A Multimodal Biometric System Using Fingerprint, Face, and Speech

Anil Jain, Lin Hong, and Yatin Kulkarni
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824-1226
{jain,honglin,kulkar10}@cse.msu.edu

Abstract

A biometric system which relies only on a single biometric identifier in making a personal identification is often not able to meet the desired performance requirements. Identification based on multiple biometrics represents an emerging trend. We introduce a multimodal biometric system, which integrates face recognition, fingerprint verification, and speaker verification in making a personal identification. This system takes advantage of the capabilities of each individual biometrics. It can be used to overcome some of the limitations of a single biometrics. Preliminary experimental results demonstrate that the identity established by such an integrated system is more reliable than the identity established by a face recognition system, a fingerprint verification system, and a fingerprint verification system.

1 Introduction

In today's electronically wired information society, there are an increasing number of situations (*e.g.*, accessing a multiuser computer account) which require an individual, as a user, to be verified by an electronic device. Traditionally, a user can be verified based on whether she is in possession of a certain *token* such as an ID card ("something that she has") and/or whether she is in possession of specific *knowledge* which only she herself is expected to know such as a *password* ("something that she knows"). These approaches have a number of significant drawbacks. Tokens may be lost, stolen, forgotten, forged, or misplaced. Passwords may be forgotten or compromised. All these approaches are unable to differentiate between an *authorized user* and an *imposter* who fraudulently acquires the "token" or "knowledge" of the authorized user. Therefore, token or knowledge-based authentication does not provide sufficient security in many critical applications involving access control and financial transactions.

Biometrics, which refers to the automatic identification of a person based on her physiological or behavioral characteristics, relies on "something which she is or she does" (*e.g.*, putting her finger on a scanner) to make a personal identification [7]. It is inherently more reliable and has a higher discrimination capability than the token-based and/or knowledge-based approaches, because the physiological or behavioral characteristics are unique to each user. Currently, nine different biometric indicators are either widely used or are under intensive evaluation, including *face, facial thermogram, fingerprint, hand geometry, hand vein, iris, retinal pattern, signature, and voice-print.* All these biometric indicators have their own advantages and disadvantages in terms of the accuracy, user acceptance, and applicability. It is the requirements of an application domain which determine the choice of a specific biometric indicator. In order to enable a biometric system to operate effectively in different applications and environments, a *multimodal biometric system* which makes a personal identification based on multiple physiological or behavioral characteristics is preferred. Consider, for example, a network logon application where a biometric system is used for user authentication. If a user cannot provide good fingerprint images (*e.g.*, due to dry finger, cuts, *etc.*) then face and voice may be better biometric indicators. If the operating environment is "noisy" then voice is not a suitable biometric indicator. If the "background" is cluttered, then the face location algorithm, which is necessary for face recognition, may not work very well.

Some work on multimodal biometric systems has already been reported in the literature. Dieckmann *et al.* [5] have proposed an abstract level fusion scheme: "2-from-3 approach" which integrates face, lip motion, and voice based on the principle that a human uses multiple clues to identify a person. Brunelli and Falavian [2] have proposed a measurement level scheme and a hybrid rank/measurement level scheme

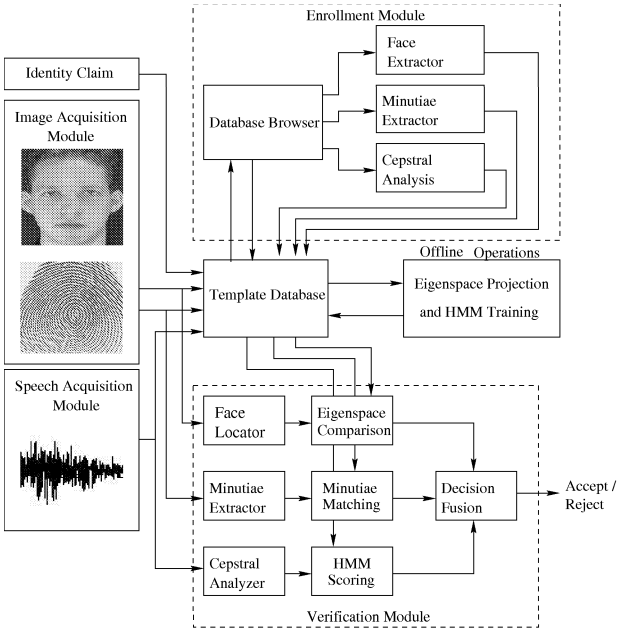

Figure 1: The block diagram of our multimodal biometric system.

to combine the outputs of sub-classifiers. Kittler *et al.* [10] have demonstrated the efficiency of an integration strategy which fuses multiple snapshots of a single biometric property using a Bayesian framework. Bigun *et al.* [1] have proposed a Bayesian integration scheme to combine different pieces of evidence. Maes *et al.* [11] have proposed to combine biometric data (*e.g.*, voice) with non-biometric data (*e.g.*, password). Hong and Jain [6] have developed a multimodal *identification* system which integrates two different biometrics (face and fingerprint) that complement each other.

