

A Face and Speech Biometric Verification System Using A Simple Bayesian Structure

ANDREW B. J. TEOH, S. A. SAMAD* AND A. HUSSAIN*

Faculty of Information Science and Technology (FIST)

Multimedia University

75450, Melaka, Malaysia

E-mail: bjteoh@mmu.edu.my

**Electrical, Electronic and System Engineering Department*

National University of Malaysia

43600, Bangi, Malaysia

E-mail: {salina; aini}@eng.ukm.my

Identity verification systems that use a mono modal biometric always have to contend with sensor noise and limitations of the feature extractor and matcher, while combining information from different biometrics modalities may well provide higher and more consistent performance levels. However, an intelligent scheme is required to fuse the decisions produced by the individual sensors. This paper presents a decision fusion technique for a bimodal biometric verification system that makes use of facial and speech biometrics. The decision fusion schemes considered have simple Bayesian structures (SBS) that particularize the univariate Gaussian density function, Beta density function or Parzen window density estimation. SBS has advantages in terms of computation speed, storage space and its open framework. The performances of SBS is evaluated and compared with that of other classical classification approaches, such as sum rule and Multi-layer Perceptron, on a bimodal database.

Keywords: bimodal biometrics, face module, speech module, simple bayesian structure, decision fusion

1. INTRODUCTION

In today's electronically wired information society, there are more and more situations which require an individual, as a user, to be verified by an electronic machine as in the case of transaction authentication on physical or virtual access control. Traditionally, these activities have mostly been conducted using ID numbers, such as a token or a password. The main problem with these numbers is that they can be used by unauthorized persons. On the other hand, biometric techniques use unique personal features of the user himself to verify the identity claimed. These techniques employ face, facial termogram, fingerprint, hand geometry, hand vein, iris, retinal pattern, signature, or voice print information. All these features have different degrees of uniqueness, permanence, measurability, user acceptability, performance, and robustness against circumvention [1].

However, there are some limitations to using just one biometric as the verification tool. For instance, it is estimated that 5% of the population does not have legible finger-

Received April 28, 2003; revised August 11, 2003; accepted July 8, 2004.
Communicated by Kuo-Chin Fan.

prints, a voice could be altered by a cold, and face recognition systems are susceptible to aging effects, changes in ambient light, and the pose of the subject. In addition, correct verification may not be guaranteed due to sensor noise and limitations of the feature extractor and matcher. One way to cope with these limitations is to combine several biometrics in a multimodal identity verification system. A multimodal biometric system uses multiple sensors to capture different types of biometrics. This allows the integration of two or more types of biometric recognition and verification systems in order to meet stringent performance requirements. The enhanced structure takes advantage of the proficiency of each individual biometric and can be used to overcome some of the limitations of a single biometric. Multimodals are generally much more vital to deal with fraud technologies because it is more difficult to forge multiple biometric characteristics than to forge a single biometric characteristic. However, an intelligent scheme is required to fuse the decisions churned out by the individual sensors.

In the literature, Brunelli and Falavigna [2] proposed a person identification system based on acoustic and visual features. They use a HyperBF network as the best performing fusion module. Dieckmann *et al.* [3] proposed an abstract level fusion scheme called the “2 from 3 approach,” which integrates the face, lip motion, and voice based on the principle that a human uses multiple clues to identify a person. Duc *et al.* proposed in [4] a simple averaging technique and compared it with the Bayesian integration scheme presented by Bigun *et al.* in [5]. In this multimodal system, the authors use a face identification expert and a text-dependent speech expert. Kittler *et al.* proposed in [6] a multimodal person verification system, using three experts: frontal face, face profile, and voice. The best combination results are obtained for a simple sum rule. Hong and Jain proposed in [7] a multimodal personal identification system which integrates face and fingerprints that complement each other. The fusion algorithm operates at the decision level, where it combines the scores from the different experts by simply multiplying them. Ben-Yacoub proposed in [8] a multimodal data fusion approach for person authentication, based on the use of Support Vector Machines (SVM) to combine the results obtained from a face identification expert and a text-dependent speech expert. Choudhury *et al.* proposed in [9] a multimodal person recognition approach using unconstrained audio and video. Combination of the two experts is performed using a Bayes net. Ross *et al.* [10] combined the matching of face, fingerprint, and hand geometry to enhance the performance of their system. Three different techniques, the sum rule, decision tree, and linear discriminant analysis, are used to combine the matching scores. Most recently, Wang *et al.* [11] proposed to combine face and iris biometrics together, with Fisher’s discriminant analysis and radial basis function network employed in the fusion module.

