Why A Statistics-based Face Recognition System Should Base Its Recognition on the Pure Face Portion: A Probabilistic Decision-based Proof*

Li-Fen Chen† Hong-Yuan Mark Liao‡† Chin-Chuan Han‡
Ja-Chen Lin†

†Department of Computer and Information Science,
National Chiao Tung University, Taiwan
‡Institute of Information Science, Academia Sinica, Taiwan
{corinna,liao}@iis.sinica.edu.tw

Abstract

Face recognition, by definition, should be a recognition process in which recognition is based on the content of a face. The problem is: what is a "face"? Goudail *et al.* [1] and Swets and Weng [2] have recently proposed state-of-the-art statistics-based face recognition systems. However, they 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 different portions of a face image will influence the recognition results, two

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

hypothesis testing models are proposed. We then implement the two above mentioned face recognition systems and use the proposed hypothesis testing models to evaluate the systems. Experimental results reflected that the influence of the "real" face portion is much less than that of the nonface portion. This outcome confirms quantitatively that a statistics-based face recognition system should base its recognition solely on the "pure" face portion.

1 Introduction

Face recognition has been a very hot research topic in recent years [3, 4, 5]. 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 [6, 7, 8, 9, 10, 11, 12, 13]. The second stage of a face recognition system involves recognizing the target faces obtained in the first stage. Recently, some successful face recognition systems have been developed and reported in the literature [1, 2, 14, 15, 16, 17]. Among these works, the systems proposed by Goudail *et al.*[1] and Swets and Weng [2] represent two state-of-the-art face recognition systems. However, Liao *et al.*[18] mentioned that these two statistics-based systems used "incorrect" databases because their face image covered face, hair, shoulders, and background, not solely face. It was pointed out in [18] that, in these two systems, the "facial" portion does not play a key role during execution of "face" recognition. From the psychological viewpoint, Hay and Young [19] also 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.

In recent years, some researchers have noticed this problem and tried to exclude those irrelevant "nonface" portions while performing face recognition. In [14], 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.*[8] tried to eliminate the near-boundary pixels of a normalized face image by using a fixed-size

mask. Moghaddam and Pentland [9] and Lin *et al.*[16] both used probabilistics-based face detectors to cut out the middle portion of a face image for correct recognition. In [18], 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. However, none of them tried to use a quantitative measure 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, the 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 solve the above mentioned problem.

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 two sets 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. 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 will influence the recognition result, two hypothesis testing models were proposed. We then implemented two state-of-the-art face recognition systems and used the proposed hypothesis testing models to evaluate the systems. Experimental results obtained from both systems 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 some of the previous statistics-based face recognition systems use "incorrect" face databases.

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

2 Two state-of-the-art Face Recognition Systems

In this section, two state-of-the-art face recognition systems which were implemented and used in the experiments will be introduced. In [1], Goudail *et al.* investigated the performance of a technique for face recognition based on the computation of 25 local autocorrelation coefficients. They used the set of transformed 25-dimensional database samples to determine the set of most discriminating projection axes based on linear discriminant analysis (LDA) and then calculated each sample's projective feature vector. When an unknow image appeared, its corresponding projective feature vector was calculated and compared with those of the database samples. For database construction, they asked all the persons to wear dark company jackets and to sit down in front of a uniform, black background. Although they kept the color of the background and cloth dark, their "face" image was actually a combination of face, hair, shoulders, and background. Basically, this kind of face image is "incorrect" in terms of "face" recognition. Another state-of-the-art system was proposed by Swets and Weng [2]. In this work, they applied the principal component analysis (PCA) technique to reduce the dimensionality of the original images. They selected the top 15 principal axes and used them to derive a 15-dimensional feature vector for every sample. These transformed samples were then used as bases to execute LDA, and they reported a peak recognition rate of more than 90%. Again, we find that their face image contained face, hair, shoulders, and background, not solely face. Since both methods mentioned above are statistics-based, we believe that inclusion of irrelevant "facial" portions, such as hair, shoulders, and background, will generate incorrect decision boundaries for recognition. Therefore, in this paper, we shall prove our argument through statistical methods. Since both of the above two face recognition systems adopted linear discriminant analysis (LDA), which is based on Fisher's criterion [20], to decide on the projection axes for the recognition purpose, we shall briefly introduce the LDA approach in the following paragraph.

Let the training set be comprised of K classes, where each class is for one person and contains M sample face images. In LDA, one determines the mapping

$$\boldsymbol{v}_m^k = A^t \boldsymbol{u}_m^k, \tag{1}$$

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 the 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. 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 [21]:

$$S_w = \sum_{k=1}^{K} \sum_{m=1}^{M} (\boldsymbol{v}_m^k - \bar{\boldsymbol{v}}^k)(\boldsymbol{v}_m^k - \bar{\boldsymbol{v}}^k)^t. \tag{2}$$

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

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

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

$$tr(S_w^{-1} S_b). \tag{4}$$

An algorithm which solves the mapping matrix A can be found in [22]. A Euclidean distance classifier is used to perform classification in the mapped space for these two face recognition systems.

3 Hypothesis Testing Models

We 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 two statistics-based hypothesis testing models 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 [23]:

$$p(\boldsymbol{x}|\omega_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Lambda|^{\frac{1}{2}}} exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Lambda^{-1} (\boldsymbol{x} - \boldsymbol{\mu})), \tag{5}$$

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. Here, 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 $\omega_k (k = 1, \ldots, K)$, two additional vector sets, $\boldsymbol{Y}_k^l$ and $\boldsymbol{Z}_l^k$ ($l = 1, \ldots, K$, $l \neq k$), are extracted and associated with it. The number of elements in $\boldsymbol{Y}_k^l$ or $\boldsymbol{Z}_l^k$ (for a specific l) is, respectively, equal to M, which is exactly the same as the number of elements in $\boldsymbol{X}^k$. The formation of the elements in $\boldsymbol{Y}_k^l$ or $\boldsymbol{Z}_l^k$ is as follows. Basically, 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. On the other hand, each element in $\boldsymbol{Z}_l^k$ is also a d-dimensional feature vector. The difference between $\boldsymbol{Y}_k^l$ and $\boldsymbol{Z}_l^k$ is that the latter is extracted from a synthesized face image which combines the middle face portion of an element in ω_k and the nonface portion of its corresponding element in ω_l. 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. In sum, for each vector set $\boldsymbol{X}^k$ of class ω_k ($k = 1, \ldots, K$), there are $2(K-1)$ 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 $2(K-1)$ feature sets which are associated with $\boldsymbol{X}^k$, defined as follows:

$$\begin{aligned}\boldsymbol{Y}_k^l \;=\; & \{\boldsymbol{y}_k^l(m), m = 1, \ldots, M| \;\; \boldsymbol{y}_k^l(m) \text{ is a } d\text{-dimensional feature vector extracted from a} \\ & \text{synthesized face image which combines the middle face portion of } \boldsymbol{w}_l^m \text{ and the} \\ & \text{nonface portion of } \boldsymbol{w}_k^m\}, \text{ and} \end{aligned} \tag{6}$$

$$\begin{aligned}\boldsymbol{Z}_l^k \;=\; & \{\boldsymbol{z}_l^k(m), m = 1, \ldots, M| \;\; \boldsymbol{z}_l^k(m) \text{ is a } d\text{-dimensional feature vector extracted from a} \\ & \text{synthesized face image which combines the middle face portion of } \boldsymbol{w}_k^m \text{ and the} \\ & \text{nonface portion of } \boldsymbol{w}_l^m\}. \end{aligned} \tag{7}$$

