To appear in Proceedings of International Conference on Biometric Authentication, Hong Kong, July 2004

Soft Biometric Traits for Personal Recognition Systems

Anil K. Jain[1], Sarat C. Dass[2], and Karthik Nandakumar[1]

[1] Department of Computer Science and Engineering
{jain,nandakum}@cse.msu.edu
[2] Department of Statistics and Probability
sdass@stt.msu.edu
Michigan State University, MI - 48824, U.S.A.

Abstract. Many existing biometric systems collect ancillary information like gender, age, height, and eye color from the users during enrollment. However, only the primary biometric identifier (fingerprint, face, hand-geometry, etc.) is used for recognition and the ancillary information is rarely utilized. We propose the utilization of "soft" biometric traits like gender, height, weight, age, and ethnicity to complement the identity information provided by the primary biometric identifiers. Although soft biometric characteristics lack the distinctiveness and permanence to identify an individual uniquely and reliably, they provide some evidence about the user identity that could be beneficial. This paper presents a framework for integrating the ancillary information with the output of a primary biometric system. Experiments conducted on a database of 263 users show that the recognition performance of a fingerprint system can be improved significantly ($\approx 5\%$) by using additional user information like gender, ethnicity, and height.

1 Introduction

Biometric systems automatically recognize individuals based on their physiological and/or behavioral characteristics like fingerprint, face, hand-geometry, iris, retina, palmprint, voice, gait, signature, and keystroke dynamics [1]. Biometric systems that use a single trait for recognition, called unimodal biometric systems, are affected by problems like noisy sensor data, non-universality and/or lack of distinctiveness of the chosen biometric trait, unacceptable error rates, and spoof attacks. Some of the problems associated with unimodal biometric systems can be overcome by the use of multimodal biometric systems that combine the evidence obtained from multiple sources [2]. A multimodal biometric system based on different biometric identifiers like fingerprint, iris, face, and hand-geometry can be expected to be more robust to noise, address the problem of non-universality, improve the matching accuracy, and provide reasonable protection against spoof attacks. However, such a system will require a longer verification time thereby causing inconvenience to the users.

A possible solution to the problem of designing a reliable and user-friendly biometric system is to use ancillary information about the user like height, weight, age, gender, ethnicity, and eye color to improve the performance of the primary biometric system. Most practical biometric systems collect such information about the users during enrollment. However, this information is not currently utilized during the automatic

identification/verification phase. Only when a genuine user is falsely rejected by the system, a human operator steps in to verify the soft biometric traits of the user. If these characteristics can be automatically extracted and utilized during the decision making process, the overall performance of the system will improve and the need for manual intervention will be reduced. The ancillary information by itself is not sufficient to establish the identity of a person because these traits are indistinctive, unreliable, and can be easily spoofed. Hence, we define *soft biometric traits* as *characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals*. The soft biometric traits can either be continuous (e.g., height and weight) or discrete (e.g., gender, eye color, ethnicity, etc.). In this paper, we describe a framework for integrating the information provided by the soft biometric indicators with the output of the primary biometric system. We also analyze the performance gains obtained by integrating the ancillary information like gender, ethnicity, and height with the output of a fingerprint biometric system.

2 Related work

The first personal identification system developed by Alphonse Bertillon [3] for identification of criminals was based on three sets of features: (i) anthropometric measurements like height and length of the arm, (ii) morphological description of the appearance and body shape like eye color and anomalies of the fingers, and (iii) peculiar marks observed on the body like moles and scars. Although the Bertillon system was useful in tracking criminals, it had an unacceptably high error rate because the features used are indistinctive (several individuals can have the same set of measurements) and non-permanent (for the same individual, the measurements can change over time). Heckathorn et al. [4] have shown that a combination of personal attributes like gender, race, eye color, height, and other visible identification marks can be used to identify an individual only with a limited accuracy. Hence, a system that is completely based on soft biometric traits cannot meet the accuracy requirements of real-world applications. However, soft biometric traits can be used to improve the performance of traditional biometric systems.