A bimodal biometric verification system based on facial and vocal modalities is described in this paper. Both face image and speech biometrics are chosen due to their complementary characteristics, physiology, and behavior. The mutual independence of these two biometrics means that the data can be collected separately, thus making augmentation easy. Based on this supposition, a measurement level decision fusion scheme that makes use of the simple Bayesian structure (SBS) is considered. The proposed scheme is optimal in the Bayesian sense when sufficient data are available to obtain reasonable estimates of the univariate densities. The relatively simplicity of SBS compared to the aforementioned approaches offers a number of advantages in terms of the verification rate, learning and classification speed, and storage space as well as its open framework.

Therefore, the univariate Gaussian density function, Beta density function, and Parzen window density estimation have been chosen to particularize the SBS in order to achieve a high verification rate. Other various classical classifier schemes, such as sum rule and MLP, were also devised for decision fusion purposes and are compared with SBS here.

2. VERIFICATION MODULES

2.1 Face Verification

In personal verification, face recognition refers to static, controlled, full frontal portrait recognition. There are two major tasks in face recognition: (i) face detection and (ii) face verification.

In our system as shown in Fig. 1, the Eigenface approach [12] is used in the face detection and face recognition modules. The main idea behind the Eigenface approach is to find the vectors that best account for the distribution of face images within the entire image space, defined as the face space. Face spaces are eigenvectors of the covariance matrix corresponding to the original face images, and since they are face-like in appearance, they are so called Eigenfaces.

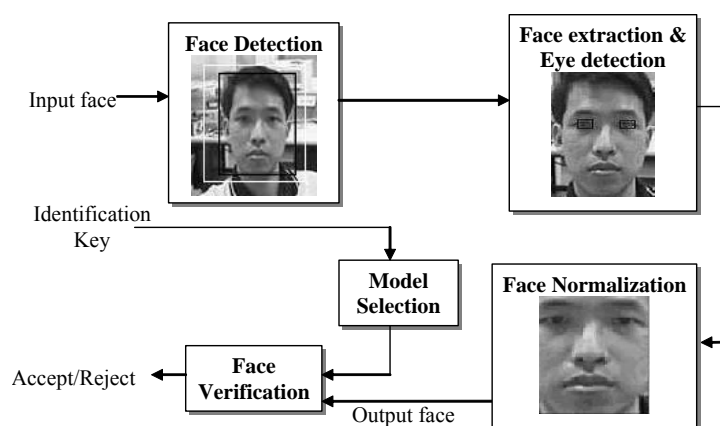


Fig. 1. Face verification system.

Now let the training set of face images be i_1, i_2, \dots, i_M ; the average face of the set is defined as

$$\bar{i} = \frac{1}{M} \sum_{j=1}^M i_j, \quad (1)$$

where M is the total number of images.

Each face differs from the average one by the vector $\phi_n = i_n - \bar{i}$. A covariance ma-

trix is constructed, where

$$\begin{aligned} C &= \sum_{j=1}^M \phi_j \phi_j^T \\ &= AA^T, \end{aligned} \quad (2)$$

where $A = [\phi_1 \ \phi_2 \ \dots \ \phi_M]$.

Then, the eigenvectors v_k and the eigenvalues λ_k with a symmetric matrix C are calculated. v_k determines the linear combination of M difference images with ϕ to form the Eigenfaces:

$$u_l = \sum_{k=1}^M v_{lk} \phi_k \quad l = 1, \dots, M. \quad (3)$$

From these Eigenfaces, $K (< M)$ Eigenfaces are selected corresponding to the K highest eigenvalues.

Face detection is accomplished by calculating the sum of the square error between a region of the scene and the Eigenface, a measure of the Distance From Face Space (DFFS) that is a measure of how face-like a region is. If a window, ψ , is swept across the scene to find the DFFS at each location, the most probable location of the face can be estimated. This will simply be the point where the reconstruction error, ε , has the minimum value:

$$\varepsilon = \|\psi - \psi_f\|, \quad (4)$$

where ψ_f is the projection into face-space.

From the extracted face, eye co-ordinates are determined using the hybrid rule based approach and contour mapping technique [13]. Based on the information obtained, scale normalization and lighting normalization are applied for a head in box format.

The Eigenface-based face recognition method is divided into two stages: (i) the training stage and (ii) the operational stage. In the training stage, a set of normalized face images, $\{i\}$, that best describe the distribution of the training facial images in a lower dimensional subspace (Eigenface) is computed by the following operation:

$$\varpi_{nk} = u_k (i_n - \bar{i}), \quad (5)$$

where $n = 1, \dots, M$ and $k = 1, \dots, K$.

Next, the training facial images are projected onto the eigenspace, Ω_n , to generate representations of the facial images in Eigenface:

$$\Omega_n = [\varpi_{n1}, \varpi_{n2}, \dots, \varpi_{nK}], \quad (6)$$

where $n = 1, 2, \dots, M$.

In the operational stage, an incoming facial image is projected onto the same eigenspace, and the similarity measure, which is the Mahalanobis distance between the input facial image and the template, is computed in the Eigenspace.