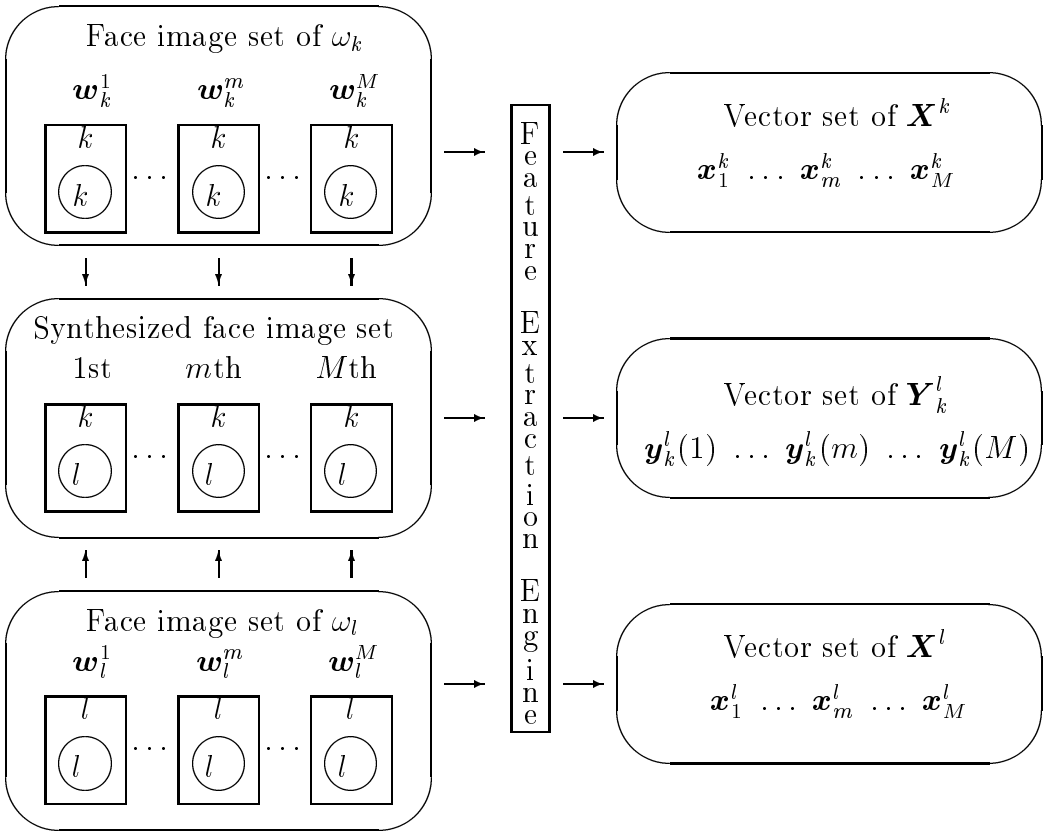

Figure 1: Each rectangle in the left column represents one face image, and the circle area is the middle face portion. The middle entry of the left column shows that each synthesized face image corresponding to vector $\boldsymbol{y}_k^l(m)$ is obtained by combining the middle face portion of $\boldsymbol{w}_l^m$ in class ω_l and the nonface portion of its counterpart $\boldsymbol{w}_k^m$ in class ω_k.

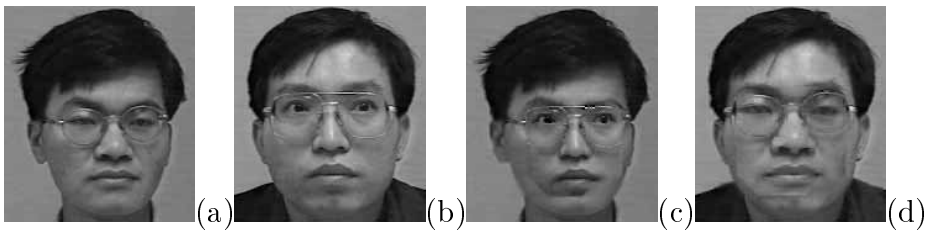

Figure 2: Examples of synthesized face images. (a) the mth face image in $\omega_k - \boldsymbol{w}_k^m$; (b) the mth face image in $\omega_l - \boldsymbol{w}_l^m$; (c) the synthesized face image obtained by combining the middle face portion of $\boldsymbol{w}_l^m$ and the nonface portion of $\boldsymbol{w}_k^m$. The extracted feature vector corresponding to this synthesized face image is $\boldsymbol{y}_k^l(m)$; (d) the synthesized face image obtained by combining the middle face portion of $\boldsymbol{w}_k^m$ and the nonface portion of $\boldsymbol{w}_l^m$. The extracted feature vector corresponding to this synthesized face image is $\boldsymbol{z}_l^k(m)$.

Figure 1 is a graphical illustration showing how $\boldsymbol{Y}_k^l$ is extracted. Figure 2 is a typical example illustrating how the synthesized face image is combined with the middle face portion of an image in ω_k and the nonface portion of its corresponding image in ω_l.

Bichsel and Pentland [15] 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 two feature vector sets, $\boldsymbol{Y}_k^l$ and $\boldsymbol{Z}_l^k$, as normal distribution functions. 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, the Bhattacharyya distance [24] is a well-known metric which is defined for measuring the similarity between two arbitrary statistical distributions. 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

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

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

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

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 [23]. In what follows, we shall define two hypothesis testing models as the tools for experiments. The Bhattacharyya distance will be used as a decision criterion for determining acceptance or rejection of our hypotheses.

3.1 First Hypothesis Testing Model

In the first 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_1^k. The

definition of D_1^k is as follows:

$$D_1^k = \{d_1^k(l), l = 1, \ldots, K; l \neq k |\ d_1^k(l) = D(\boldsymbol{X}^k, \boldsymbol{Y}_k^l) - D(\boldsymbol{X}^l, \boldsymbol{Y}_k^l)\}, \tag{10}$$

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_1^k. The physical meaning of every constituent of D_1^k, i.e., $d_1^k(l)$ $(l = 1, \ldots, K; l \neq k)$, is a statistical measure that can evaluate the importance, quantitatively, between the middle face portion and the nonface portion. Figure 3 illustrates how $d_1^k(l)$ is calculated in a graphical illustrative manner. Figure 3(a) shows how the first term that defines $d_1^k(l)$ is calculated. The top row of Figure 3(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 3(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 3(a) and (b) represent, respectively, the difference of 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 $d_1^k(l)$.

