. \tag{8}$$

When both ω_1 and ω_2 are normal distributions, the Bhattacharyya distance can be simplified into a new form as follows:

$$D(\omega_1, \omega_2) = \frac{1}{8}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \left(\frac{\Lambda_1 + \Lambda_2}{2}\right)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2} \ln \frac{\left|\frac{\Lambda_1 + \Lambda_2}{2}\right|}{(|\Lambda_1||\Lambda_2|)^{1/2}}, \tag{9}$$

where $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Λ_1, Λ_2 are the mean vectors and covariance matrices of ω_1 and ω_2, respectively [33]. In what follows, we shall define a hypothesis testing model for use as a tool in experiments. The Bhattacharyya distance will be used as a decision criterion for determining acceptance or rejection of our hypothesis.

3.1 The Hypothesis Testing Model

In the hypothesis testing, our goal was to prove that the influence of the nonface portions of face images on the recognition result is larger than that of the middle face portions of face images; that is, the nonface portion of a face image dominates the recognition result.

In what follows, we shall define a metric based on the above mentioned Bhattacharyya distance. The metric to be defined for a specific class k is a real-number set, D^k. The definition of D^k is as follows:

$$D^k = \{d^k(l), l = 1, \ldots, K; l \neq k |\ d^k(l) = D(\boldsymbol{X}^k, \boldsymbol{Y}_k^l) - D(\boldsymbol{X}^l, \boldsymbol{Y}_k^l)\}, \tag{10}$$

where $D(\bullet)$ represents the Bhattacharyya distance between two distributions as defined in Equation (9).

For a specific class k, there are in total $K-1$ elements contained in D^k. The physical meaning of every constituent of D^k, i.e., $d^k(l)$ $(l = 1, \ldots, K; l \neq k)$, is a statistical measure that can be used to evaluate the importance, quantitatively, between the middle face portion and the nonface portion. Figure 4 illustrates how $d^k(l)$ is calculated in a graphically illustrative manner. Figure 4(a) shows how the first term that defines $d^k(l)$ is calculated. The top row of Figure 4(a) contains two rectangles, each of which includes a circle region. The rectangle region together with the circle region inside represents a face image. The left hand side combination contains 2 k's. This means that the middle face portion (the circle region) and the nonface portion (the rectangle region excluding the circle region) belong to the same person. The right hand side combination, on the other hand, contains the nonface portion belonging to person k and the middle face portion belonging to person l, respectively. The middle row of Figure 4(a) shows the corresponding feature vectors extracted from the (pure) face image on the left hand side and the synthesized face image on the right hand side, respectively. Both assemblages of $\boldsymbol{x}_m^k$ and $\boldsymbol{y}_k^l(m)$ contain, respectively, M elements. The bottom rows of Figure 4(a) and (b) represent, respectively, the difference between the two distributions, which can be computed using the Bhattacharyya distance as defined in Equation (9). In what follows, we shall report how the degree of importance between the middle face portion and the nonface portion can be determined based on the value of $d^k(l)$.

From Equation (10), it is obvious that when $d^k(l) \geq 0$, the distribution of $\boldsymbol{Y}_k^l$ is closer to

that of $\boldsymbol{X}^l$ than to that of $\boldsymbol{X}^k$. Otherwise, the distribution of $\boldsymbol{Y}_k^l$ is closer to that of $\boldsymbol{X}^k$ than to that of $\boldsymbol{X}^l$. According to the definition of face recognition, the recognition process should be dominated by the middle face portion. In other words, the normal situation should result in a $d^k(l)$ which has a value not less than zero. If, unfortunately, the result turns out to be $d^k(l) < 0$, then this means that the nonface portion dominates the face recognition process. We have mentioned that for a specific class k, there are in total $K-1$ possible synthesized face image sets. Therefore, we shall have $K-1$ $d^k(l)$ values (for $l = 1, \ldots, K, l \neq k$). From the statistical viewpoint, if more than half of these $d^k(l)$ values are less than zero, then this means that the face recognition process regarding person k is dominated by the nonface portion. The formal definition of the test values for person k is as follows:

$$\begin{aligned} \bar{H}^k &: \; p(d^k(l) \geq 0; d^k(l) \in D^k) \;\geq\; 0.5, \\ H^k &: \; p(d^k(l) \geq 0; d^k(l) \in D^k) \;<\; 0.5, \end{aligned} \tag{11}$$

where $\bar{H}^k$ represents the null hypothesis, H^k stands for the alternative hypothesis, and $p(\bullet)$ here represents the probability decided under a predefined criterion $\bullet$. According to the definition of D^k, it contains $K-1$ $d^k(l)$ real values. Therefore, the rules defined in Equation (11) will let the null hypothesis $\bar{H}^k$ be accepted whenever the amount of $d^k(l)$ which has a value not less than zero is more than one half of $K-1$; otherwise, the alternative hypothesis H^k will be accepted.