Let φ_1^o denote the representation of the input face image with claimed identity C and let φ_1^c denote the representation of the C^{th} template. The similarity measure between φ_1^o and φ_1^c is defined as follows:

$$F_1(\varphi_1^o, \varphi_1^c) = \left\| \varphi_1^o - \varphi_1^c \right\|_m, \quad (7)$$

where $\|\bullet\|_m$ denotes the Mahalanobis distance.

2.2 Speaker Verification

The speaker verification module includes three important stages: endpoint detection, feature extraction, and pattern comparison. The endpoint detection stage aims to remove silent parts from the raw audio signal, as these parts do not convey speaker dependent information.

Noise reduction techniques are used to reduce the noise from the speech signal. Simple spectral subtraction [14] is first used to remove additive noise prior to endpoint detection. Then, in order to cope with channel distortion or convolution noise that is introduced by a microphone, the zero'th order cepstral coefficients are discarded, and the remaining coefficients are appended with delta feature coefficients [15]. In addition, the cepstral components are weighted adaptively to emphasize the narrow-band components and suppress the broadband components [16]. The cleaned audio signal is converted to 12th order linear prediction cepstral coefficients (LPCC), using the autocorrelation method, resulting in a 24-dimensional vector for every utterance. The significant improvement of verification rate achieved by using this combination is reported in paper [17]. Fig. 2 shows the process used in the front end module.

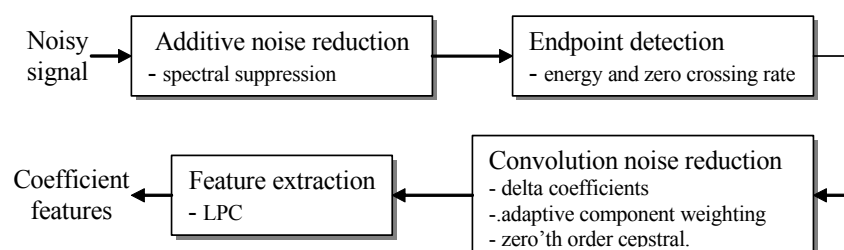


Fig. 2. The front-end of the speaker verification module.

As with the face recognition module, the speaker verification module also consists of two stages: (i) the training stage and (ii) the operational stage. In the training phase, two sample utterances with the same words from the same speaker are collected and trained using the modified k -Mean algorithm [15]. The main advantages of this algorithm are the statistical consistency of the generated templates and their ability to cope with a

wide range of individual speech variations in a speaker-independent environment.

In the operational stage, we employ for a well-known pattern-matching algorithm, Dynamic Time Warping (DTW) [15] to compute the distance between the trained template and the input sample.

Let φ_2^o represent the input speech sample with the claimed identity C , and let φ_2^c denote the representation of the C^{th} template. The similarity function between φ_2^o and φ_2^c is defined as follows:

$$F_2(\varphi_2^o, \varphi_2^c) = \|\varphi_2^o - \varphi_2^c\|, \quad (8)$$

where $\|\cdot\|$ denotes the distance score result obtained from DTW.

3. DECISION FUSION MODULE

In general, there are three possible levels of fusion when combining multiple biometric systems: One is fusion at the data extraction level, where raw data extracted using multiple sensors are concatenated. Another is fusion at the feature extraction level, which involves combining features extracted from raw measurements. For instance, linear prediction cepstral coefficients (LPCC) are extracted from speech signal, and projected face vectors as described in Eq. (5) are transformed from the normalized face image. The third level of fusion is at the measurement level. The output from each module is a set of possible labels with associated confidence values. In this case, more accurate decisions can be made by integrating different confidence measures into a more informative confidence measure [18].

In this paper, the third approach is adopted using the SBS to fuse the confidence measures. The output scores from Eqs. (7) and (8) are a measure of their respective belief in the acceptability of the identity claimed; the higher the scores, the higher the belief that the identity claimed is genuine, not an imposter. This allows for developing generic decision fusion rules, which are application independent and can alleviate the dimensionality problem. This reduction in dimensionality is beneficial since it comes along with a reduction in the number of training examples needed for training the different possible fusion modules.

3.1 Simple Bayesian Decision Theory in Bimodal Biometrics Decision Fusion

Bayes' theorem shows how to optimally predict the genuine (G) class or imposter (I) class, $\omega_i, i = 1, 2$, from the confidence scores, $F_j, j = 1, 2$, that are contributed by the face module and the speech module, respectively. The chosen class should be the one which maximizes the conditional probability $P(\omega_i | F_j)$ as shown below (it is called the maximum *a posteriori* (MAP) decision rule [19]):

$$P(\omega_i | F_j) = \frac{p(F_j | \omega_i)P(\omega_i)}{p(F_j)}, \quad (9)$$

where $p(F_j | \omega_i)$ is the conditional probability density function of the likelihood that a