From Equation (10), it is obvious that when $d_1^k(l) \geq 0$, this means that 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 $d_1^k(l)$ which has a value not less than zero. If, unfortunately, the result turns out to be $d_1^k(l) < 0$, 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

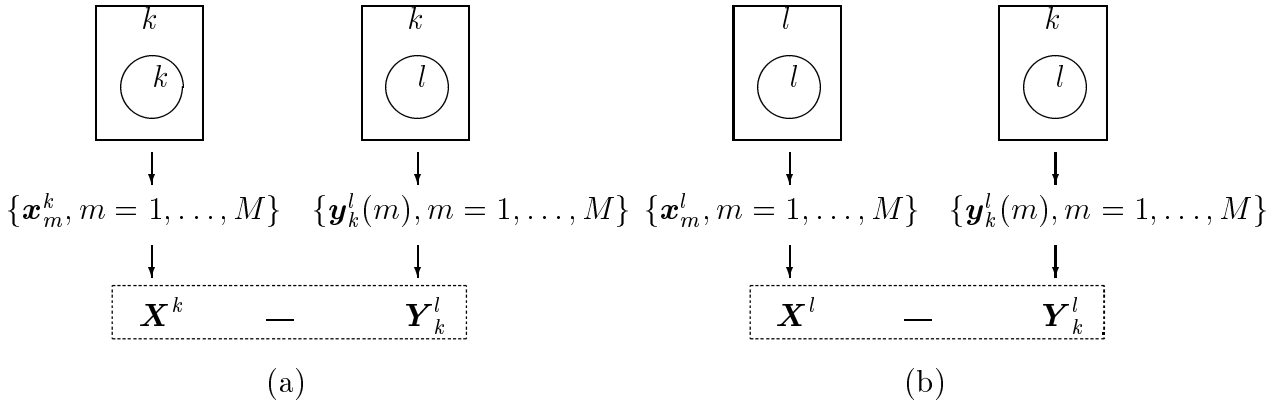

Figure 3: In the top rows of (a) and (b), each rectangle region together with the circle region inside represent a face image. The mark k or l denotes the class to which that region belongs. The feature vectors in the middle rows of (a) and (b) are extracted from the corresponding face images (pure or 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 of the two distributions, which can be computed using the Bhattacharyya distance.

$K-1$ possible synthesized face image sets. Therefore, we shall have $K-1$ $d_1^k(l)$ values (for $l = 1, \ldots, K, l \neq k$). From the statistical viewpoint, if more than half of these $d_1^k(l)$ 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:

$$\begin{aligned} \bar{H}_1^k &: p(d_1^k(l) \geq 0; d_1^k(l) \in D_1^k) \geq 0.5, \\ H_1^k &: p(d_1^k(l) \geq 0; d_1^k(l) \in D_1^k) < 0.5, \end{aligned} \tag{11}$$

where $\bar{H}_1^k$ represents the null hypothesis, H_1^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_1^k, it contains $K-1$ $d_1^k(l)$ real values. Therefore, the rules defined in Equation (11) will let the null hypothesis $\bar{H}_1^k$ be accepted whenever the amount of $d_1^k(l)$ which has a value not less than zero is more than one half of $K-1$; otherwise, the alternative hypothesis H_1^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:

$$\begin{aligned} \bar{H}_1 &: p(\bar{H}_1^k \text{ is accepted, } k = 1, \ldots, K) \geq 0.5, \\ H_1 &: p(\bar{H}_1^k \text{ is accepted, } k = 1, \ldots, K) < 0.5. \end{aligned} \tag{12}$$

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 $\bar{H}_1$ 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.

3.2 Second Hypothesis Testing Model

The objective of the second hypothesis testing model is to prove our assertion in an alternative manner. In order to achieve this goal, we used the two previously defined synthesized face image databases, $\boldsymbol{Y}_k^l$ and $\boldsymbol{Z}_l^k$, to conduct the testing process. The metric to be defined here is similar to D_1^k. That is, the metric defined for a specific feature set $\boldsymbol{X}^k$ is a real-number set, D_2^k. The definition of D_2^k is as follows:

$$D_2^k = \{d_2^k(l), l = 1, \ldots, K, l \neq k |\ d_2^k(l) = D(\boldsymbol{X}^k, \boldsymbol{Y}_k^l) - D(\boldsymbol{X}^k, \boldsymbol{Z}_l^k)\}. \tag{13}$$

Again, for a specific feature set $\boldsymbol{X}^k$ corresponding to ω_k, there are in total $K-1$ elements contained in D_2^k. The main difference between D_2^k and D_1^k is that each $d_2^k(l)$ ($l = 1, \ldots, K$, $l \neq k$) in D_2^k is a statistical measure that compares the distance between the distribution of $\boldsymbol{X}^k$ and that of $\boldsymbol{Y}_k^l$ (extracted from a synthesized face image set which combines the middle face portions of elements in ω_l and the nonface portions of their counterpart elements in ω_k) with the distance between the distribution of $\boldsymbol{X}^k$ and that of $\boldsymbol{Z}_l^k$ (extracted from another synthesized face image set which combines the middle face portions of elements in ω_k and the nonface portions of their counterpart elements in ω_l). Figure 4 illustrates how $d_2^k(l)$ is calculated in a graphical illustrative manner. The representation of Figure 4 is the same as that of Figure 3 except for the definition of the second term. In what follows, we shall show why deciding either the middle face portion or the nonface portion is more important for face recognition based on the value of $d_2^k(l)$.

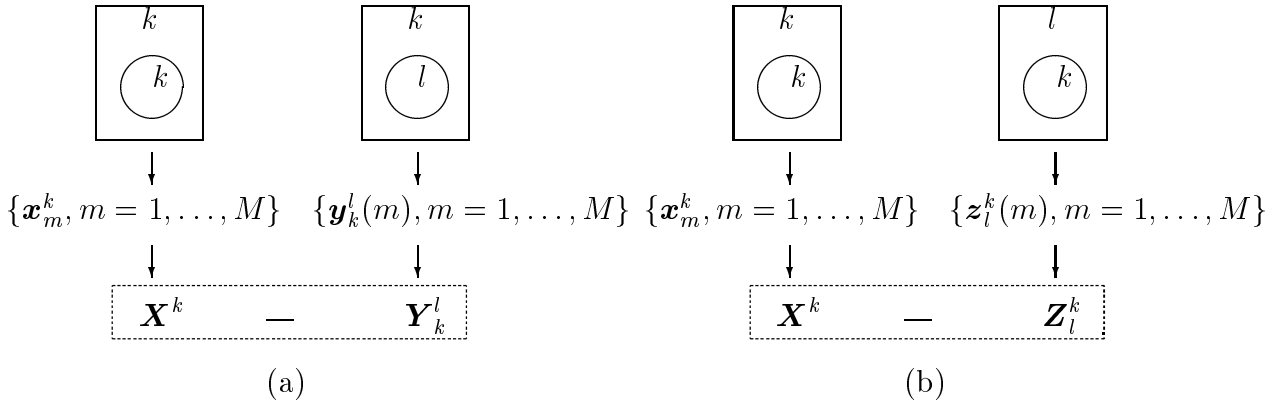

Figure 4: In the top rows of (a) and (b), each rectangular region together with the circle region inside represents a face image. The mark k or l denotes the class to which that region belongs. The feature vectors in the middle rows of (a) and (b) are extracted from the corresponding face images (pure 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 of two distributions, which can be computed using the Bhattacharyya distance.