Wayman [5] proposed the use of soft biometric traits like gender and age, for filtering a large biometric database. Filtering refers to limiting the number of entries in a database to be searched, based on characteristics of the interacting user. For example, if the user can somehow be identified as a middle-aged male, the search can be restricted only to the subjects with this profile enrolled in the database. This greatly improves the speed or the search efficiency of the biometric system. In general, filtering reduces the time required for identification but errors in filtering can degrade the recognition performance. Some studies [6],[7] have shown that factors such as age, gender, race, and occupation can affect the performance of a biometric system. For example, a young female Asian mine-worker is seen as the most difficult subject for a fingerprint system [7]. This provides the motivation for tuning the biometric system parameters like threshold on the matching score in a unimodal biometric system, and thresholds and weighting of the different modalities in a multimodal biometric system to obtain the optimum performance for a particular user or a class of users. Filtering and system parameters tuning require an accurate classification of a user into a particular class or bin

(e.g., male or female, blue or brown eyes, Caucasian or Asian or African). This requires a pre-identification module that can accurately perform this classification.

3 Framework for integration of soft biometrics

In our framework, the biometric recognition system is divided into two subsystems. One subsystem is called the primary biometric system and it is based on traditional biometric identifiers like fingerprint, face and hand-geometry. The second subsystem, referred to as the secondary biometric system, is based on soft biometric traits like age, gender, and height. Figure 3 shows the architecture of a personal identification system that makes use of both primary and soft biometric measurements. Let $\omega_1, \omega_2, \cdots, \omega_n$ represent the n users enrolled in the database. Let **x** be the feature vector corresponding to the primary biometric. Without loss of generality, let us assume that the output of the primary biometric system is of the form $P(\omega_i \mid \mathbf{x}),\ i = 1, 2, \cdots, n$, where $P(\omega_i \mid \mathbf{x})$ is the probability that the test user is ω_i given the feature vector **x**. If the output of the primary biometric system is a matching score, it is converted into posteriori probability using an appropriate transformation. For the secondary biometric system, we can consider $P(\omega_i \mid \mathbf{x})$ as the prior probability of the test user being user ω_i.

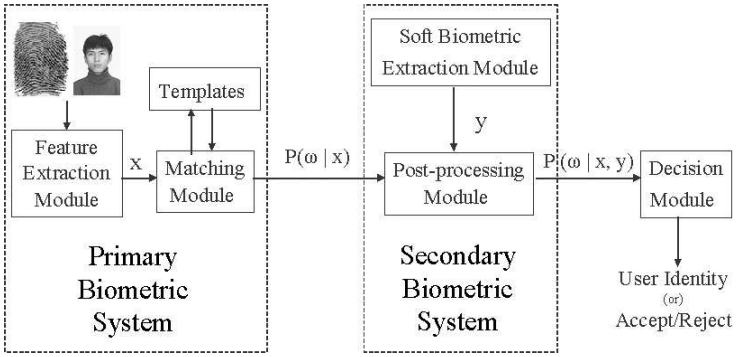


Fig. 1. Integration of Soft Biometric Traits with a Fingerprint Biometric System. (x is the fingerprint feature vector, y is the soft biometric feature vector)

Let $\mathbf{y} = [y_1,\ \cdots,\ y_k,\ y_{k+1},\ \cdots,\ y_m]$ be the soft biometric feature vector, where y_1 through y_k are continuous variables and y_{k+1} through y_m are discrete variables. The updated probability of user ω_i, given the primary biometric feature vector **x** and the soft biometric feature vector **y**, i.e., $P(\omega_i \mid \mathbf{x},\ \mathbf{y})$ can be calculated using the Bayes rule as

$$P(\omega_i|\mathbf{x},\ \mathbf{y}) = \frac{p(\mathbf{y}|\omega_i)\ P(\omega_i|\mathbf{x})}{\sum_{i=1}^{n} p(\mathbf{y}|\omega_i)\ P(\omega_i|\mathbf{x})}\ . \tag{1}$$