vector F_j is observed in class ω_i , $p(F_j)$ is the unconditional probability density function for all classes, and $P(\omega_i)$ is the a priori probability of F_j . Since $p(F_j)$ does not depend on the class index, i , the MAP decision also can be stated as

$$\text{MAP} = \max_i p(F_j | \omega_i) P(\omega_i). \quad (10)$$

From the MAP rule, an equivalent rule is derived using Bayes' theorem and an assumption that the *a priori* probabilities of F_j , $P(\omega_i)$ are the same. This rule is called the maximum likelihood (ML) decision rule, in which a vector F_j is assigned to a class of highest $P(\omega_i | F_j)$:

$$\text{ML} = \max_i p(F_j | \omega_i). \quad (11)$$

Again, with the assumption that the confident scores F_j , $j = 1, 2$ are mutually independent in the class given, $p(F_j | \omega_i)$ can be decomposed into the product of $p(F_1 | G)$ and $p(F_2 | I)$, respectively, where F_1 and F_2 are the scores that are obtained from Eqs. (7) and (8). In addition, every $p(F_j | \omega_i)$ should be only an univariate density function. In other words, each biometric module will be considered separately rather than treated as a compound unit. This will make the computation and tabulation of $p(F_j | \omega_i)$ easier and more manageable.

The MAP and ML decision rules can be restated as

$$\text{MAP}_s = \max_i P(\omega_i) \prod_{k=1}^2 p(F_k | \omega_i). \quad (12)$$

and

$$\text{ML}_s = \max_i \prod_{k=1}^2 p(F_k | \omega_i). \quad (13)$$

In the literature, the above treatment leads to the Simple Bayesian decision scheme [19].

Note that the SBS can not only be applied to two biometric modalities as discussed in this paper, but it also can be extended to n biometric modalities without causing a high increment of the computation cost. The cost of computation can be measured based on the training time and the memory requirements.

3.1.1 Particularization using the univariate Gaussian density function (BUGD)

The structure of the simple Bayesian decision fusion scheme in Eqs. (12) and (13) is determined by the conditional densities $p(F_j | \omega_i)$ as well as by $P(\omega_i)$ for MAP_s . Of the various density functions that have been investigated, the univariate Gaussian density would be the first option:

$$p(F_j | \omega_i) = \frac{1}{\sqrt{2\pi}\sigma_{j,\omega_i}} e^{-\frac{1}{2}\left(\frac{F_j - \mu_{j,\omega_i}}{\sigma_{j,\omega_i}}\right)^2}, \quad (14)$$

where μ and σ are the mean and variance, respectively, F_j , $j = 1, 2$ are the speech and face confident scores, respectively, and ω_i , $i = 1, 2$ correspond to genuine and imposter classes, respectively.

To a large extent, the attention given to this function is due to its analytical tractability. In addition, one must consider the simplification that occurs in the problem of estimating parameters of a univariate Gaussian density function. The required computations are merely those of the sample mean and sample variance. These easily computed and easily updated statistics are contained in the biometrics training data set. The *a priori* probability $P(\omega_i)$ in Eq. (12) also can be obtained from the training data set. If equality of $P(\omega_i)$ for both classes are assumed, then the ML_s decision scheme in Eq. (13) may be used.

3.1.2 Particularization using the beta density function (BBD)

Besides the Gaussian density function, there are many other density functions that may be used to estimate the conditional densities, $p(F_j | \omega_i)$. However, not every density function can be applied because it may have poor estimators of its parameters, as in the case of the Cauchy distribution function. The key point in choosing the proper density lies in whether the density function contains simple *sufficient statistics* [19]. The *sufficient statistic* is a function that conveys all the information relevant to estimating the density function parameters. Therefore, it is appropriate to select a density function from the exponential family that contains simple *sufficient statistics*. Members of the exponential family include the Gaussian, Rayleigh, Beta, Maxwell and many other familiar distributions. Among these density functions, the Beta distribution is chosen due to its shape diversity and to the domain of its densities [0, 1]. The shape of the beta distribution is quite variable, depending on the values of the parameters, a and b , as illustrated in Fig. 3.

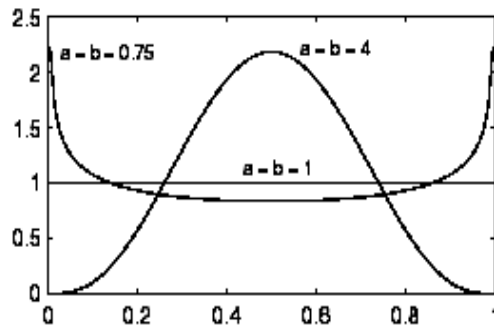


Fig. 3. Beta distribution.