From Equation (13), it is obvious that when $d_2^k(l) \geq 0$, this means that the distribution of $\boldsymbol{X}^k$ is closer to that of $\boldsymbol{Z}_l^k$ than to that of $\boldsymbol{Y}_k^l$. Otherwise, the distribution of $\boldsymbol{X}^k$ is closer to that of $\boldsymbol{Y}_k^l$ than to that of $\boldsymbol{Z}_l^k$. According to the definition of face recognition, a person should be recognized solely based on his/her own face, no matter how he/she changes his/her hair style or dresses. In other words, the middle face portion should dominate the distribution of the face image set more than the nonface portion. That is, a normal face recognition process should result in a $d_2^k(l)$ which has a value not less than zero. If unfortunately, the result turns out to be $d_2^k(l) < 0$, then it means that the nonface portion dominates the recognition process.

We have mentioned that for a specific class k, there are in total $K-1$ paired synthesized face image sets (corresponding to $\boldsymbol{Y}_k^l$ and $\boldsymbol{Z}_l^k$, $l = 1, \ldots, K$, $l \neq k$). Therefore, we shall have $K-1$ $d_2^k(l)$ values. Again, if more than half of these $d_2^k(l)$ values are less than zero, then this means that the nonface portion dominates the face recognition process. The formal definition of the rules for person k is as follows:

$$\begin{aligned} \bar{H}_2^k &: p(d_2^k(l) \geq 0; d_2^k(l) \in D_2^k) \geq 0.5, \\ H_2^k &: p(d_2^k(l) \geq 0; d_2^k(l) \in D_2^k) < 0.5, \end{aligned} \tag{14}$$

where $\bar{H}_2^k$ represents the null hypothesis, and H_2^k stands for the alternative hypothesis. Ac-

cording to the definition of D_2^k, it is a set containing $K-1$ $d_2^k(l)$ real values. Therefore, the rules defined in Equation (14) will let the null hypothesis $\bar{H}_2^k$ be accepted whenever the number of $d_2^k(l)$ which has a value not less than zero is more than one half of $K-1$; otherwise, the alternative hypothesis H_2^k will be accepted.

The rules described in Equation (14) are for a specific class k. If they are extended to the whole population, a global hypothesis test rule is required. The extension is as follows:

$$\begin{aligned} \bar{H}_2 &: p(\bar{H}_2^k \text{ is accepted, for } k = 1, \ldots, K) \geq 0.5, \\ H_2 &: p(\bar{H}_2^k \text{ is accepted, for } k = 1, \ldots, K) < 0.5. \end{aligned} \tag{15}$$

The physical meaning of the rules described in Equation (15) is that when over half of the population pass the null hypothesis, the global null hypothesis $\bar{H}_2$ is accepted; otherwise, the global alternative hypothesis will be accepted. When the latter is true, this means that the nonface portion of a face image plays a major role in the face recognition process.

4 Experimental Results

In the experiments, the two above mentioned statistics-based state-of-the-art face recognition systems [1, 2] were implemented and tested against the two proposed hypothesis testing models. 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. In the autocorrelation plus LDA approach proposed by Goudail *et al.* [1], each projective feature vector obtained from a face image is 24-dimensional. As to the PCA plus LDA approach proposed by Swets and Weng [2], each projective feature vector extracted from a face image is 15-dimensional. Based on these feature vectors of training samples, the two hypothesis models were tested. Since the projection axes derived through linear discriminant analysis were ordered according to their discriminating capability, the first projection axis was most discriminating and then the second projection axis. For the convenience of visualization, all samples were projected onto the first two projection axes and are shown in

Figures 5 and 6, respectively, for the first and second hypotheses models.

Figure 5 shows the three related distributions covered in D_1^k (the first hypothesis model). '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 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. The two most discriminating projection axes shown in Figure 5(a) were determined using the autocorrelation plus LDA approach. It is obvious that the distribution of $\boldsymbol{Y}_k^l$ was 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. As to the case of PCA plus LDA, which is shown in Figure 5(b), the above mentioned phenomenon was even stronger. That is, the distribution of $\boldsymbol{Y}_k^l$ was completely disjointed from that of $\boldsymbol{X}^l$ and almost completely overlapped that of $\boldsymbol{X}^k$. In sum, the experiments shown in Figure 5(a) and (b) both confirmed that the nonface portion of a face image did dominate the distributions of the 2-dimensional projective feature vectors. The experiments shown in Figure 6 are the results associated with the second hypothesis test. From Equation (13), it is seen that the three distributions covered in D_2^k are $\boldsymbol{X}^k$, $\boldsymbol{Y}_k^l$, and $\boldsymbol{Z}_l^k$. They are represented by 'o', '+', and '*', respectively, in Figure 6. The result shown in Figure 6(a) was the outcome obtained by applying the autocorrelation plus LDA approach. From this experiment, we find that the distribution of $\boldsymbol{X}^k$ was closer to that of $\boldsymbol{Y}_k^l$ than to that of $\boldsymbol{Z}_l^k$. As to the PCA plus LDA approach (Figure 6(b)), the above mentioned phenomenon was, again, stronger. This means that the distribution of $\boldsymbol{Z}_l^k$ was completely disjointed from that of $\boldsymbol{X}^k$. Both experimental results shown in Figure 6 also confirm that the real face portion (middle face) of a face image "did not" play any (or only a small) role in the face recognition process.

Figures 7 and 8 showed the experimental results obtained by applying the first hypothesis testing model. The data shown in Figure 7 are the results extracted by executing the autocorrelation plus LDA approach. The data shown in Figure 8, on the other hand, are the results extracted by performing the PCA plus LDA approach. In both cases, k was set to 1. That is, l ranged from 2 to 90 in both sets of experiments. 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 $d_1^k(l)$. 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 $d_1^k(l)$. The results shown in Figure 7(a) reflect that from l=2 to 90, the second term ('+') of $d_1^k(l)$ 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 $d_1^k(l)$ $(l = 2, \ldots, 90)$ was larger than the second term of $d_1^k(l)$ is zero. Figure 7(b) shows, from class 1 to class 90, the individual probability that the first term of $d_1^k(l)$ $(l = 2, \ldots, 90)$ was larger than the second term of $d_1^k(l)$. 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 zeros (less than 0.05). Figure 8 shows the results obtained by performing the PCA plus LDA approach. The definition of the 'o' sign and that of the '+' sign are the same as in Figure 7(a). One thing worth noticing is that the PCA plus LDA approach had the ability to extract more "discriminating" projection axes than the autocorrelation plus LDA approach did. Therefore, the phenomenon whereby the nonface portion dominated the face recognition process was even more apparent in the PCA plus LDA approach. This conclusion is confirmed by the individual probability values shown in Figure 8(b). We can see that all the individual probabilities were equal to zero when the PCA plus LDA approach was applied. From the individual probabilities shown in Figures 7(b) and 8(b), we can draw a conclusion that all the null hypotheses $\bar{H}_1^k$'s $(k = 1, \ldots, 90)$ were rejected, and that the probability of accepting $\bar{H}_1^k$ $(k = 1, \ldots, 90)$ was equal to zero.

