Chapter 4

SCORE LEVEL FUSION

4.1 Introduction

The match score is a measure of similarity between the input and template biometric feature vectors. When match scores output by different biometric matchers are consolidated in order to arrive at a final recognition decision, fusion is said to be done at the match score level. This is also known as fusion at the measurement level or confidence level. Apart from the raw data and feature vectors, the match scores contain the richest information about the input pattern. Also, it is relatively easy to access and combine the scores generated by different biometric matchers. Consequently, information fusion at the match score level is the most commonly used approach in multibiometric systems.

The general flow of information in a match score level fusion scheme is shown in Figure 4.1. It must be noted that the match scores generated by the individual matchers may not be homogeneous. For example, one matcher may output a distance or dissimilarity measure (a smaller distance indicates a better match) while another may output a similarity measure (a larger similarity value indicates a better match). Furthermore, the outputs of the individual matchers need not be on the same numerical scale (range). Finally, the match scores may follow different probability distributions. These three factors make match score level fusion a challenging problem. In this chapter, we will analyze some of the techniques to perform match score level fusion. We first present a mathematical framework that describes classifier combination from a pattern recognition perspective.

Face Matcher

Fingerprint Matcher

User Identity	Match Score
Alice	0.2
Bob	**0.9**
Charlie	-0.4
Dick	0.6

User Identity	Match Score
Alice	**87**
Bob	85
Charlie	62
Dick	20

Score Fusion Module

User Identity	Fused Score
Alice	1.47
Bob	**1.80**
Charlie	0.92
Dick	1.00

Figure 4.1. Flow of information in a match score level fusion scheme. In this example, the match scores have been combined using the sum of scores fusion rule after min-max normalization of each matcher's output. Note that the match scores generated by the face and fingerprint matchers are similarity measures. The range of match scores is assumed to be $[-1, +1]$ and $[0, 100]$ for the face and fingerprint matchers, respectively.

4.2 Classifier combination rules

In the context of statistical pattern recognition, Kittler et al., 1998 developed a theoretical framework for consolidating the evidence obtained from multiple classifiers, where each classifier makes use of a different representation derived from the same input pattern. Consider the problem of classifying an input pattern X into one of M possible classes $\{\omega_1, \omega_2, \ldots, \omega_M\}$ based on the evidence provided by R different classifiers. Let $\boldsymbol{x}_j$ be the feature vector (derived from the input pattern X) presented to the j^{th} classifier. In general, each of the R classifiers can have its own multidimensional feature vector derived from the input pattern X. In the chosen feature space, each class ω_k can be modeled by a probability density function $p(\boldsymbol{x}_j|\omega_k)$ and its prior probability of occurrence is denoted by $P(\omega_k)$.

According to the Bayesian decision theory (Duda et al., 2001), given the feature vectors $\boldsymbol{x}_j, j = 1, \ldots, R$, the input pattern X should be assigned to the class ω_r that maximizes the posteriori probability, i.e.,

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$P(\omega_r|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R) \geq P(\omega_k|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R), \tag{4.1}$$

where $k = 1, \ldots, M$. The Bayesian decision rule stated in Equation 4.1 is known as the minimum error-rate classification rule in the pattern recognition literature. This rule assumes a *zero-one* loss function which assigns no loss to a correct decision and assigns a unit loss to any misclassification error. The posterior probabilities in Equation 4.1 can be expressed in terms of the conditional joint probability densities of the feature vectors, $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R|\omega_k)$, by using the Bayes rule as follows:

$$P(\omega_k|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R) = \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R|\omega_k)P(\omega_k)}{\sum_{l=1}^{M} p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R|\omega_l)P(\omega_l)}. \tag{4.2}$$

Kittler et al., 1998 suggest many approximations to simplify the computation of the posterior probability in Equation 4.2 which lead to five classifier combination strategies. All the five strategies are based on the assumption of statistical independence of the R feature representations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R$. Under this assumption, the conditional joint probability density $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R|\omega_k)$ can be expressed as the product of the marginal conditional densities, i.e.,