If we assume that the soft biometric variables are independent, equation (1) can be rewritten as

$$P(\omega_i|\mathbf{x},\ \mathbf{y}) = \frac{p(y_1|\omega_i)\cdots\ p(y_k|\omega_i)\ P(y_{k+1}|\omega_i)\cdots\ P(y_m|\omega_i)\ P(\omega_i|\mathbf{x})}{\sum_{i=1}^{n} p(y_1|\omega_i)\cdots\ p(y_k|\omega_i)\ P(y_{k+1}|\omega_i)\cdots\ P(y_m|\omega_i)\ P(\omega_i|\mathbf{x})}. \quad (2)$$

In equation (2), $p(y_j|\omega_i),\ j = 1, 2, \cdots, k$ represents the conditional probability of the continuous variable y_j given user ω_i. This can be evaluated from the conditional density of the variable j for user ω_i. On the other hand, discrete probabilities $P(y_j|\omega_i), j = k+1, k+2, \cdots, m$ represents the probability that user ω_i is assigned to the class y_j. This is a measure of the accuracy of the classification module in assigning user ω_i to one of the distinct classes based on biometric indicator y_j. In order to simplify the problem, let us assume that the classification module performs equally well on all the users and therefore the accuracy of the module is independent of the user.
The logarithm of $P(\omega_i|\mathbf{x},\ \mathbf{y})$ in equation (2) can be expressed as

$$\log P(\omega_i|\mathbf{x},\ \mathbf{y}) = \log p(y_1|\omega_i) + \ \cdots\ + \log p(y_k|\omega_i) + \log P(y_{k+1}|\omega_i) + \ \cdots$$
$$+ \log P(y_m|\omega_i) + \log P(\omega_i|\mathbf{x}) - \log p(\mathbf{y}), \quad (3)$$

where $p(\mathbf{y}) = \sum_{i=1}^{n} p(y_1|\omega_i)\ \cdots\ p(y_k|\omega_i)\ P(y_{k+1}|\omega_i)\ \cdots\ (y_m|\omega_i)\ P(\omega_i|\mathbf{x})$.

This formulation has two main drawbacks. The first problem is that all the m soft biometric variables have been weighed equally. In practice, some soft biometric variables may contain more information than the others. For example, the height of a person may give more information about a person than gender. Therefore, we must introduce a weighting scheme for the soft biometric traits based on an index of distinctiveness and permanence, i.e., traits that have smaller variability and larger distinguishing capability will be given more weight in the computation of the final matching probabilities. Another potential pitfall is that any impostor can easily spoof the system because the soft characteristics have an equal say in the decision as the primary biometric trait. It is relatively easy to modify/hide one's soft biometric attributes by applying cosmetics and wearing other accessories (like mask, shoes with high heels, etc.). To avoid this problem, we assign smaller weights to the soft biometric traits compared to those assigned to the primary biometric traits. This differential weighting also has another implicit advantage. Even if a soft biometric trait of a user is measured incorrectly (e.g., a male user is identified as a female), there is only a small reduction in that user's posteriori probability and the user is not immediately rejected. In this case, if the primary biometric produces a good match, the user may still be accepted. Only if several soft biometric traits do not match, there is significant reduction in the posteriori probability and the user could be possibly rejected. If the devices that measure the soft biometric traits are reasonably accurate, such a situation has very low probability of occurrence. The introduction of the weighting scheme results in the following discriminant function for user ω_i:

$$g_i(\mathbf{x},\ \mathbf{y}) = a_0 \log P(\omega_i|\mathbf{x})\ +\ a_1 \log p(y_1|\omega_i) +\ \cdots\ +\ a_k \log p(y_k|\omega_i) +$$
$$a_{k+1} \log P(y_{k+1}|\omega_i)\ +\ \cdots\ +\ a_m \log P(y_m|\omega_i), \tag{4}$$

where $\sum_{i=0}^{m} a_i = 1$ and $a_0 >> a_i,\ i = 1, 2, \cdots, m$. Note that a_i's, $i = 1, 2, \cdots, m$ are the weights assigned to the soft biometric traits and a_0 is the weight assigned to the primary biometric identifier. It must be noted that the weights $a_i,\ i = 1, 2, \cdots, m$ must be made small to prevent the domination of the primary biometric by the soft biometric traits. On the other hand, they must large enough so that the information content of the soft biometric traits is not lost. Hence, an optimum weighting scheme is required to maximize the performance gain.