We are interested in designing a multimodal biometric system which integrates face, fingerprint, and speech to make a personal identification. Our choice of these three specific biometrics is based on the fact that they have been used routinely in law enforcement community. Most of the successful commercial biometric systems currently rely on either fingerprint, face, or voice. Further, these biometric indicators complement one another in their advantages and strengths. While fingerprint provides an extremely high verification accuracy, it is difficult for an untrained human to match fingerprints. Face and speech, on the other hand, are routinely used by all of us in our daily recognition tasks. Our system is targeted for verification applications to authenticate the identity claimed by a user such as in a multiuser account authentication. The block diagram of our system is shown in Figure 1, which mainly consists of four components: (*i*) acquisition module, (*ii*) template database, (*iii*) enrollment module, and (*iv*) verification module. The acquisition module is responsible for acquiring fingerprint images, face images, and speech signal of a user who intends to access the system. The template database is a physical database which contains all the template records of the users who are enrolled in the system. The task of the enrollment module is *system management* which includes user enrollment, user deletion, user update, training, system parameter specification, *etc.* The verification module is responsible for authenticating the identity claimed by a user at the point-of-access. The verification process essentially consists of four stages: (*i*) fingerprint verification, (*ii*) face recognition, (*iii*) speaker verification, and (*iv*) decision fusion. Fingerprint verification is responsible for matching the input fingerprint against the fingerprint template(s) stored in the database to obtain the fingerprint matching score. Face recognition is responsible for matching the input face against the face template to obtain the face matching score. Speaker verification is responsible for obtaining the matching score of the input speech signal. The decision fusion integrates the matching scores from fingerprint verification, face recognition, and speaker verification to establish the final decision.

2 The Multimodal Biometric System

Each individual biometrics in our multimodal system has a very different characteristic and a different matching scheme. Therefore, it is more reasonable to integrate the multiple biometrics at the decision level instead of at the sensor level.

2.1 Formulation

Let $\mathcal{B}$ denote a given biometric system, and let $\mathbf{\Phi^1}, \mathbf{\Phi^2}, ..., \mathbf{\Phi^N}$ denote the templates of the N users enrolled in $\mathcal{B}$, who are labeled by numerical indicators, $1, 2, ..., N$. Assume, for simplicity, that each enrolled user has only one template (for each type of indicator) stored in the system. So the template for the *ith* user, $\mathbf{\Phi^i} = \{\Phi_1^i, \Phi_2^i, \Phi_3^i\}$, has three components, where $\Phi_1^i, \Phi_2^i, \Phi_3^i$ are the templates for fingerprint, face, and speech biometrics, respectively. Let $(\mathbf{\Phi^0}, I)$ denote the biometric indicator and the identity claimed by a user. Again $\mathbf{\Phi^0}$ has three components, $\mathbf{\Phi^0} = \{\Phi_1^0, \Phi_2^0, \Phi_3^0\}$, corresponding to the measurements of the three biometric indicators. The claimed identity, I, either belongs to category w_1 or category w_2, where w_1 indicates that the user claims a true identity (a genuine user) and w_2 indicates that the user claims a false identity (an impostor). The biometric system $\mathcal{B}$ matches $\mathbf{\Phi^0}$ against $\mathbf{\Phi^I}$ to determine which category, w_1 or w_2,

the claimed identity I falls in, *i.e.*

$$I \in \begin{cases} w_1, & if\ \mathcal{F}(\mathbf{\Phi^0}, \mathbf{\Phi^I}) > \epsilon, \\ w_2, & otherwise, \end{cases} \quad (1)$$

where $\mathcal{F}(\mathbf{\Phi^0}, \mathbf{\Phi^I})$ is a function which measures the similarity between $\mathbf{\Phi^0}$ and $\mathbf{\Phi^I}$ and ϵ is a threshold.