The rules described in Equation (11) are only for a specific class k. If they are extended to the whole population, a global hypothesis test rule is required. The extension is trivial and can be written as follows:

$$\begin{aligned} \bar{H} &: \; p(\bar{H}^k \text{ is accepted}, \, k = 1, \ldots, K) \;\geq\; 0.5, \\ H &: \; p(\bar{H}^k \text{ is accepted}, \, k = 1, \ldots, K) \;<\; 0.5. \end{aligned} \tag{12}$$

The physical meaning of the rules described in Equation (12) is that when over half of the population passes the null hypothesis, the global null hypothesis $\bar{H}$ is accepted; otherwise, the global alternative hypothesis will be accepted. When the latter occurs, this means that the nonface portion of a face image dominates the face recognition process among the majority of the whole population.

4 Experimental Results

Before showing our experimental results, we will first introduce a series of experiments conducted in [18]. Liao et al. conducted a series of experiments using the synthesized images Y_k^l. From the results, they found that in the PCA plus LDA approach, the nonface portion dominated the whole recognition process. It is obvious that the results shown in Figure 5 indicate that the face portion did not play a key role in the recognition process. The similar experiment has been conducted in [2]. Belhumeur et al. partitioned the images into two different scales: one included the full face and part of the background while the other one was cropped and included only internal facial structures such as brow, eyes, nose, mouth, and chin. They found that the recognition rate using a full-face database is much better than that using a closely cropped database; however, if the background or hair style of full face images have been varied, the recognition rate would have been much lower and even worse than that using closely cropped images. These exciting results encourage us to make a formal (quantitative) proof of the problem.

In the experiments described below, the statistics-based state-of-the-art face recognition system proposed by Swets and Weng [1] was implemented and tested against the proposed hypothesis testing model. The training database contained 90 persons (classes), and each class contained 30 different face images of the same person. The 30 face images of each class were labelled and ordered according to the orientations in which they were obtained. These orientations included ten frontal views, ten frontal views with 15 degrees to the right, and ten frontal views with 15 degrees to the left. The process for collecting facial images was as follows: after asking the persons to sit down in front of a CCD camera, with neutral expression and slightly head moving in three different orientations, a 30-second period for each orientation was recorded on videotape under well-controlled lighting condition. Later, a frame grabber was used to grab 10 image frames for each orientation from the videotape and stored them with resolution of 155×175 pixels. Since these images were obtained under the same conditions, the synthesized images used in our hypothesis testing would look very similar to real images visually. For the PCA plus LDA approach proposed by Swets and Weng [1], each projective feature vector extracted from a face image is 15-dimensional. Based on these feature vectors of

training samples, the proposed hypothesis model was tested. Since the projection axes derived through linear discriminant analysis were ordered according to their discriminating capability, the first projection axis was most discriminating followed by the second projection axis. For the convenience of visualization, all the samples were projected onto the first two projection axes and are shown in Figure 6 for the proposed hypothesis model.

Figure 6 shows the three related distributions covered in D^k . 'o' and 'x' represent $\boldsymbol{X}^k$ of person k and $\boldsymbol{X}^l$ of person l, respectively, and '+' represents $\boldsymbol{Y}_k^l$, whose element combines the middle face portion of person l and the nonface portion of person k. The distributions of $\boldsymbol{X}^k$, $\boldsymbol{X}^l$, and $\boldsymbol{Y}_k^l$ all covered 30 elements (2-dimensional vectors). Each distribution was enclosed by an ellipse, which was drawn based on the distribution's scaled variance on each dimension. Therefore, most of the feature vectors belonging to the same class were enclosed in the same ellipse. In Figure 6, it is obvious that the distribution of $\boldsymbol{Y}_k^l$ is closer to that of $\boldsymbol{X}^k$. This means that the nonface portions of the set of face images dominated the distribution of the projective feature vector set. That is, the distribution of $\boldsymbol{Y}_k^l$ was completely disjointed from that of $\boldsymbol{X}^l$ and almost completely overlapped that of $\boldsymbol{X}^k$. From the view of classification, each element in $\boldsymbol{Y}_k^l$ would be classified into class k, the one which contains the nonface portion of the test image. In sum, the results of experiments shown in Figure 6 confirm that the nonface portion of a face image did dominate the distributions of the 2-dimensional projective feature vectors.