The conditional densities, $p(F_j | \omega_i)$, can be modeled by using the Beta distribution as follows:

$$p(F_j | \omega_i) = \frac{\Gamma(a_{j,\omega_i} + b_{j,\omega_i})}{\Gamma(a_{j,\omega_i})\Gamma(b_{j,\omega_i})} F_j^{a_{j,\omega_i}} (1 - F_j)^{b_{j,\omega_i}} \quad \text{for } a > 0 \text{ and } b > 0, \quad (15)$$

where Γ is the gamma function, a and b are the shape parameters, $F_j = 1, 2$ are the speech and face confident scores, respectively, and $\omega_i = 1, 2$ correspond to genuine and imposter classes, respectively. Values of parameters a and b can be obtained by using Maximum likelihood estimation from training data sets [19].

3.1.3 Particularization using Parzen window density estimation (BPW)

The assumption of whether $p(F_j | \omega_i)$ is Gaussian or Beta may not lead to the optimal decision since it may deviate from the Gaussian or Beta distribution assumption. Hence, $p(F_j | \omega_i)$ can be anticipated by using Parzen windows density estimation with the Gaussian kernel [19].

In the Parzen windows (PW) technique, a density function (the kernel), $\phi(x)$, is used, for which $\phi(x) > 0$, $\int \phi(x) dx = V$. With sample F_{j,ω_i} , the estimated $\hat{p}(F_j | \omega_i)$ can be obtained from

$$\hat{p}_n(F_j | \omega_i) = \frac{1}{n_{\omega_i}} \sum_{k=1}^{n_{\omega_i}} \frac{1}{h_{n_{\omega_i}}^d} \phi\left(\frac{F_j - F_{k,\omega_i}}{h_{n_{\omega_i}}}\right), \quad (16)$$

where n_{ω_i} is the total amount of training data in the respective genuine and imposter populations and $h_n^d (= V)$ is a d -dimensional hypercube of length h_n .

Although PW is a theoretically sound technique, in practice, the amount of training data is limited; therefore, the kernel cannot be chosen too small (otherwise, “holes” and “spikes” will appear in the estimated probability density function). The optimal choice for the kernel width, h_n (often referred to as the smoothing-parameter), is still being researched. However, in practice, the Gaussian kernel is often used, and a width is chosen which optimizes the classification performance. In this paper, the kernel Gaussian $\phi(u) = (2\pi)^{-1/2} \exp(-u^2/2)$ is used.

3.2 Others Decision Fusion Schemes

3.2.1 Sum rule

The sum rule method of integration takes the weighted average of the individual score values. This strategy is applied to all possible combinations of two or more biometric modules. Equal weights are assigned to each modality, since the bias of each matcher is not available.

3.2.2 Multilayer perceptron (MLP)

An MLP is a neural classifier that separates the training data of the two classes by implementing a separation surface, which can have any arbitrary flexible shape. The

flexibility of the separating surface is determined by the complexity of the architecture. In this study, we used MLP with two neurons in the input layer (two scores coming from two module biometrics), 5 neurons in the hidden layer, and one neuron (two classes) in the output layer, and we used sigmoidal activation functions for all neurons and the Backpropagation training algorithm. With sigmoidal activation functions, the value of the output neuron lies in the interval $[0, 1]$, and the optimal decision threshold was fixed at 0.5.

4. EXPERIMENTS AND DISCUSSION

4.1 Distance Score Normalization

Normalization typically involves mapping the similarity measure values obtained from multiple domains into a common framework before combining them. In this case, the similarity measure values from Eqs. (7) and (8) cannot be fused directly since they have different ranges. They have to be mapped into the common score interval $[0, 1]$.

The opinions from the each of the biometrics modules are used in a fusion stage. The system considers the opinions and makes a final decision to either accept or reject the claim. The bimodal biometric system proposed here is designed as shown in Fig. 4.

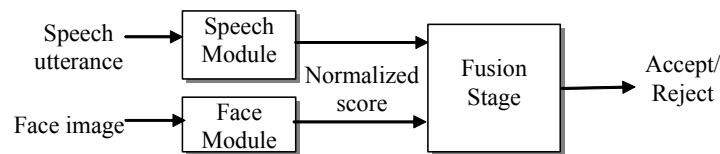


Fig. 4. The building blocks of the bimodal face and speech verification system.

From the distance scores, F_j , $j = 1, 2$ produced by the speech and face databases, respectively, the mean, μ , and the variance, σ^2 , of the distance values of the speech and face experts are found by performing validation experiments on the database. The distance score is then normalized by mapping it to the range $[-1, 1]$ using

$$y = \frac{F_j - \mu}{\sigma}. \quad (17)$$

The $[-1, 1]$ interval corresponds to the approximately linearly changing portion of the sigmoid function used to map the values to the $[0, 1]$ interval:

$$f(F_j) = \frac{1}{1 + \exp(-F_j)}. \quad (18)$$

Fig. 5 shows the distribution plot for the genuine and imposter reference points obtained for the system using mapping. The plot indicates that the two distributions are reasonably well separated in 2-dimensional space; therefore, the mutually independent assumption between the face and speech models is appropriate.