For a claimed identity I which can be in either w_1 or w_2, the biometric system may determine whether I is in w_1 or w_2. Therefore, there are a total of four possible outcomes: (i) a claimed identity in w_1 is determined to be in w_1, (ii) a claimed identity in w_1 is determined to be in w_2, (iii) a claimed identity in w_2 is determined to be in w_2, and (iv) a claimed identity in w_2 is determined to be in w_1. Outcome (i) corresponds to a genuine user being accepted, outcome (ii) corresponds to a genuine user being rejected, outcome (iii) corresponds to an impostor being rejected, and outcome (iv) corresponds to an impostor being accepted. Obviously, outcomes (i) and (iii) are correct whereas outcomes (ii) and (iv) are incorrect. Ideally, a biometric system should make only correct decisions. In practice, due to large *intraclass* variations in the acquired digital representation of the biometric indicator, incorrect decisions are unavoidable. Typically, (i) *false acceptance rate* (FAR) and (ii) *false reject rate* (FRR) are used to characterize the performance of a biometric system. The false acceptance rate corresponds to the probability of outcome (iv) and the false reject rate is defined as the probability of outcome (ii). The lower the values of the FAR and FRR, the more reliable is the decision made by the system. The FAR and FRR values of a given biometric system are determined by the inherent *interclass* and *intraclass* variations of the indicator and the design (*e.g.*, feature extraction, decision making) of the system.

2.2 Fingerprint Verification

A *fingerprint* is the pattern of ridges and furrows on the surface of a fingertip. The uniqueness of a fingerprint is exclusively determined by the local ridge characteristics and their spatial relationships. The two most prominent ridge characteristics, called minutiae, are (i) *ridge ending* and (ii) *ridge bifurcation*. Fingerprint verification depends on the comparison of minutiae and their relationships to make a personal identification, which usually consists of two stages [8]: (i) minutiae extraction and (ii) minutiae matching. The minutiae extraction module extracts minutiae from input fingerprint images and the minutiae matching module determines the similarity of two minutiae patterns.

Let Φ_1^0 denote the minutiae pattern extracted from the input fingerprint image with claimed identity I and Φ_1^I be the Ith fingerprint template stored in the database. The similarity function between an input fingerprint Φ_1^0 and a template Φ_1^I is defined as follows:

$$\mathcal{F}_1(\Phi_1^0, \Phi_1^I) = \frac{100C^2}{PQ}, \quad (2)$$

where P and Q are the total number of minutiae in Φ_1^0 and Φ_1^I, respectively and C is the total number of *corresponding* minutiae pairs between Φ_1^0 and Φ_1^I established by the minutiae matching algorithm [8].

2.3 Face recognition

In personal identification, *face recognition* refers to static, controlled full frontal portrait recognition. There are two major tasks in face recognition: (i) face location and (ii) face recognition. Face location finds whether there is a face in the input image and if so, the location of the face in the image. Face recognition finds the similarity between the located face and the stored templates to determine the identity of the user. A number of face recognition approaches have been reported in the literature [4]. The performance of some of the proposed face recognition approaches is very impressive. In our system, the eigenface approach [9] is used.

The eigenface-based face recognition method is divided into two stages: (i) training stage and (ii) operational stage. In the training stage, a set of orthonormal images that best describe the distribution of the training facial images in a lower dimensional subspace (eigenspace) is computed. Then, the training facial images are projected onto the eigenspace to generate the representations of the facial images in the eigenspace. In the operational stage, a detected facial image is projected onto the same eigenspace and the similarity between the input facial image and the template is, thus, computed in the eigenspace. Let Φ_2^0 denote the representation of the input face image with claimed identity I and Φ_2^I denote the representation of the Ith template. The similarity function between Φ_2^0 and Φ_2^I is defined as follows:

$$\mathcal{F}_2(\Phi_2^0, \Phi_2^I) = -\|\Phi_2^I - \Phi_2^0\|, \quad (3)$$

where $\| \bullet \|$ denotes the L_2 *norm*.

2.4 Speaker Verification

Anatomical variations that naturally occur amongst different people and the differences in their learned speaking habits manifest themselves as differences in the acoustic properties of the speech signal. By analyzing and identifying these differences, it is possible to discriminate among speakers [3]. We have implemented a text-dependent

speaker recognition system, which uses the left-to-right *hidden markov model* (HMM) of the 10th order linear prediction coefficients (LPC) of the cepstrum to make a verification [3]. Input to the system consists of a random combination of four spoken digits (1,2,7 and 9) visually prompted to the speaker on a video monitor.