Figure 7 show the experimental results obtained by applying the proposed hypothesis testing model. In this case, k was set to 1. That is, l ranged from 2 to 90 (horizontal axis). The 'o' sign shown in Figure 7(a) represents the Bhattacharyya distance (vertical axis) between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$, which is the first term of $d^k(l)$. The '+' sign shown in Figure 7(a), on the other hand, represents the Bhattacharyya distance (vertical axis, too) between $\boldsymbol{X}^l$ and $\boldsymbol{Y}_k^l$ and is the second term of $d^k(l)$. The results shown in Figure 7(a) reflect that from l=2 to 90, the second term ('+') of $d^k(l)$ was always larger than its first term ('o'). Therefore, we can say that for $k = 1$ (class 1), the probability that the first term of $d^k(l)$ $(l = 2, \ldots, 90)$ was larger than the second term of $d^k(l)$ is zero. This means that the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$ was always smaller than the distance between $\boldsymbol{X}^l$ and $\boldsymbol{Y}_k^l$ for $k = 1$, $l = 2, \ldots, 90$. One thing worth noticing is that the PCA plus LDA approach had the ability to extract

very "discriminating" projection axes since the distributions, $\boldsymbol{X}^l$ and $\boldsymbol{X}^k$, of different persons were far away. Therefore, the phenomenon whereby the nonface portion dominated the face recognition process was very apparent in the results obtained by using the PCA plus LDA approach. This conclusion is confirmed by the individual probability values shown in Figure 7(b). Figure 7(b) shows, from class 1 to class 90, the individual probability that the first term of $d^k(l)$ $(l = 2, \ldots, 90)$ was larger than the second term of $d^k(l)$. From this figure, it is obvious that most of the individual probabilities (ranging from 1 to 90) were zero. Only a few individual probabilities had values very close to zero (less than 0.05). From the data shown in Figure 7(b), we can draw a conclusion that all the individual null hypotheses $\bar{H}^k$'s $(k = 1, \ldots, 90)$ were rejected, and that the probability of accepting $\bar{H}^k$ $(k = 1, \ldots, 90)$ was equal to zero. Moreover, since $p(\bar{H}^k$ is accepted, $k = 1, \ldots, K) = 0$, the global alternative hypothesis H is accepted. That means for the whole population in this database, the nonface portions of a face image, including hair, shoulder and background, dominate the face recognition process. A possible reason for the above phenomenon is that the number of pixels of the background could be more than the number of pixels of the face portion in the synthesized images. (Therefore, with a PCA-based algorithm, it is not unfair to expect that the synthesized face could match the background better than the face.)

5 Conclusion

In this paper, we have proposed a statistics-based technique to quantitatively prove that the previously proposed face recognition system used "incorrect" databases. According to the definition of face recognition, the recognition process should be dominated by the "pure" face portion. However, after implementing a state-of-the-art statistics-based face recognition system based on PCA plus LDA, we have shown, quantitatively, that the influence of the middle face portion on the recognition process in their system was much smaller than that of the nonface portion. That is, the nonface portion of a face image dominated the recognition result. This outcome is very important because it proves, quantitatively or statistically, that some of the previous statistics-based face recognition systems have used "incorrect" face databases. This outcome also reminds us that if we adopt databases established by other people, a pre-

processing stage has to be introduced. The purpose of the preprocessing stage is to guarantee the correct use of a face database.

References

[1] D. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kiregman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[3] W. Zhao, R. Chellappa, and A. Kirshnaswamy, "Discriminant analysis of principal components for face recognition", in *Proceedings of the third Conference on Automatic Face and Gesture Recognition*, 1998, pp. 336–340.

[4] R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey", *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, 1995.

[5] D. Valentin, H. Abdi, A. O'Toole, and G. Cottrell, "Connectionist models of face processing: A survey", *Pattern Recognition*, vol. 27, no. 9, pp. 1209–1230, 1994.

[6] A. Samal and P. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey", *Pattern Recognition*, vol. 25, no. 1, pp. 65–77, 1992.

[7] S. A. Sirohey, "Human face segmentation and identification", Master's thesis, University of Maryland, 1993.

[8] G. Yang and T. S. Huang, "Human face detection in a complex backgroud", *Pattern Recognition*, vol. 27, no. 1, pp. 53–63, 1994.

[9] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection", A.I. Memo 1521, M.I.T., 1994.

[10] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object detection", *Proceedings of the 5th IEEE Conference on Computer Vision*, pp. 786–793, 1995.

[11] P. Juell and R. Marsh, "A hierarchical neural network for human face detection", *Pattern Recognition*, vol. 29, no. 5, pp. 781–787, 1996.

[12] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, July 1997.

[13] S. H. Jeng, H. Y. Mark Liao, C. C. Han, M. Y. Chern, and Y. T. Liu, "Facial feature detection using geometrical face model:an efficient approach", *Pattern Recognition*, vol. 31, no. 3, pp. 273–282, 1998.

[14] C. C. Han, H. Y. Mark Liao, G. J. Yu, and L. H. Chen, "Fast face detection via morphology-based pre-processing", to appear in *Pattern Recognition*, 1999.