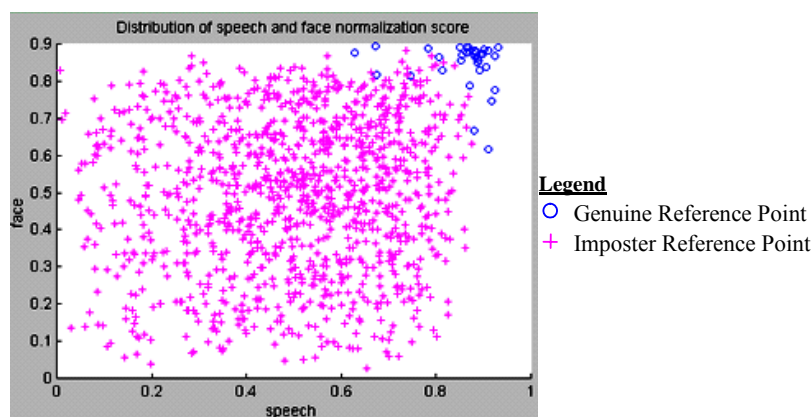


Fig. 5. The distribution plot for the genuine and imposter reference points.

4.2 Performance Criteria

The basic error measures of a verification system are the false acceptance rate (FAR) and false rejection rate (FRR) as defined in Eqs. (19) and (20):

$$\text{FAR} = \frac{\text{Number of rejected genuine claims}}{\text{Total number of genuine accesses}} \times 100\%. \quad (19)$$

$$\text{FRR} = \frac{\text{Number of rejected genuine claims}}{\text{Total number of genuine accesses}} \times 100\%. \quad (20)$$

A unique measure can be obtained by combining these two errors into the total error rate (TER) or total success rate (TSR), where

$$\text{TER} = \frac{\text{FAR} + \text{FRR}}{\text{Total number of accesses}} \times 100, \quad (21)$$

$$\text{TSR} = 100 - \text{TER}. \quad (22)$$

For the individual experts, the equal error rate EER (i.e FAR = FRR) is used to obtain their corresponding threshold values.

In practical applications that use bimodal biometric verification systems, the Minimum Total Misclassification Error (MTME) criterion is used, which means that the system always tries to minimize ε as shown in Eq. (23) [20]:

$$\varepsilon = \min(\text{FA} + \text{FR}). \quad (23)$$

In order to apply this criterion, FAR < 0.1% is set while the FRR is kept to a minimum possible value.

4.3 Experimental Setup

All the experiments were performed using a face database obtained from the Olivetti Research Lab (ORL) [21] and Digit multi-modal Database [22]. Obviously, ORL face database does not have corresponding speech samples, so to each face image, an arbitrary but fixed speech class from the Otago speech database [23] was assigned, which yielded 60 subjects in total. For face experts, six face images from a subject were randomly selected to be trained and projected into Eigenspace, and the other two samples were used for the subsequent validation and testing. Similarly, three samples were used in speech experts for the modeling (training); two samples were used for the subsequent validation and testing. Three sessions of the face database and speech database were used separately. The first enrollment session was used for training. This means that each access was used to model the respective genuine, yielding 60 different genuine models. In the second enrollment session, the accesses from each person were used to generate the validation data in two different ways. The first way was to derive a single genuine access by matching the shot or utterance template of a specific person with his own reference model, and the other way was to generate 59 impostor accesses by matching it with the 59 models of the other persons of the database. This simple strategy thus resulted in 60 genuine and 3540 impostor accesses, which were used to validate the performance of the individual verification system and to calculate the thresholds for the EER criterion and the parameters for the fusion decision schemes. The third enrollment session was used to test these verification systems, using the thresholds calculated with the validation data set.

4.4 Experimental Results

The performance results for the speech and face expert are shown in Table 1.

Table 1. Individual performance of the face and speech expert.

Biometric Module	FRR(%)	FAR(%)	TER(%)	TSR(%)
Speech	8.33	6.50	6.53	93.47
Face	5.00	6.02	6.00	94.00

From the TSR values shown in Table 1, we can observe that the experts worked equally well individually.

Table 2. Results for BUGD.

	FRR(%)	FAR(%)	TER(%)	TSR(%)
MAP _s	21.67	0.00	0.36	99.64
ML _s	5.00	0.51	0.58	99.42

Table 2 shows the results that were obtained by applying the univariate Gaussian density function (BUGD). For MAP_s, the *priori* probability for the genuine class was $P(\omega_1) = 1/60$ and for imposter class was $P(\omega_2) = 59/60$. For the ML_s decision scheme, $P(\omega_1) = P(\omega_2)$ was assumed. The parameter values of the mean, μ , and variance, σ^2 , were obtained from the data in the validation session.