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R|\omega_k) = \prod_{j=1}^{R} p(\boldsymbol{x}_j|\omega_k), \tag{4.3}$$

where $k = 1, \ldots, M$. In a multimodal biometric system, each one of the R classifiers uses features from a different biometric trait. Different biometric

traits of an individual (e.g., face, fingerprint and hand geometry) generally tend to be mutually independent. Hence, the underlying assumption in Equation 4.3 is reasonable in most multimodal biometric systems. On the other hand, the independence assumption may not be true for a multi-sample biometric system (e.g., two impressions of the same finger) that uses the same representation scheme (e.g., minutiae) for each sample. This is because different samples of the same biometric trait usually tend to be correlated.

Product Rule: This rule is a direct implication of the assumption of statistical independence between the R feature representations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R$. The product decision rule can be stated as

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$P(\omega_r) \prod_{j=1}^{R} p(\boldsymbol{x}_j|\omega_r) \geq P(\omega_k) \prod_{j=1}^{R} p(\boldsymbol{x}_j|\omega_k), \tag{4.4}$$

where $k = 1, \ldots, M$. The product rule can also be expressed in terms of the product of the posteriori probabilities of the individual classifiers as follows.

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\frac{\prod_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j)}{(P(\omega_r))^{(R-1)}} \geq \frac{\prod_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j)}{(P(\omega_k))^{(R-1)}}, \tag{4.5}$$

where $k = 1, \ldots, M$. Further, in most practical biometric systems all classes (M users in the identification mode and "genuine" and "impostor" classes in the verification mode) are assigned equal prior probabilities. Under this assumption, the product rule can be simplified as

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\prod_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j) \geq \prod_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j). \tag{4.6}$$

One of the main limitations of the product rule is its sensitivity to errors in the estimation of the posteriori probabilities. Even if one of the classifiers outputs a probability close to zero, the product of the R posteriori probabilities is rather small and this often leads to an incorrect classification decision.

Sum Rule: The sum rule is more effective than the product rule when the input X tends to be noisy, leading to errors in the estimation of the posteriori

probabilities. In such a scenario, we can assume that the posteriori probabilities do not deviate dramatically from the prior probabilities for each class, i.e.,

$$P(\omega_k|\boldsymbol{x}_j) = P(\omega_k)(1 + \delta_{j,k}), \tag{4.7}$$

where $\delta_{j,k}$ is a constant, $0 < \delta_{j,k} << 1$; $j = 1, \ldots, R$; $k = 1, \ldots, M$. Substituting Equation 4.7 for the posteriori probabilities in Equation 4.5, we get

$$\frac{\prod_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j)}{(P(\omega_k))^{(R-1)}} = P(\omega_k) \prod_{j=1}^{R} (1 + \delta_{j,k}). \tag{4.8}$$

Expanding the product on the right hand side in Equation 4.8 and neglecting the higher order terms, we can approximate the product in terms of a summation as follows.

$$\prod_{j=1}^{R} (1 + \delta_{j,k}) \approx 1 + \sum_{j=1}^{R} \delta_{j,k}. \tag{4.9}$$

Further, by substituting for $\delta_{j,k}$ from Equation 4.7, we get

$$\sum_{j=1}^{R} \delta_{j,k} = \frac{\sum_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j)}{P(\omega_k)} - R. \tag{4.10}$$

Finally, substituting Equations 4.9 and 4.10 into Equation 4.5 for the product rule, we obtain the sum decision rule which can be stated as follows:

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\left\{ (1 - R)P(\omega_r) + \sum_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j) \right\} \geq \left\{ (1 - R)P(\omega_k) + \sum_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j) \right\}, \tag{4.11}$$

where $k = 1, \ldots, M$. When the prior probabilities are equal, the sum rule can be expressed as follows.

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\sum_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j) \geq \sum_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j), \tag{4.12}$$

where $k = 1, \ldots, M$. The decision rule in Equation 4.12 is also known as the *mean* or *average* decision rule because it is equivalent to assigning the input

pattern to the class that has the maximum average posteriori probability over all the R classifiers.