[15] M. Turk and A Pentland, "Eigenfaces for recognition", *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[16] M. Bichsel and A. P. Pentland, "Human face recognition and the face image set's topology", *CVGIP: Image Understanding*, vol. 59, no. 2, pp. 254–261, 1994.

[17] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face recognition/detection by probabilistic decision-based neural network", *IEEE Trans. on Neural Networks*, vol. 8, no. 1, pp. 114–132, 1997.

[18] H. Y. Mark Liao, C. C. Han, G. J. Yu, H. R. Tyan, M. C. Chen, and L. H. Chen, "Face recognition using a face-only database: A new approach", *Proceedings of the 3rd Asian Conference on Computer Vision,* Hong Kong. Lecture Notes in Computer Science, vol. 1352, pp. 742–749, Jan. 1998.

[19] *Proceedings of the second Internation Conference on Automatic Face and Gesture Recognition*, 1996.

[20] *Proceedings of the third Internation Conference on Automatic Face and Gesture Recognition*, 1998.

[21] "Theme section - face and gesture recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, 1997.

[22] D. Hay and A. W. Young, *Normality and Pathology in Cognitive Functions*, chapter The Human Face, Academic Press, 1982.

[23] H. Y. Mark Liao, C. C. Han, and G. J. Yu, "Face + hair + shoulders + background $\neq$ face", in *Proceedings of Workshop on 3D Computer Vision '97*, The Chinese University of Hong Kong, pp.91-96, 1997 (Invited paper.).

[24] V. Bruce, P. J. B. Hancock, and A. M. Burton, "Comparisons between human and computer recognition of faces", *Proceedings of the third Internatonal Conference on Automatic Face and Gesture Recognition*, pp. 408–413, Apr, 1998.

[25] A. Pentland, B. Moghaddam, and T. Starber, "View-basd and modular eigenspaces for face recognition", *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 84–91, 1994.

[26] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lage, C. von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture", *IEEE Trans. on Computers*, vol. 42, pp. 300–311, 1994.

[27] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775–779, 1997.

[28] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, and N. Otsu, "Face recognition system using local autocorrelations and multiscale integration", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 1024–1028, 1996.

[29] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.

[30] R. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches*, John Wiley, 1992.

[31] K. Liu, Y. Cheng, and J. Yang, "Algebraic feature extraction for image recognition based on an optimal discriminant criterion", *Pattern Recognition*, vol. 26, no. 6, pp. 903–911, 1993.

[32] L. F. Chen, H. Y. M. Liao, J. C. Lin, M. D. Kao, and G. J. Yu, "A new lda-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, to appear, 1999.

[33] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic, New York, 1990.

[34] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 388–444, April 1993.

[35] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions", *Bull. Calcutta Math. Soc.*, pp. 99–110, 1943.

[36] X. Tnag and W. K. Stewart, "Texture classification using principal component analysis techniques", *Proceedings of SPIE - Int. Soc. Opt. Eng.*, vol. 2315, no. 13, pp. 22–35, 1994.

[37] G. Xuan, P. Chai, and M. Wu, "Bhattacharyya distance feature selection", *Proceedings of the 13th International Conference on Pattern Recognition*, vol. 2, pp. 195–199, 1996.

[38] C. Lee and D. Hong, "Feature extraction using the bhattacharyya distance", *IEEE International Conference on Systems, man, and Cybernetics. Computational Cybernetics and Simulation*, vol. 3, pp. 2147–50, 1997.

[39] L. F. Chen, H. Y. M. Liao, C. C. Han, and J. C. Lin, "Why a statistics-based face recognition system should base its recognition on the pure face portion: A probabilistic decision-based proof", *Proc. 1998 Symposium on Image, Speech, Signal Processing, and Robotics, The Chinese University of Hong Kong*, pp. 225–230, September 3-4, 1998 (invited).