From Table 2, it can be observed that MAP_s outperformed ML_s in terms of TSR. However, ML_s showed more balanced results in terms of FRR and FAR. Since $P(\omega_2) > P(\omega_1)$ for MAP_s, FAR decreased and FRR increased. Changing the ML_s rule into a MAP_s led to more rejections of genuine claims, which is not desirable for genuine users (thus the increase in FRR). In spite of MAP_s fulfilling the MTME criteria with FAR < 0.1% and achieved a high TSR of 99.64%, the high FRR value (21.67%) was unfavorable for real applications.

For MAP_s and ML_s decision fusion schemes that use the Beta distribution function, the parameters a_{j,ω_j} and b_{j,ω_j} in Eq. (15) could be estimated from the training data using the maximum likelihood technique, and the values are shown in Table 3.

Table 4 shows the results; MAP_s still outperformed ML_s when the Beta distribution was applied. MAP_s fulfilled the MTME criteria, in that FAR < 0.1% and the FRR value was also acceptable.

Table 3. Estimation of Beta distribution parameters.

Parameters	Biometric Module	Class Population	Values
a	Speech	Genuine	73.79
		Imposter	2.54
	Face	Genuine	72.89
		Imposter	2.50
b	Speech	Genuine	8.83
		Imposter	2.62
	Face	Genuine	13.74
		Imposter	2.56

Table 4. Results from BBD.

	FRR(%)	FAR(%)	TER(%)	TSR(%)
MAP _s	5.67	0.09	0.19	99.81
ML _s	3.33	0.68	0.72	99.28

To apply the simple Bayesian decision fusion scheme, which particularizes Parzen windows density estimation (BPW), the optimum kernel width, h_n , in Eq. (16) has to be estimated. However, a criterion for determining an optimum h_n is rather subjective and application dependent; therefore, the MTME criteria was used here as guidance to get an optimum result. From the result of experiments conducted in the training session, $h_n = 5$ for MAP_s and $h_n = 25$ for ML_s were taken and applied in the testing session. The obtained results are shown in Table 5.

Table 5. Results for BPWD.

	FRR(%)	FAR(%)	TER(%)	TSR(%)
MAP_s	3.33	1.13	1.17	98.83
ML_s	30.00	0.79	1.28	98.72

The results shown in the Table 5 shows that MAP_s performed better than ML_s , but that neither of them satisfied the MTME criteria. The overall performance also was poor if compared with that of BUGD and BBD. This indicates the difficulty of estimating an unknown underlying probability density function due to (1) the small number of sample data, only 60, for the genuine conditional probability density function $p(F_j | \omega_1)$, (2) and the poor generalization caused by h_n .

The other fusion methods described in section 3.2 were also used, and the results are shown in Table 6.

Table 6. Results for the other fusion methods.

Method	FRR(%)	FAR(%)	TER(%)	TSR(%)
Sum Rule	1.67	1.16	1.17	98.83
MLP	8.33	0.14	0.28	99.72

Table 6 shows that the sum rule method does not meet the MTME criteria and has poor performance in terms of TSR but not MLP.

In general, all the fusion techniques outperformed both of the individual modal experts. SBS with the Beta density function performed best using the MAP decision rule. This indicates that the assumption that genuine and imposter populations are Beta distributed is appropriate, since the shape factors a and b in the Beta density function can yield a rich variety of shapes. Its domain bound in the interval 0-1 also works well in the experimental setting. This enables SBS to deliver the optimal decision in the Bayesian sense. However, the results obtained when univariate Gaussian distribution was assumed are also relatively good.

It is interesting to compare MLP with SBS with the univariate Gaussian density function and Beta density function. MLP uses a brute force approach to directly estimate the underlying probability distribution. This is appealing because it can lead to the optimum decision and does not rely on any assumption. However, training MLP may not always

produce the optimal estimation of $p(F_j | \omega_i)$. Good tuning of the various parameters during the validation phase is required. Therefore, the classifier performance will deteriorate in terms of the training time, training speed, and storage space. Furthermore, if extra biometrics modals are added into the system, the cost of computation starts to increase exponentially. On the other hand, SBS with appropriate particularizations only needs a few parameters to be estimated (only two for BUGD and BBD). Regardless of how many biometric modules are present in the system, it is still able to achieve a very high verification rate. Table 7 compares the training and recognition times of BBD and MLP.

Table 7. Computation speed comparison between BBD and MLP.

Approach	Training time	Recognition Time
BBD	< 2 seconds *	< 2 seconds *
MLP	10 minutes *	< 5 seconds *
* The experiments were conducted on a 800MHz IBM compatible PC with 256 RAM		

5. CONCLUSIONS

The paper has presented a decision fusion technique that employs the Simple Bayesian Structure (SBS) with three chosen particularizations: the univariate Gaussian density function, Beta density function, and Parzen window density estimation. SBS has been shown to possess advantages in terms of computation speed, storage space, and its open framework.