As mentioned earlier, the sum rule is primarily based on the assumption that the posteriori probabilities $P(\omega_k|\boldsymbol{x}_j)$ do not deviate much from the prior probabilities $P(\omega_k)$. In general, this assumption is unrealistic because the feature vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R$ contain significant discriminatory information about the pattern class. However, Kittler et al., 1998 showed that the sum rule is robust to errors in the estimation of the posteriori probabilities. Therefore, the sum decision rule usually works quite well in practice and is commonly used in multibiometric systems.

Max Rule: The max rule approximates the mean of the posteriori probabilities by their maximum value, i.e.,

$$\frac{1}{R}\sum_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j) \approx \max_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j). \tag{4.13}$$

Hence, the max rule can be stated as follows:

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\left\{(1-R)P(\omega_r) + R\max_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j)\right\} \geq \left\{(1-R)P(\omega_k) + R\max_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j)\right\}, \tag{4.14}$$

where $k = 1, \ldots, M$. Under the assumption of equal priors, the max rule can be simplified as

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\max_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j) \geq \max_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j), \; k = 1, \ldots, M. \tag{4.15}$$

Min Rule: It is well known that the product of probabilities is always less than or equal to the minimum value of probability in the product. Hence,

$$\prod_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j) \leq \min_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j). \tag{4.16}$$

By substituting this upper bound in place of the product term in Equation 4.5, we obtain the min rule which can be stated as

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\frac{\min_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j)}{(P(\omega_r))^{(R-1)}} \geq \frac{\min_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j)}{(P(\omega_k))^{(R-1)}}, \; k = 1, \ldots, M. \tag{4.17}$$

If the prior probabilities of all the classes are equal, the min rule reduces to

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$\min_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j) \geq \min_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j), \; k = 1, \ldots, M. \tag{4.18}$$

Median Rule: If we assume equal priors for all the classes, the sum rule in Equation 4.12 can be viewed as the mean rule. The mean rule assigns a pattern to the class that has the maximum average posteriori probability over all the classifiers. Since the average posteriori probability is sensitive to outliers, it is often replaced by the median value. The median decision rule can be stated as

$$\text{Assign} \quad X \rightarrow \omega_r \quad \text{if}$$

$$median_{j=1}^{R} P(\omega_r|\boldsymbol{x}_j) \geq median_{j=1}^{R} P(\omega_k|\boldsymbol{x}_j), \; k = 1, \ldots, M. \tag{4.19}$$

The classifier combination rules developed by Kittler et al., 1998 can be used in a multibiometric system only if the output of each biometric matcher is of the form $P(\omega_k|\boldsymbol{x}_j)$ i.e., the posteriori probability of class ω_k given the features extracted by the j^{th} modality from the input biometric sample X. In practice, most biometric matchers output only a match score $s_{j,k}$. Verlinde et al., 1999 proposed that the match score $s_{j,k}$ is related to $P(\omega_k|\boldsymbol{x}_j)$ as follows:

$$s_{j,k} = g\left(P(\omega_k|\boldsymbol{x}_j)\right) + \beta(\boldsymbol{x}_j), \tag{4.20}$$

where g is a monotonic function and β is the error made by the biometric matcher that depends on the input features. This error could be due to the noise introduced by the sensor during the acquisition of the biometric signal, and the errors made by the feature extraction and matching processes. If we assume that β is zero, it is reasonable to approximate $P(\omega_k|\boldsymbol{x}_j)$ by $P(\omega_k|s_{j,k})$. In this scenario, the classifier combination rules can be applied for fusion of match scores from different biometric matchers. On the other hand, if we assume that the value of β is non-zero, $P(\omega_k|s_{j,k})$ may not be a good estimate of $P(\omega_k|\boldsymbol{x}_j)$. Hence, it is not possible to directly apply the classifier combination rules in such a scenario.