As for testing of the second hypothesis model against the two state-of-the-art systems, the results are reported in Figures 9 and 10, respectively. The results shown in Figure 9 were obtained by performing the autocorrelation plus LDA approach. The 'o' sign and the '+' sign represent, respectively, the Bhattacharyya distances between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$ and between $\boldsymbol{X}^k$ and $\boldsymbol{Z}_l^k$. Again, the experimental results show that the nonface portion dominated the face recognition process. On the other hand, the experimental results shown in Figure 10 (the PCA plus LDA approach) also agreed with the above mentioned assertion. Following the convention commonly adopted in the hypothesis testing process, the testing results for both state-of-the-art systems [1, 2] are listed in Table 1. All the results shown in Table 1 confirm that the nonface portions of all the testing images did play a discriminating role in the face

recognition systems used in [1] and [2].

Face Recognition Systems	First Hypothesis Testing			Second Hypothesis Testing		
	n_0	Z	Accept	n_0	Z	Accept
autocorrelation + LDA	0	-9.49	H_1	0	-9.49	H_2
PCA + LDA	0	-9.49	H_1	0	-9.49	H_2

Table 1: The experimental results for our two hypotheses models. n_0 is the number of successes and Z is the test statistic.

5 Conclusions

In this paper, we have proposed a statistics-based technique to quantitatively prove that two previously proposed face recognition systems used "incorrect" databases. According to the definition of face recognition, the recognition process should be dominated by the "pure" face portion. However, after implementing two state-of-the-art statistics-based face recognition systems, we have shown, quantitatively, that the influence of the middle face portion on the recognition process in their systems was much less 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. Our suggestion for future research is that a statistics-based face recognition system should base its recognition solely on a face-only database.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object detection", *Proc. 5th IEEE Conf. on Computer Vision*, pp. 786–793, 1995.

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

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

[12] 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", to appear in *Pattern Recognition*, 1997.

[13] C. C. Han, H. Y. Mark Liao, G. J. Yu, and L. H. Chen, "Fast face detection via morphology-based pre-processing", in *Proc. 9th International Conference on Image Analysis and Processing*, pp. 469-476, 1997.

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

[15] 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.

[16] 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.

[17] H. Y. Mark Liao, C. C. Han, G. J. Yu, M. C. Chen H. R. Tyan, and L. H. Chen, "Face recognition using a face-only database: A new approach", *Proc. 3rd Asian Conference on Computer Vision*, Jan. 1998.

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

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

[20] A. Fisher, *The Mathematical Theory of Probabilities*, Macmillan, New York, 1923.

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

[22] 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.

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

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

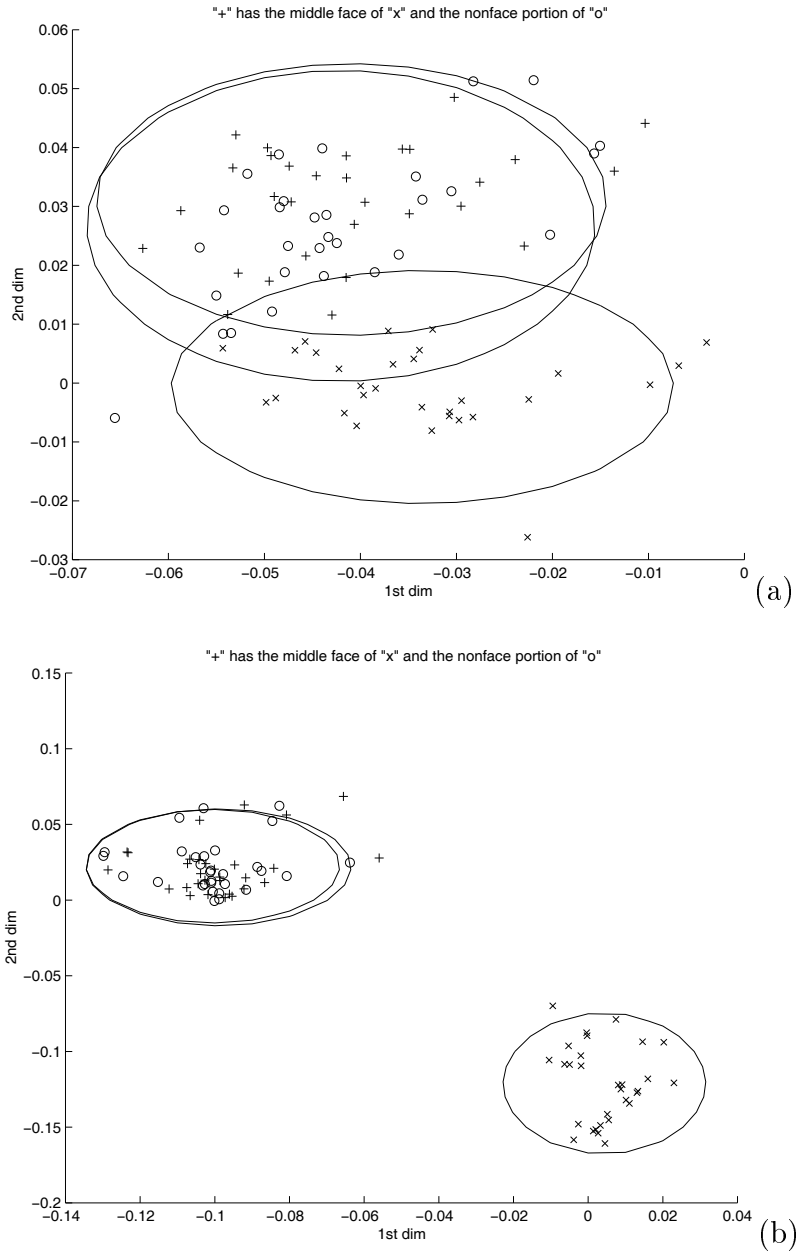

Figure 5: The distributions of 2-dimensional vectors associated with the first hypothesis model. Each node represents the feature vector extracted from a face image, and there are 30 nodes for each person. 'o' and 'x' represent $\boldsymbol{X}^k$ and $\boldsymbol{X}^l$ of persons k and l, respectively. '+' stands for $\boldsymbol{Y}_k^l$, which represents the synthesized image by combining the middle face of person l and the nonface portion of person k. The horizontal axis and vertical axis in (a) and (b) are, respectively, the most discriminating and the second most discriminating projection axes in the feature space. (a) shows the distributions of feature vectors extracted by the autocorrelation plus LDA approach; (b) shows the distributions of feature vectors extracted by the PCA plus LDA approach. This figure shows that '+' ($\boldsymbol{Y}_k^l$) was classified into class 'o' ($\boldsymbol{X}^k$).