4 Experimental Results

Our experiments demonstrate the benefits of utilizing the gender, ethnicity, and height information of the user in addition to the fingerprint. Our fingerprint database consisted of impressions of 160 users obtained using a Veridicom sensor. Each user provided four impressions of each of the four fingers, namely, the left index finger, the left middle finger, the right index finger, and the right middle finger. Of these 640 fingers, 263 were selected and assigned uniquely to the users in the face database described in [8]. Gender and ethnicity information of users were automatically extracted from their face images. Fingerprint matching was done using minutia features [9]. Two fingerprint impressions of each user were used as templates and the other two impressions were used for testing. The fingerprint matching score for a particular user was computed as the average of the scores obtained by matching the test impression against the two templates of that user. The separation of the fingerprint database into training and test sets, was repeated 20 times and the results reported are the average for the 20 trials.

The ethnicity classifier proposed in [8] was used in our experiments. This classifier identifies the ethnicity of a test user as either Asian or non-Asian with an accuracy of 96.3%. If a "reject" option is introduced, the probability of making an incorrect classification is reduced to less than 1%, at the expense of rejecting 20% of the test images. A gender classifier was built following the same methodology used in [8] for ethnicity classification. The accuracy of the gender classifier without the "reject" option was 89.6% and the introduction of the "reject" option reduces the probability of making an incorrect classification to less than 2%. In cases where the ethnicity or the gender classifier cannot make a reliable decision, the corresponding information is not utilized for updating the matching score of the primary biometric system.

Since we did not have the height information about the users in the database, we randomly assigned a height 'H_i' to user ω_i, where H_i is drawn from a Gaussian distribution with mean 165 cm and standard deviation 15 cm. The height of a person can be measured during the recognition phase using a sequence of real-time images as described in [10]. However, the measured height will not be equal to the true height of the user stored in the database due to the errors in measurement and the variation in the

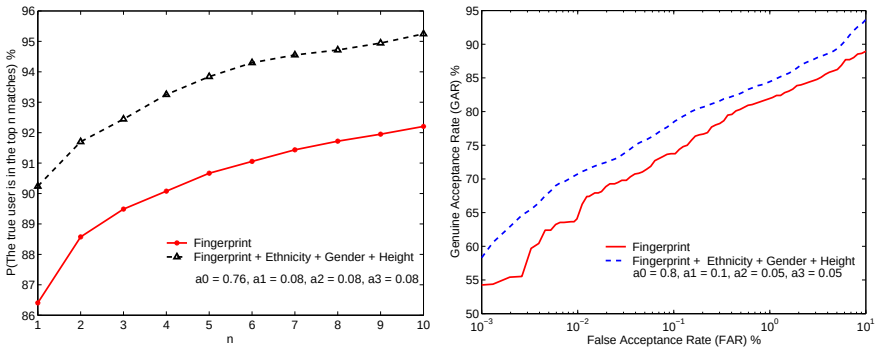
user's height over time. Therefore, it is reasonable to assume that the measured height H_i^* will follow a Gaussian distribution with a mean H_i cm and a standard deviation of 5 cm.

Let $P(\omega_i|s)$ be the posterior probability that the test user is user ω_i given the fingerprint matching score 's' of the test user. Let $y_i = (G_i,\ E_i,\ H_i)$ be the soft biometric feature vector corresponding to the user ω_i, where G_i, E_i, and H_i are the true values of gender, ethnicity, and height of ω_i. Let $y^* = (G^*,\ E^*,\ H^*)$ be the observed soft biometric feature vector of the test user, where G^* is the observed gender, E^* is the observed ethnicity, and H^* is the observed height. Now the final score after considering the observed soft biometric characteristics is computed as

$$g_i(s,\ y^*) = a_0 \log P(\omega_i|s) + a_1 \log p(H^*|H_i) + a_2 \log P(G^*|G_i) + a_3 \log P(E^*|E_i)\ ,$$

where $a_2 = 0$ if G^* ="reject", and $a_3 = 0$ if E^* ="reject".