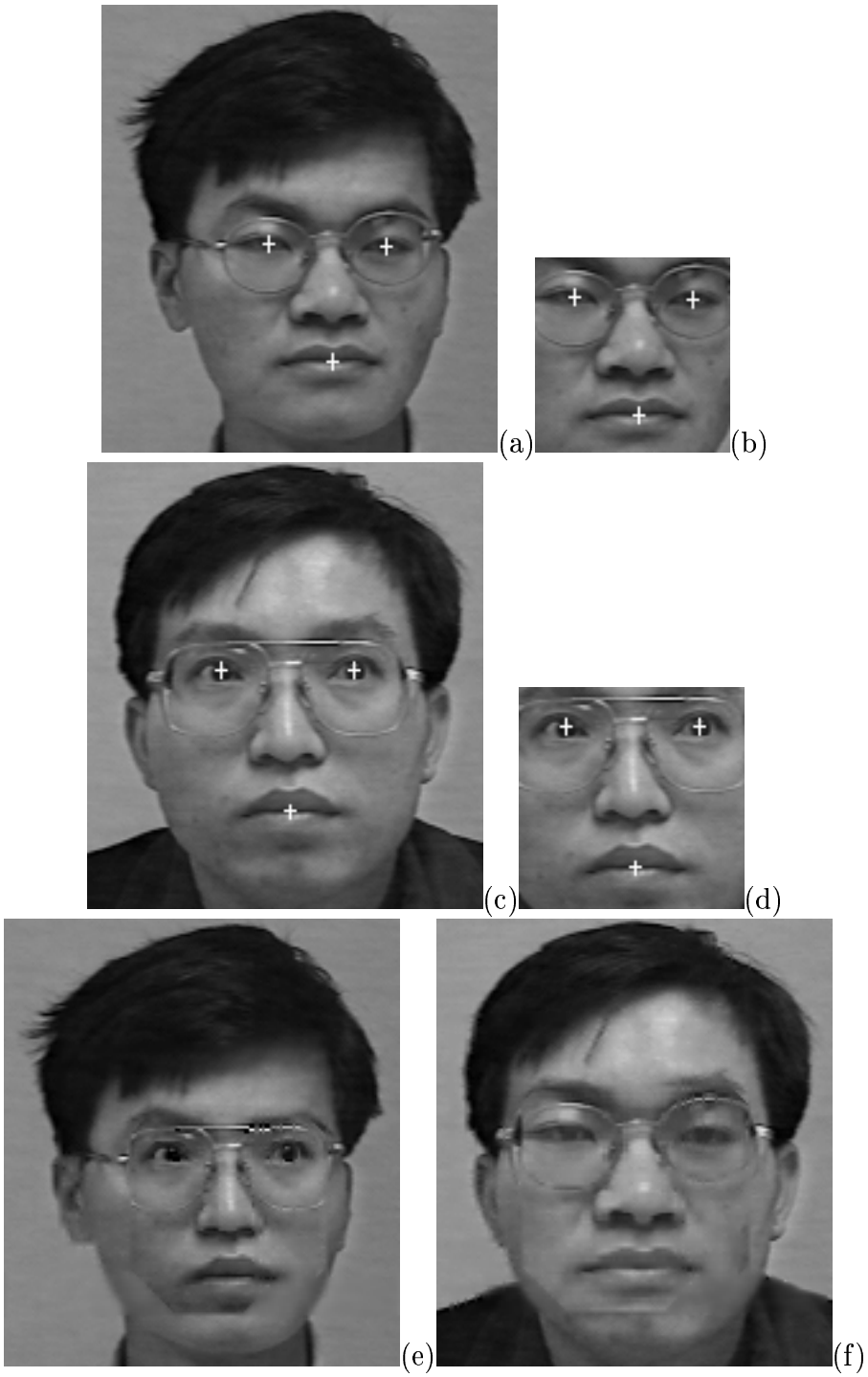

Figure 1: Examples of detecting middle face portions from face images. (a) and (c) show two face images, each of which contains three landmarks. (b) and (d) show the corresponding middle face portions of (a) and (b), respectively. (e) and (f) are the synthetic face images obtained by exchanging the middle face portions of (a) and (c), respectively.