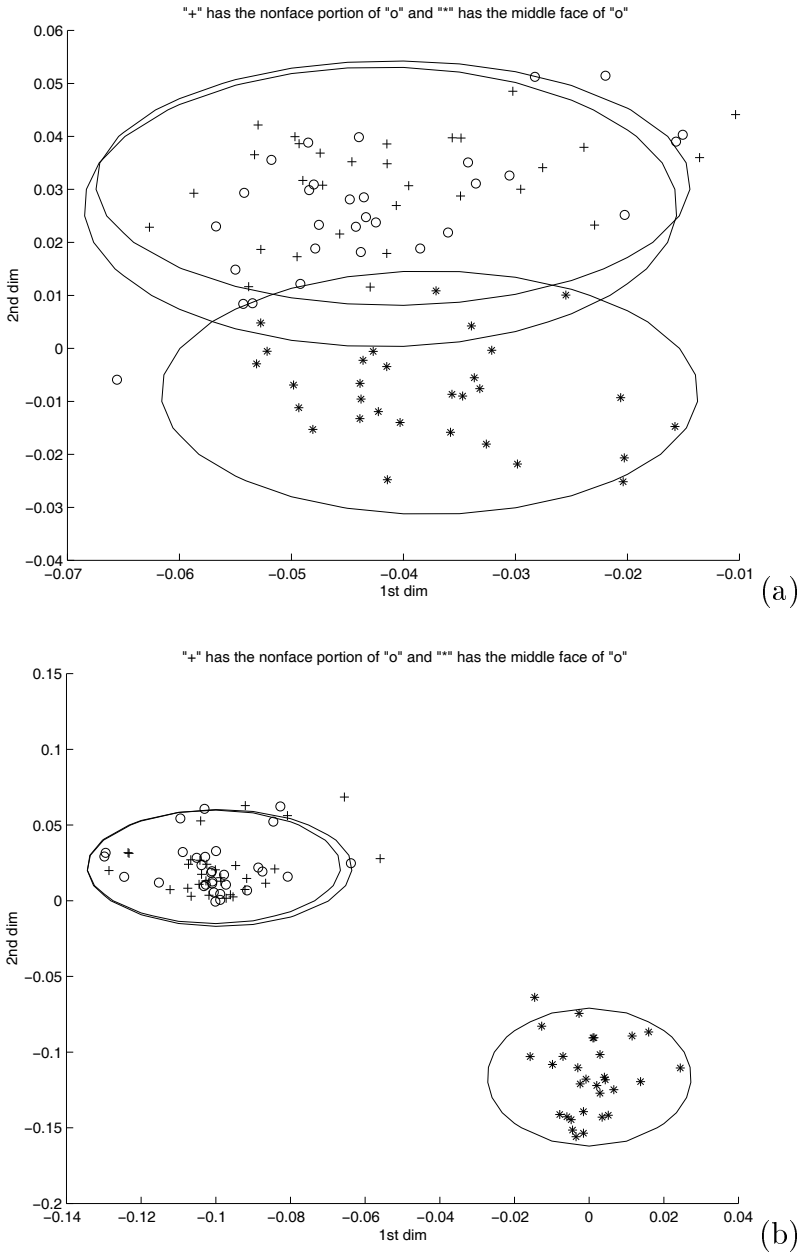

Figure 6: The distributions of 2-dimensional vectors associated with the second hypothesis model. Each node represents the feature vector extracted from a face image, and there are 30 nodes for each person. 'o' represents $\boldsymbol{X}^k$ of person k, '+' stands for $\boldsymbol{Y}_k^l$, which represents the synthesized face image by combining the middle face portion of person l and the nonface portion of person k, and '*' stands for $\boldsymbol{Z}_l^k$, which represents the synthesized face image by combining the middle face portion of person k and the nonface portion of person l. (a) shows the distributions of feature vectors extracted by the autocorrelation plus LDA approach; (b) shows the distributions of feature vectors extracted by the PCA plus LDA approach. Both (a) and (b) confirm that the nonface portion dominated the distribution of a face image set.