4.3 Score fusion techniques

Let us consider a multibiometric system operating in the verification mode where the output of each biometric matcher is a match score (the formulation presented below can be trivially extended to the identification scenario also). Since the goal of the multibiometric system is to determine whether the input biometric sample X belongs to a "genuine" user or an "impostor", the number of classes (M) is now reduced to two. The minimum error-rate decision rule in Equation 4.1 is based on the assumption that all types of errors (misclassifying a sample from class ω_k as $\omega_{k'}, \forall\, k,\ k' = 1, \ldots, M,\ k \neq k'$) are equally costly. Most practical verification systems assign different costs to the false accept and false reject errors. Let λ_1 and λ_2 be the cost (or loss) associated with the false accept and false reject errors, respectively, and let $\eta = \lambda_1/\lambda_2$ be the ratio of the two cost values. Therefore, when a biometric system operating in the verification mode has different costs for the false accept and false reject errors, the modified Bayesian decision rule is

$$\text{Assign} \quad X \rightarrow genuine \quad \text{if}$$

$$\frac{P(genuine|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)}{P(impostor|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)} \geq \eta. \tag{4.21}$$

In score level fusion, it is assumed that the feature representations of the R biometric matchers $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R$ are not available. Hence, the posteriori probabilities $P(genuine \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)$ and $P(impostor|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)$ must be estimated from the vector of match scores $\boldsymbol{s} = [s_1, s_2, \ldots, s_R]$, where s_j is the match score provided by the j^{th} matcher, $j = 1, \ldots, R$ (note that since the class is fixed as either "genuine" or "impostor", we drop the subscript k that represents the class information). Techniques that have been proposed for estimating these posteriori probabilities can be divided into three broad categories listed below.

1 The first approach assumes that the posteriori probabilities $P(genuine|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)$ and $P(impostor|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)$ can be approximated by $P(genuine|\boldsymbol{s} = [s_1, s_2, \ldots, s_R])$ and $P(impostor|\boldsymbol{s} = [s_1, s_2, \ldots, s_R])$, respectively. Conversion of the vector of scores, $\boldsymbol{s}$, into the probabilities $P(genuine|\boldsymbol{s})$ and $P(impostor|\boldsymbol{s})$ requires explicit estimation of the underlying conditional densities $p(\boldsymbol{s}|genuine)$ and $p(\boldsymbol{s}|impostor)$. Hence, this approach is referred to as *density-based score fusion*. After estimating the densities, the probabilities $P(genuine|\boldsymbol{s})$ and $P(impostor|\boldsymbol{s})$ are computed, and the Bayesian decision rule in Equation 4.21 can be used to make a decision.

2 Accurate estimation of the class conditional densities $p(\boldsymbol{s}|genuine)$ and $p(\boldsymbol{s}|impostor)$ is possible only when the number of match scores available

for training the fusion module is large. Also, the assumption that the posteriori probabilities $P(genuine|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)$ and $P(impostor|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)$ can be approximated by $P(genuine|\boldsymbol{s})$ and $P(impostor|\boldsymbol{s})$ is valid only when the value of β in Equation 4.20 is zero. Hence, in cases where the number of training match scores is limited and/or β's are non-zero, an alternative approach is to transform the match scores obtained from the different matchers into a common domain in order to make them compatible. This transformation is known as score normalization and the resulting fusion approach is known as *transformation based score fusion*. In the transformed domain, the sum, max and min classifier combination rules can be directly applied. In general, the normalized scores do not have any probabilistic interpretation. Therefore, the product rule given by Equation 4.6 cannot be applied.

3 The third approach is *classifier based score fusion* where the relationship between the vector of match scores $[s_1, s_2, \ldots, s_R]$ and the posteriori probabilities, $P(genuine|s_1, s_2, \ldots, s_R)$ and $P(impostor|s_1, s_2, \ldots, s_R)$, is indirectly learned using a pattern classifier.