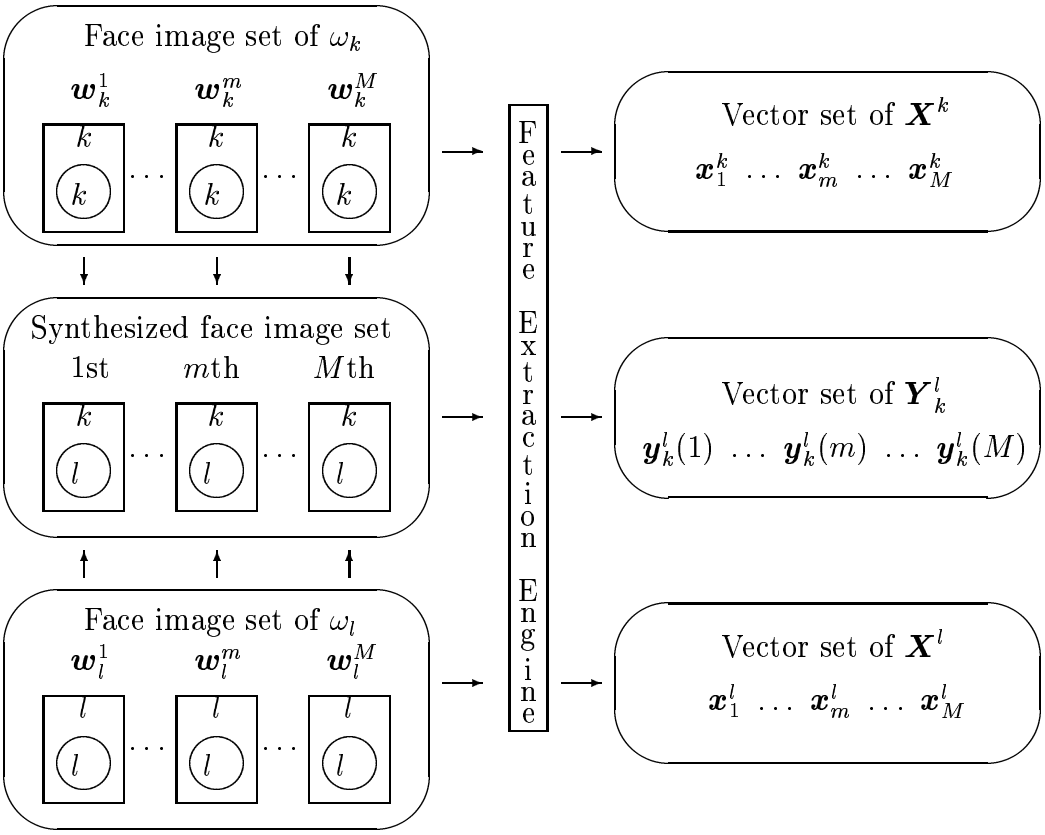

Figure 2: Each rectangle in the left column represents one face image, and the circle area is the middle face portion. The middle entry in the left column shows that each synthesized face image corresponding to vector $\boldsymbol{y}_k^l(m)$ is obtained by combining the middle face portion of $\boldsymbol{w}_l^m$ in class ω_l with the nonface portion of its counterpart $\boldsymbol{w}_k^m$ in class ω_k.

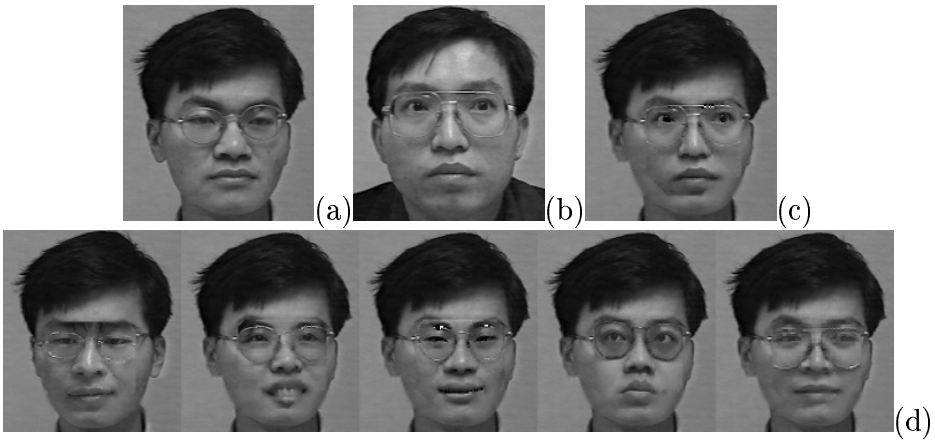

Figure 3: Examples of synthesized face images. (a) The mth face image in ω_k – $\boldsymbol{w}_k^m$; (b) the mth face image in ω_l – $\boldsymbol{w}_l^m$; (c) the synthesized face image obtained by combining the middle face portion of $\boldsymbol{w}_l^m$ and the nonface portion of $\boldsymbol{w}_k^m$. The extracted feature vector corresponding to this synthesized face image is $\boldsymbol{y}_k^l(m)$; (d) some other examples with 5 different l's (persons).