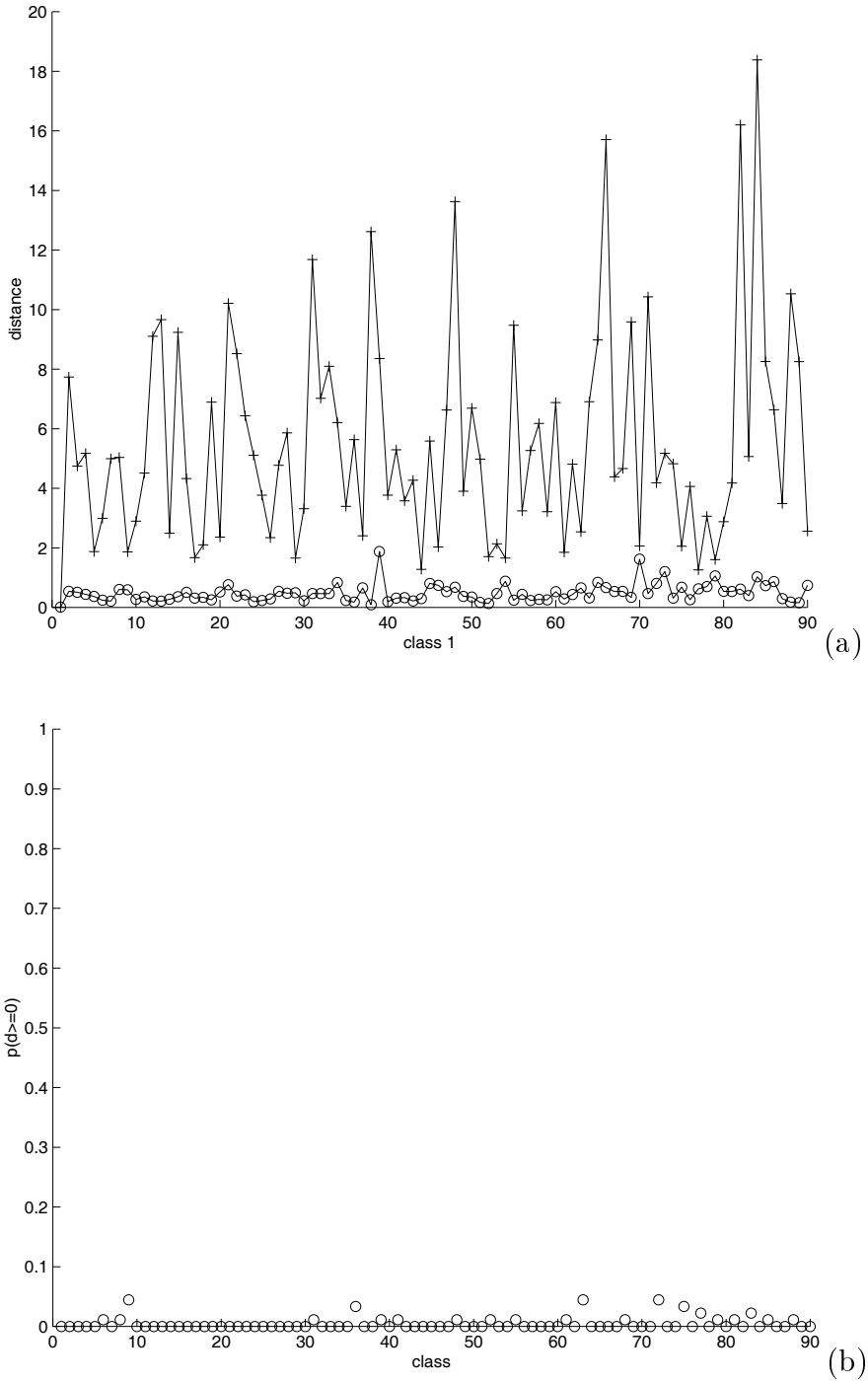

Figure 7: The experimental results for D_1^k using the autocorrelation plus LDA approach. 'o' is the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$, and '+' is the distance between $\boldsymbol{X}^l$ and $\boldsymbol{Y}_k^l$. (a) shows the values of the first term ('o') and the second term ('+') of every $d_1^k(l)$ in D_1^k, $l = 2, \ldots, 90$, where $k = 1$; (b) shows the individual probabilities of $p(d_1^k(l) \geq 0; d_1^k(l) \in D_1^k)$, $k = 1, \ldots, 90$.

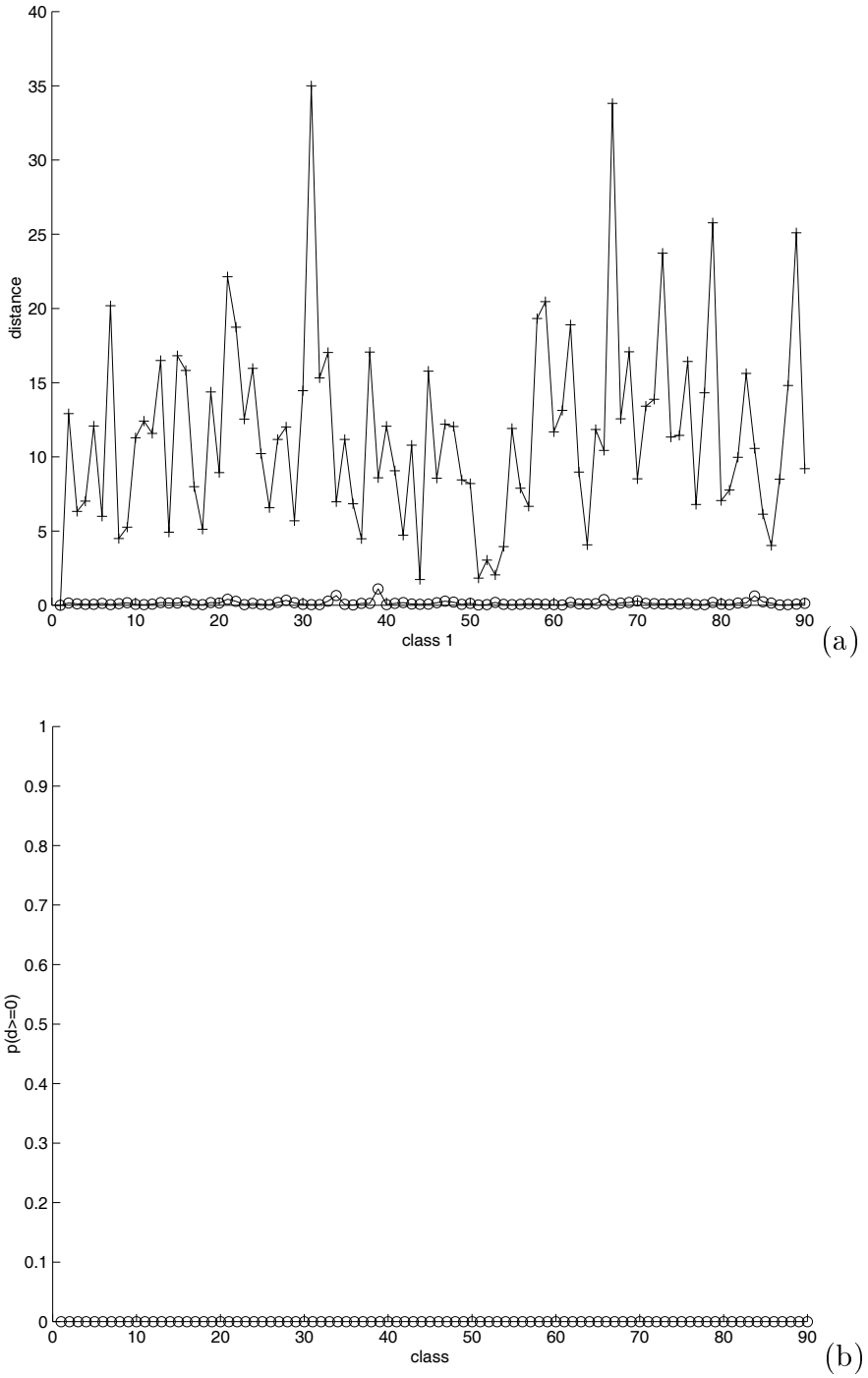

Figure 8: The experimental results for D_1^k using the PCA plus LDA approach. 'o' is the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$, and '+' is the distance between $\boldsymbol{X}^l$ and $\boldsymbol{Y}_k^l$. (a) shows the values of the first term ('o') and the second term ('+') of every $d_1^k(l)$ in D_1^k, $l = 2, \ldots, 90$, where $k = 1$; (b) shows the individual probabilities of $p(d_1^k(l) \geq 0; d_1^k(l) \in D_1^k)$, $k = 1, \ldots, 90$.