Let $\Phi_3^0 = s[1:L]$ denote the input feature vector of length L and Φ_3^I denote the Ith template which is characterized by $q(s|t_i)$ and $p(t_i|t_{i-1})$. The similarity function between the input Φ_3^0 and the Ith template Φ_3^I is defined as:

$$\mathcal{F}_3(\Phi_3^0, \Phi_3^I) = \log\{p(s(1:L)|\Phi_3^I)\} \quad (4)$$

$$= \max_{i_1 \ldots i_L} \{\prod_{k=1}^{L} q(s_{i_k}|t_{i_k}) p(t_{i_k}|t_{i_{k-1}})\}, \quad (5)$$

where p and q are the conditional probabilities used to characterize Φ_3^I.

Note that other approaches to face recognition and speaker verification could be used as well. Our choice of these specific face recognition and speaker verification algorithms was determined by what was available in our laboratory. The main purpose of this paper is to demonstrate the improvement in the overall system performance by integrating multiple biometric indicators. As such, we have not made any attempts to optimize the performance of individual biometrics.

2.5 Decision Fusion

The final decision made by our system is based on the integration of the decisions made by the fingerprint verification module, the face recognition module, and the speaker verification module. If the output of each module is only a category label, either w_1 (claimed identity is true) or w_2 (claimed identity is not true), which is not associated with any confidence value, then the integration of these multiple decisions can only be performed at an *abstract level*, in which a majority rule can be employed to reach a more reliable decision [12]. If the output of each module is a similarity value, then a more accurate decision can be made at a rank level or at a measurement level by accumulating the confidence associated with each individual decision.

Let X_1, X_2, and X_3 be the random variables used to indicate the similarity (dissimilarity) between an input and a template for fingerprint verification, face recognition, and speaker verification, respectively. Let $p_j(X_j|w_i)$, where $j = 1, 2, 3$ and $i = 1, 2$, be the class-conditional probability density functions of X_1, X_2, and X_3. Assume that X_1, X_2, and X_3 are statistically independent. Then, the joint class-conditional probability density function of X_1, X_2, and X_3, has the following form:

$$p(X_1, X_2, X_3|w_i) = \prod_{j=1}^{3} p_j(X_j|w_i), \quad i = 1, 2. \quad (6)$$

Depending on the application requirement on verification accuracy, any one of a number of different statistical decision theory frameworks can be used. In biometrics, the performance requirement is usually specified in terms of the FAR. In this case, the decision fusion scheme should establish a decision boundary which satisfies the FAR specification and minimizes the FRR. Let R^3 denote the three-dimensional space spanned by (X_1, X_2, X_3); R_1^3 and R_2^3 denote the w_1-region and-w_2 region, respectively $(R_1^3 + R_2^3 = R^3)$; ϵ_0 denote the pre-specified FAR. According to the Neyman-Pearson rule, a given observation, $X^0 = (X_1^0, X_2^0, X_3^0)$, is classified as:

$$(X_1^0, X_2^0, X_3^0) \in \begin{cases} w_1, & \text{if} \frac{p_1(X_1^0, X_2^0, X_3^0|w_1)}{p_2(X_1^0, X_2^0, X_3^0|w_2)} > \lambda \\ w_2, & otherwise, \end{cases} \quad (7)$$

where λ is the minimum value that satisfies the following

$$\lambda = \frac{p_1(X_1, X_2, X_3|w_1)}{p_2(X_1, X_2, X_3|w_2)} \; and \quad (8)$$

$$\epsilon_0 = \int_{R_1} p_2(X_1, X_2, X_3|w_2) dX_1 dX_2 dX_3. \quad (9)$$

3 Performance Evaluation

The performance benchmark assesses the capability of the system at the point-of-identification, which depends heavily on how the system is used, whether the users are willing to cooperate, *etc.* A test which simulates the operating environment is needed to assess the performance benchmark of an implemented system. We have evaluated the performance of our multimodal biometric system on a small set of data which is acquired in a laboratory environment.