Figure 2(a) shows the Cumulative Match Characteristic (CMC) of the fingerprint biometric system operating in the identification mode, and the improvement in performance achieved after the utilization of soft biometric information. The weights assigned to the primary and soft biometric traits were selected intuitively such that the performance gain is maximized. However, no formal procedure was used and an exhaustive search of all possible sets of weights was not attempted. The use of ethnicity, gender, and height information along with the fingerprint leads to an improvement of approximately 5% over the primary biometric system. Figure 2(b) shows the Receiver Operating Characteristic (ROC) of a biometric system operating in the verification mode, using fingerprint as the primary biometric identifier and ethnicity, gender, and height as the soft biometric traits. An improvement of about 4% in the Genuine Acceptance Rate (GAR) can be observed over a wide range of values of False Acceptance Rate (FAR).



(a) (b)

Fig. 2. Improvement in recognition performance of a fingerprint system after utilization of soft biometric traits (a) Identification mode (b) Verification mode.

5 Conclusions

We have formulated a mathematical framework based on the Bayesian decision theory for integrating the soft biometric information with the output of the primary biometric system. We have demonstrated that the utilization of ancillary user information like gender, height, and ethnicity can improve the performance of the traditional biometric systems like fingerprint. Although these soft biometric characteristics are not as permanent and reliable as the traditional biometric identifiers like fingerprint, they provide some information about the identity of the user that leads to higher accuracy in establishing the user identity. However, an optimum weighting scheme based the discriminative abilities of the primary and the soft biometric traits is needed to achieve an improvement in recognition performance.

Our future research work will involve establishing a more formal procedure to determine the optimal set of weights for the soft characteristics based on their distinctiveness and permanence. Methods to incorporate time-varying soft biometric information such as age and weight into the soft biometric framework will be studied. The effectiveness of utilizing the soft biometric information for "indexing" and "filtering" of large biometric databases must be studied. Finally, more accurate mechanisms must be developed for automatic extraction of soft biometric traits.

References

1. Jain, A.K., Bolle, R., Pankanti, S., eds.: Biometrics: Personal Identification in Networked Security. Kluwer Academic Publishers (1999)
2. Hong, L., Jain, A.K., Pankanti, S.: Can Multibiometrics Improve Performance? In: Proceedings of IEEE Workshop on Automatic Identification Advanced Technologies, New Jersey, U.S.A. (1999) 59–64
3. Bertillon, A.: Signaletic Instructions including the theory and practice of Anthropometrical Identification, R.W. McClaughry Translation. The Werner Company (1896)
4. Heckathorn, D.D., Broadhead, R.S., Sergeyev, B.: A Methodology for Reducing Respondent Duplication and Impersonation in Samples of Hidden Populations. In: Annual Meeting of the American Sociological Association, Toronto, Canada (1997)
5. Wayman, J.L.: Large-scale Civilian Biometric Systems - Issues and Feasibility. In: Proceedings of Card Tech / Secur Tech ID. (1997)
6. Givens, G., Beveridge, J.R., Draper, B.A., Bolme, D.: A Statistical Assessment of Subject Factors in the PCA Recognition of Human Subjects. In: Proceedings of CVPR Workshop: Statistical Analysis in Computer Vision. (2003)
7. Newham, E.: The Biometrics Report. SJB Services (1995)
8. Jain, A.K., Lu, X.: Ethnicity Identification from Face Images. In: Proceedings of SPIE International Symposium on Defense and Security : Biometric Technology for Human Identification (To appear). (2004)
9. Jain, A.K., Hong, L., Pankanti, S., Bolle, R.: An identity authentication system using fingerprints. Proceedings of the IEEE **85** (1997) 1365–1388
10. Kim, J.S., et al.: Object Extraction for Superimposition and Height Measurement. In: Proceedings of Eighth Korea-Japan Joint Workshop on Frontiers of Computer Vision. (2002)