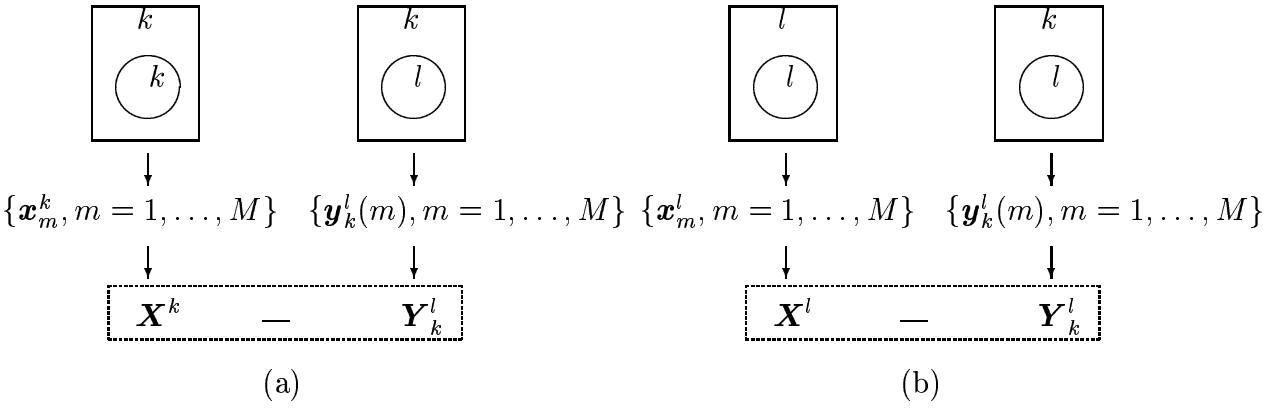

Figure 4: In the top rows of (a) and (b), each rectangle region together with the circle region inside it represent a face image. The mark k or l denotes the class to which that region belongs. The feature vectors in the middle rows of (a) and (b) are extracted from the corresponding face images (pure or synthesized). The assemblages of all vectors (e.g., $\boldsymbol{x}_m^k$) form normal distributions of corresponding vector sets (e.g., $\boldsymbol{X}^k$). The bottom rows of (a) and (b) represent the difference between the two distributions, which can be computed using the Bhattacharyya distance.

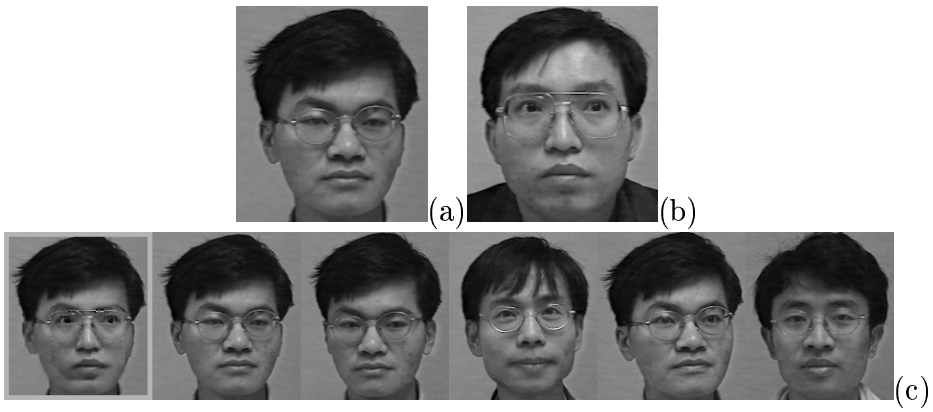

Figure 5: PCA plus LDA based face recognition using a synthesized face image as the query image . (a) and (b) are the original face images of two persons. The leftmost image of (c) is the query image synthesized from (a) and (b), and the other images are the top 5 closest retrieved images (ordered from left to right.)

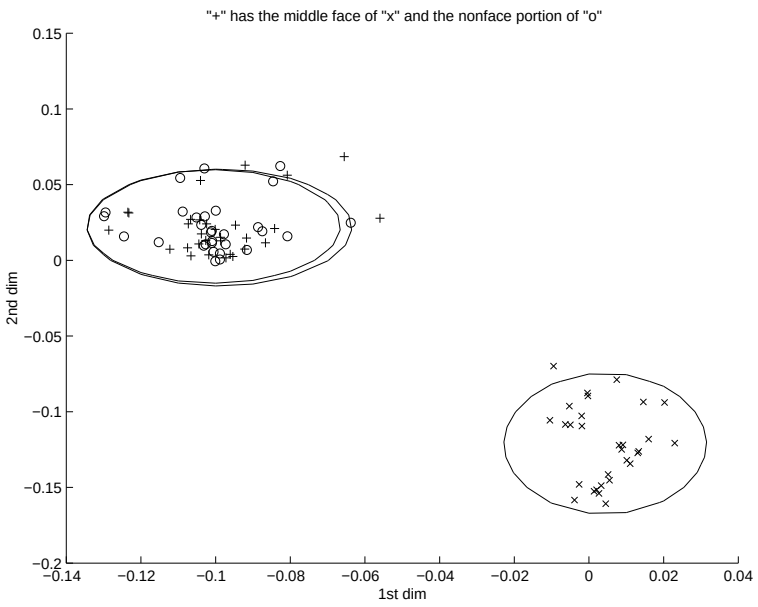

Figure 6: The distributions of 2-dimensional vectors which were extracted using the PCA plus LDA approach. Each node represents the feature vector extracted from a face image, and there were 30 nodes for each person. 'o' and 'x' represent $\boldsymbol{X}^k$ and $\boldsymbol{X}^l$ of persons k and l, respectively. '+' stands for $\boldsymbol{Y}_k^l$, which represents the synthesized image obtained by combining the middle face portion of person l with the nonface portion of person k. The horizontal axis and vertical axis are, respectively, the most discriminating and the second most discriminating projection axes in the projective feature space. This figure shows that '+' ($\boldsymbol{Y}_k^l$) was very close to class 'o' ($\boldsymbol{X}^k$).