It must be emphasized that these three methodologies are essentially different approaches to solving the same problem, namely, deciding whether the input pattern X belongs to the "genuine" or the "impostor" class based on the match score vector $[s_1, s_2, \ldots, s_R]$ generated by the R different biometric matchers. Each method has its own advantages and limitations. Further, each method requires estimation of some parameters from the training data and exhibits different levels of sensitivity to problems like lack of sufficient training data and noisy training samples. Finally, none of these three methods is guaranteed to provide optimum performance under all scenarios. In the following sections, we will describe these three approaches in detail.

4.4 Density-based score fusion

Let S_{gen} and S_{imp} be the random variables denoting the genuine and impostor match scores, respectively. Let $F_{gen}(s)$ be the distribution function of S_{gen} and $f_{gen}(s)$ be the corresponding density, i.e.,

$$P(S_{gen} \leq s) = F_{gen}(s) = \int_{-\infty}^{s} f_{gen}(v)dv. \tag{4.22}$$

Similarly, let $F_{imp}(s)$ be the distribution function of S_{imp} and $f_{imp}(s)$ be the corresponding density, i.e.,

$$P(S_{imp} \leq s) = F_{imp}(s) = \int_{-\infty}^{s} f_{imp}(v)dv. \tag{4.23}$$

The densities $f_{gen}(s)$ and $f_{imp}(s)$ are known as the class conditional densities because they represent the probability density functions of the match score given that the score comes from the genuine or impostor class ($p(s|genuine)$ and $p(s|impostor)$), respectively. The densities $f_{gen}(s)$ and $f_{imp}(s)$ are usually not known and have to be estimated from a set of training scores from the genuine and impostor classes.

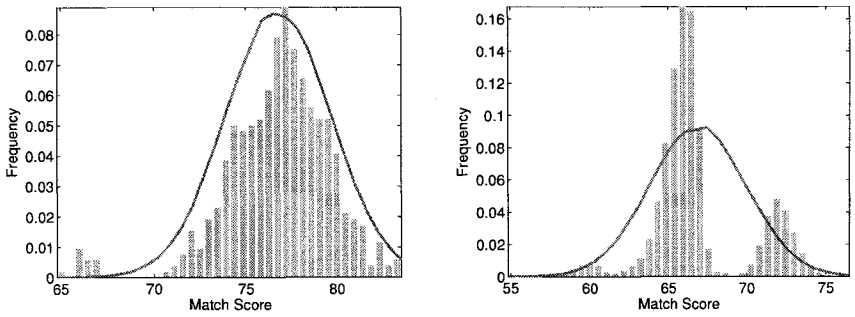
Density estimation can be done either by parametric or non-parametric methods (Duda et al., 2001). In parametric density estimation techniques, the form of the density function is assumed to be known and only the parameters of this density function are estimated from the training data. For example, if we assume a Gaussian (normal) density function, only the mean and the standard deviation parameters that characterize this density are estimated during training. On the other hand, non-parametric techniques do not assume any standard form for the density function and are essentially data-driven. The Parzen window and K-NN density estimation schemes fall in this category. In the context of multibiometric systems, it is very difficult to choose a specific parametric form for the density of genuine and impostor scores. It is well known that the commonly assumed Gaussian density approximation is usually not appropriate for genuine and impostor scores of a biometric matcher. The match score distributions generally have a large tail and may have more than one mode (see Figure 4.2). However, the Gaussian distribution is unimodal and does not capture the information contained in the tails of the distribution very well, making it inappropriate for modeling genuine and impostor score distributions. Another major problem that biometric researchers are facing is that they do not have access to large amounts of training data (especially genuine match scores) to reliably estimate the genuine and impostor densities. For example, if a multibiometric database has n users and if each user provides m biometric samples, then the maximum number of genuine scores, N_{gen}, that can be obtained from this database is $nm(m-1)/2$. On the other hand, $n(n-1)m^2$ impostor matches, N_{imp}, can be performed using the same database. Suppose that $n = 100$ and $m = 4$, the number of genuine scores available is only 600 while the number of impostor scores is $158,400$. Due to the limited availability of training data, especially genuine scores, the density estimation method must be chosen carefully.