From the experiments, it has been found that the best result is obtained by using SBS with the particularized Beta density function as this leads to lower FAR and FRR, compared to other SBS schemes that particularize the univariate Gaussian and estimated density function by using the Parzen window as well as other classical fusion schemes.

REFERENCES

1. A. Jain, R. Bolle, and S. Pankanti, *Biometrics: Personal Identification in Networked Society*, 2nd Printing, Kluwer Academic Publishers, 1999.
2. R. Brunelli and D. Falavigna, "Personal identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, 1995, pp. 955-966.
3. U. Dieckmann, P. Plankensteiner, and T. Wagner, "Sesam: a biometric person identification system using sensor fusion," *Pattern Recognition Letters*, Vol. 18, 1997, pp. 827-833.
4. B. Duc, G. Maýtre, S. Fischer, and J. Bigun, "Person authentication by fusing face and speech information," in *Proceedings of 1st International Conference on Audio- and Video-based Biometric Person Authentication*, LNCS 1206, Springer Verlag, 1997, pp. 311-318.
5. E. Bigun, J. Bigun, B. Duc, and S. Fisher, "Expert conciliation for multi modal person authentication systems by Bayesian statistics," in *Proceedings of 1st*

- International Conference on Audio- and Video-based Biometric Person Authentication*, 1997, pp. 327-334.
6. J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, 1998, pp. 226-239.
 7. L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, 1998, pp. 1295-1307.
 8. S. Ben-Yacoub, "Multimodal data fusion for person authentication using SVM," IDIAP-RR 7, IDIAP, 1998.
 9. T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in *Proceedings of 2nd International Conference on Audio and Video-based Biometric Person Authentication*, 1999, pp. 176-181.
 10. A. Ross, A. K. Jain, and S. Pankanti, "Information fusion in biometrics," in *Proceeding of 3rd International Conference on Audio and Video-based Biometric Person Authentication (AVBPA '01)*, 2001, pp. 354-359.
 11. Y. H. Wang, T. N. Tan, and A. K. Jain, "Combining face and iris biometrics for identity verification," in *Proceedings of 4th International Conference on Audio and Video-based Biometric Person Authentication (AVBPA '03)*, pp. 805-813.
 12. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuro Science*, Vol. 3, 1991, pp. 71-86.
 13. S. A. Samad, A. Hussein, and A. Teoh, "Eye detection using hybrid rule based method and contour mapping," in *Proceeding of 6th International Symposium on Signal Processing and Its Applications*, 2001, pp. 631-634.
 14. R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of 7th European Signal Processing Conference*, 1994, pp. 1182-1185.
 15. L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, United State: Prentice-Hall International, Inc., 1993.
 16. M. S. Zilovic, R. P. Ramachandran, and R. J. Mammone, "A fast algorithm for finding the adaptive component weighting cepstrum for speaker recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, 1997, pp. 84-86.
 17. S. A. Samad, A. Hussein, and A. Teoh, "Increasing robustness in a speaker verification system with template training and noise reduction techniques," in *Proceedings of the International Conference on Information Technology and Multimedia*, 2001.
 18. B. V. Dasarathy, *Decision Fusion*, IEEE Computer Society Press, 1994.
 19. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, Inc., New York, 2001.
 20. C. Sanderson and K. K. Paliwal, "Training method of a piecewise linear classifier for a multi-modal person verification system," in *Proceedings of 8th Australian International Conference on Speech Science and Technology*, 2000, pp. 312-317.
 21. Database of faces, <http://www.cam-orl.co.uk/facedatabase.html>.
 22. C. Sanderson, Digit Database 1.0 – multimodal database for speaker identification/recognition, <http://spl.me.gu.edu.au/digit/>.
 23. Otago Speech Corpus, <http://kel.otago.ac.nz/hyspeech/corpusinfo.html>.



Andrew Beng Jin Teoh obtained his BE (Electronic) in 1999 and Ph.D. degree in 2003 from National University of Malaysia. He is currently an associate dean and lecturer of Faculty of Information Science and Technology, Multimedia University. He held the post of co-chair (Biometrics Division) in Center of Excellent in Biometrics and Bioinformatics in the same university. His research interests are in multimodal biometrics, pattern recognition, multimedia signal processing, and Internet security.



Salina Abdul Samad obtained her Bachelor of Science in Electrical Engineering from the University of Tennessee, U.S.A. and a Ph.D. from the University of Nottingham, England. Her research interest is in the field of digital signal processing, from algorithm design to software and hardware implementation. She is now employed by Universiti Kebangsaan Malaysia as an Associate Professor.



Aini Hussain received the B.S. (Electrical) from Louisiana State University, U.S.A.; M.S. (Systems & Control) from UMIST, England and Ph.D. from Universiti Kebangsaan Malaysia in 1985, 1989 & 1997, respectively. She is currently an Associate Professor in the EESE Dept. at Universiti Kebangsaan Malaysia. Her research interests include signal processing, pattern recognition, and soft computing.