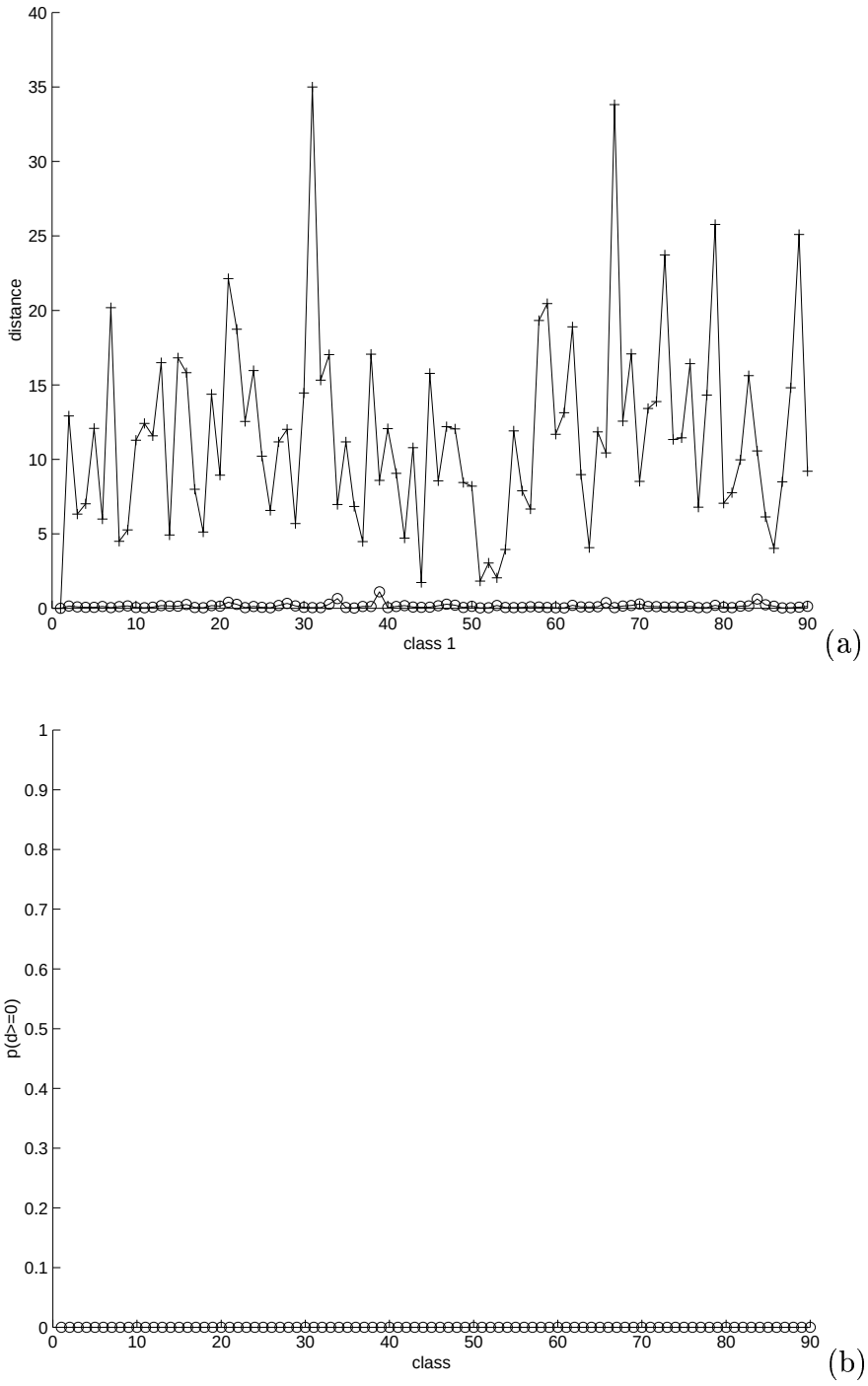

Figure 7: The experimental results for D^k obtained using the PCA plus LDA approach. 'o' is the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$, and '+' is the distance between $\boldsymbol{X}^l$ and $\boldsymbol{Y}_k^l$. (a) shows the values of the first term ('o') and the second term ('+') of every $d^k(l)$ in D^k, $l = 2, \ldots, 90$, where $k = 1$; (b) shows the individual probabilities of $p(d^k(l) \geq 0; d^k(l) \in D^k)$, $k = 1, \ldots, 90$. These figures show that $\boldsymbol{Y}_k^l$ will be classified into class k, which includes the nonface portion of $\boldsymbol{Y}_k^l$.