3.1 Databases

A training database of fingerprints, faces, and speech samples of 50 users was collected. For each user, 10 fingerprint images (a total of 500 images), 9 face images (a total of 450 images), and 12 speech samples (a total of 600 samples) were acquired. The fingerprint images were acquired using an optical fingerprint scanner manufactured by Digital Biometrics with the restriction that fingers be placed approximately at the center of the scanner and the orientation of fingers be within 90^o. The face images were acquired using a

Panasonic video camera under normal indoor lighting conditions. The rotation of the face was restricted to $[-30^o, +30^0]$ and the scaling factor was allowed to be in the interval $[0.90, 1.10]$. The speech samples were collected using Labtec microphone in a laboratory environment. Examples of acquired fingerprint images, face images, and speech waveforms are shown in Figures 2. A test database involving 25 users (a subset

(a) (b) (c)

Figure 2: Fingerprint, face, and speech samples.

of the 50 people in the training set) that were available during the second round of data collection was collected. For each user, 15 fingerprint images (a total of 375 images), 15 face images (a total of 375 images), and 15 speech samples (a total of 375 samples) were collected over 3 sessions which lasted for about two weeks.

3.2 Benchmarks

Neither the genuine distribution nor the impostor distribution for each individual biometric indicator can be precisely formulated by a known statistical model. They need to be estimated from empirical data. In our test, an "all against all" verification test on the training database was used to generate the genuine and impostor distributions; each distribution was discretized into 100 bins. Figure 3 shows the genuine and impostor distributions estimated from the training database. With the estimated distributions, the decision boundary that satisfies a given FAR specifications is derived according to the Neyman-Pearson rule described in section 2.

Since the pre-specified FAR for a biometric system is usually very small (< 0.001), we need a large number of representative samples to demonstrate that a biometric system does meet such a performance specification. Unfortunately, obtaining a large number of test samples is both expensive and time consuming. Our test database consists of only a very limited number of samples. In order to overcome this *insufficient-testing-sample* problem, we use different assignment practices - each time, a different fingerprint, a different face, and a different speech sample are combined to form a *probe*. Obviously, such a scheme might result in unjustified performance improvement. How-

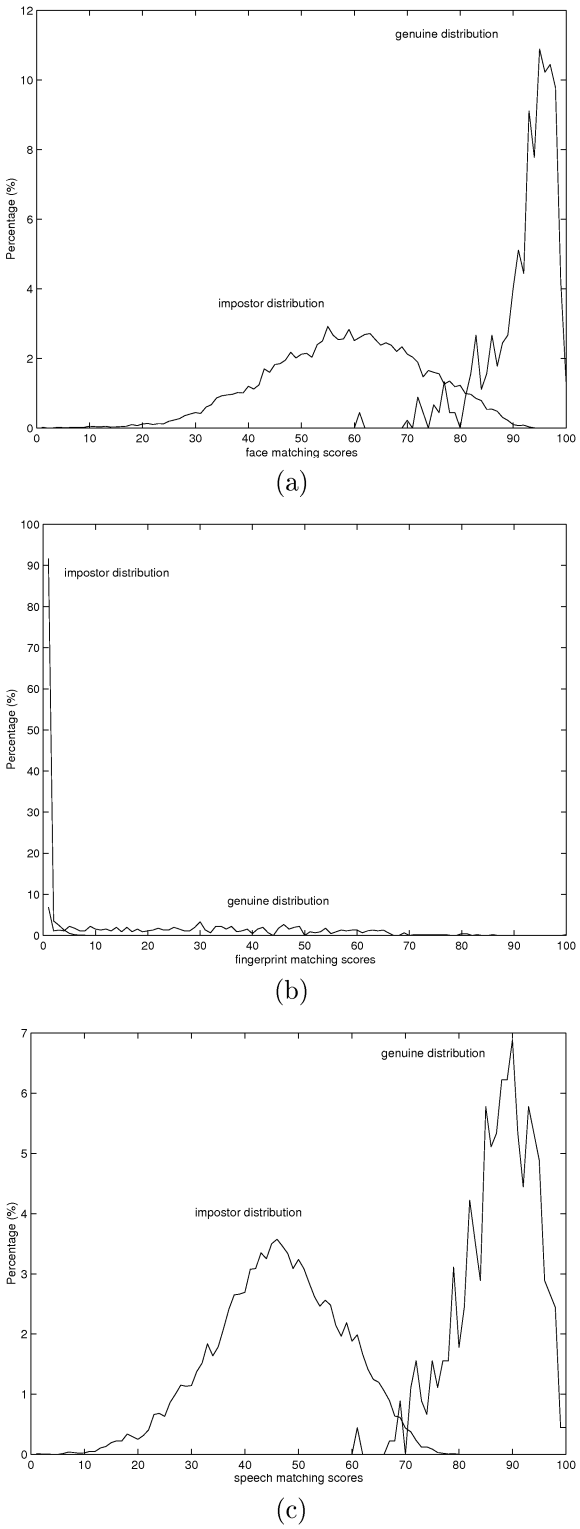

Figure 3: Genuine and impostor distributions of face recognition, fingerprint verification, and speaker verification; matching scores are normalized to $[0, 100]$.