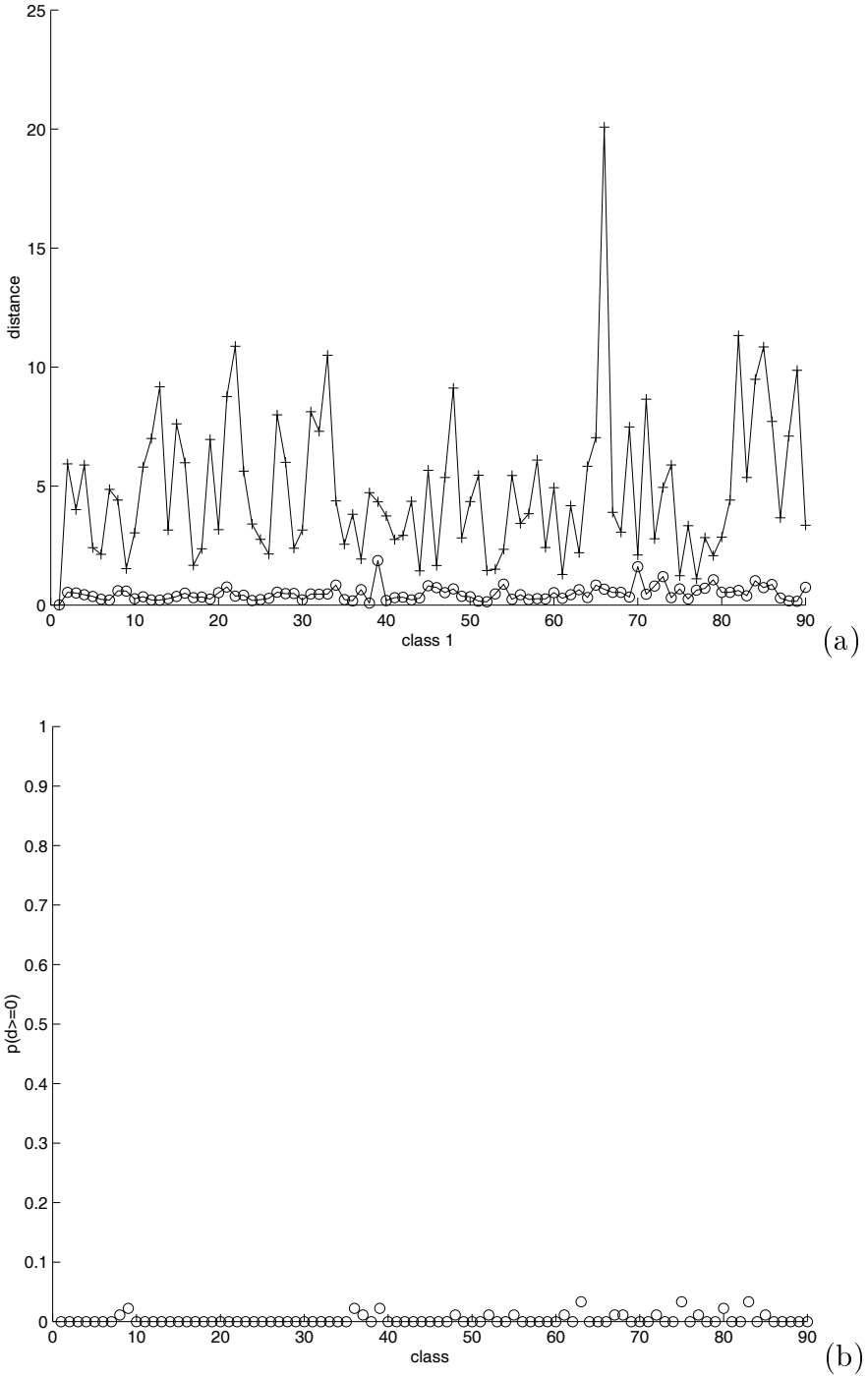

Figure 9: The experimental results for D_2^k using the autocorrelation plus LDA approach. 'o' is the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$, and '+' is the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Z}_l^k$. (a) shows the values of the first term ('o') and the second term ('+') of every $d_2^k(l)$ in D_2^k, $l = 2, \ldots, 90$, where $k = 1$; (b) shows the individual probabilities of $p(d_2^k(l) \geq 0; d_2^k(l) \in D_2^k)$, $k = 1, \ldots, 90$.

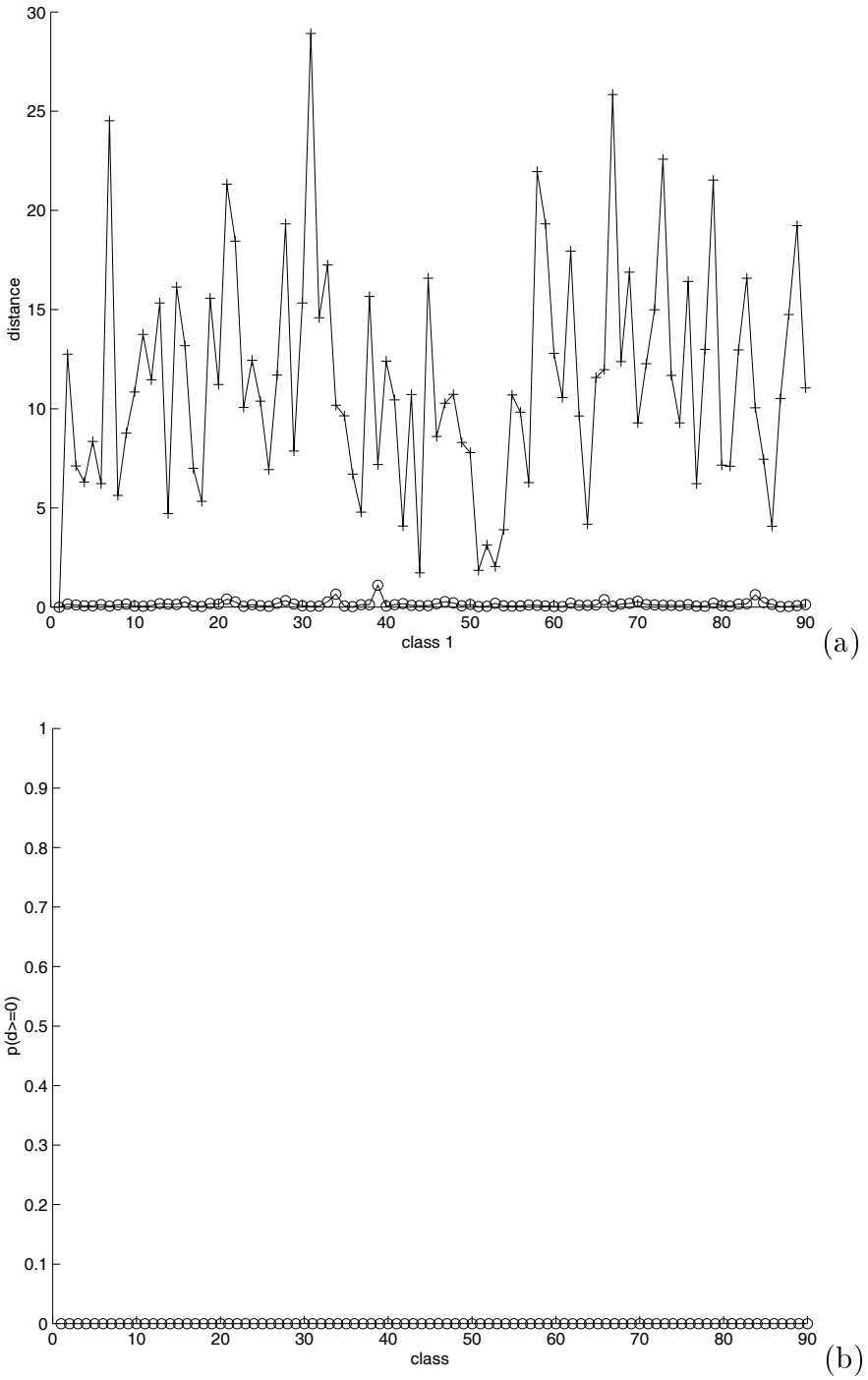

Figure 10: The experimental results for D_2^k using the PCA plus LDA approach. 'o' is the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Y}_k^l$, and '+' is the distance between $\boldsymbol{X}^k$ and $\boldsymbol{Z}_l^k$. (a) shows the values of the first term ('o') and the second term ('+') of every $d_2^k(l)$ in D_2^k, $l = 2, \ldots, 90$, where $k = 1$; (b) shows the individual probabilities of $p(d_2^k(l) \geq 0; d_2^k(l) \in D_2^k)$, $k = 1, \ldots, 90$.