ever, since the three biometric indicators are independent, such a scheme appears to be reasonable. In our test, a total of 36,796 impostor probes and 358 genuine probes were generated and tested. The receiver operating curve of decision fusion, along with the receiver operating curves of face recognition, fingerprint verification, and speaker verification are plotted in Figure 4, in which the authentic acceptance rate (the percentage of genuine individuals being accepted, *i.e.*, $1 - FRR$) is plotted against FAR. We can conclude from these test results that the integration of fingerprint, face and speech leads to an improvement in verification performance.

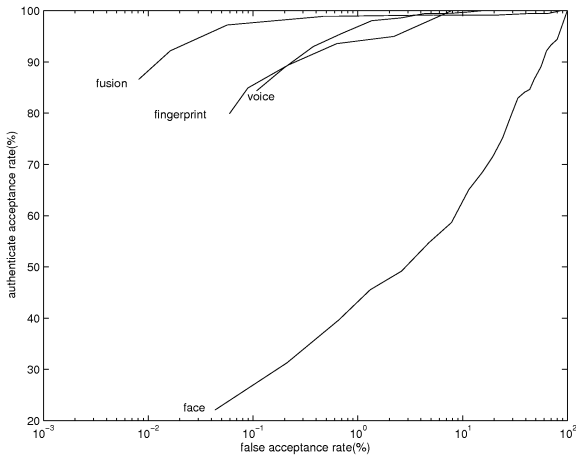

Figure 4: Receiver Operating Curves using Neyman-Pearson rule.

4 Summary and Conclusions

A *multimodal biometrics* technique, which combines multiple biometrics in making a personal identification, can be used to overcome the limitations of individual biometrics. We have developed a multimodal biometric system which integrates decisions made by face recognition, fingerprint verification, and speaker verification to make a personal identification. In order to demonstrate the efficiency of such an integrated system, experiments which simulate the operating environment on a small data set which is acquired in a laboratory environment were performed. The experimental results show that our system performs very well. However, the system needs to be tested on a large dataset in a real operating environment.

References

[1] E. S. Bigun, J. Bigun, B. Duc, and S. Fischer. Expert conciliation for multi modal person authentication systems by Bayesian statistics. In *Proc. 1st Int. Conf. on Audio Video-Based Personal Authentication*, pages 327–334, Crans-Montana, Switzerland, March 1997.

[2] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

[3] J. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437 – 1462, September 1997.

[4] R. Chellappa, C. Wilson, and A. Sirohey. Human and machine recognition of faces: A survey. *Proceedings IEEE*, 83(5):705–740, 1995.

[5] U. Dieckmann, P. Plankensteiner, and T. Wagner. Sesam: A biometric person identification system using sensor fusion. *Pattern Recognition Letters*, 18(9):827–833, 1997.

[6] L. Hong and A. Jain. Integrating faces and fingerprints for personal identification. In *Proc. 3rd Asian Conference on Computer Vision*, pages 16–23, Hong Kong, China, 1998.

[7] A. Jain, R. Bolle, and S. Pankanti. *Biometrics: Personal Identification in Networked Society.* Kluwer Academic Publishers, Boston, 1998.

[8] A. Jain, L. Hong, and R. Bolle. On-line fingerprint verification. *IEEE Trans. Pattern Anal. and Machine Intell.*, 19(4):302–314, 1997.

[9] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. PAMI*, 12(1):103–108, 1990.

[10] J. Kittler, Y. Li, J. Matas, and M. U. Sanchez. Combining evidence in multimodal personal identity recognition systems. In *Proc. 1st Int. Conf. on Audio Video-Based Personal Authentication*, pages 327–334, Crans-Montana, Switzerland, March 1997.

[11] S. Maes and H. Beigi. Open sesame! speech, password or key to secure your door? In *Proc. 3rd Asian Conference on Computer Vision*, pages 531–541, Hong Kong, China, 1998.

[12] Y. A. Zuev and S. K. Ivanov. The voting as a way to increase the decision reliability. In *Proc. Foundations of Information/Decision Fusion with Applications to Engineering Problems*, pages 206–210, Washington, D.C., August 1996.