Snelick et al., 2003 adopt a parametric approach to estimate the conditional densities of the match scores. They assume a normal distribution for the conditional densities of the match scores, i.e.,

$$p(s_j|genuine) \sim \mathcal{N}(\mu_{j,gen}, \sigma_{j,gen}) \tag{4.24}$$

and

$$p(s_j|impostor) \sim \mathcal{N}(\mu_{j,imp}, \sigma_{j,imp}), \tag{4.25}$$



(a) (b)

Figure 4.2. Histograms of match scores and the corresponding Gaussian density estimates for the `Face-G` matcher in the NIST BSSR1 database. (a) Genuine and (b) Impostor. Note that the Gaussian density does not account well for the tail in the genuine score distribution and the multiple modes in the impostor score distribution.

where $\mu_{j,gen}$ ($\mu_{j,imp}$) and $\sigma_{j,gen}$ ($\sigma_{j,imp}$) are the mean and standard deviation of the genuine (impostor) match scores of the j^{th} matcher, respectively. Based on the training data containing N_{gen} genuine scores and N_{imp} impostor scores for each matcher, the maximum likelihood estimates of the parameters $\mu_{j,gen}$, $\sigma_{j,gen}$, $\mu_{j,imp}$, and $\sigma_{j,imp}$ are obtained as follows.

$$\hat{\mu}_{j,gen} = \frac{1}{N_{gen}} \sum_{i=1}^{N_{gen}} s^i_{j,gen}, \tag{4.26}$$

$$\hat{\mu}_{j,imp} = \frac{1}{N_{imp}} \sum_{i=1}^{N_{imp}} s^i_{j,imp}, \tag{4.27}$$

$$\hat{\sigma}_{j,gen} = \frac{1}{N_{gen}} \sum_{i=1}^{N_{gen}} (s^i_{j,gen} - \hat{\mu}_{j,gen})^2, \tag{4.28}$$

$$\hat{\sigma}_{j,imp} = \frac{1}{N_{imp}} \sum_{i=1}^{N_{imp}} (s^i_{j,imp} - \hat{\mu}_{j,imp})^2, \tag{4.29}$$

where $s^i_{j,gen}$ ($s^i_{j,imp}$) represents the i^{th} genuine (impostor) training score of the j^{th} matcher, $i = 1, \ldots, N_{gen}$ for the genuine class and $i = 1, \ldots, N_{imp}$ for the impostor class.

Given a test match score s_j^t for the j^{th} matcher, the posteriori probabilities of the score belonging to a genuine user and an impostor are computed as follows.

$$P(genuine|s_j^t) = \frac{p(s_j^t|genuine)P(genuine)}{p(s_j^t)} \tag{4.30}$$

and

$$P(impostor|s_j^t) = \frac{p(s_j^t|impostor)P(impostor)}{p(s_j^t)}, \tag{4.31}$$

where $p(s_j^t) = (p(s_j^t|genuine)P(genuine) + p(s_j^t|impostor)\,P(impostor))$ and $P(genuine)$ and $P(impostor)$ are the prior probabilities of a genuine user and an impostor, respectively. Snelick et al., 2003 assumed that prior probabilities of the genuine and impostor classes are equal and the matchers are conditional independent. Hence, the posteriori probabilities based on the scores from the different matchers can be computed as follows.

$$P(genuine|s_1^t, \ldots, s_R^t) = \prod_{j=1}^{R} P(genuine|s_j^t) \tag{4.32}$$

and

$$P(impostor|s_1^t, \ldots, s_R^t) = \prod_{j=1}^{R} P(impostor|s_j^t). \tag{4.33}$$

The final accept/reject decision is based on the Bayesian decision rule in Equation 4.21 which can be stated as follows.

$$\text{Assign} \quad X^t \rightarrow genuine \quad \text{if}$$

$$\frac{P(genuine|s_1^t, \ldots, s_R^t)}{P(impostor|s_1^t, \ldots, s_R^t)} \geq \eta, \tag{4.34}$$

where X^t is the given test sample and η is the decision threshold which is a tradeoff between the false accept and false reject error rates. When the goal is to minimize the total error rate (sum of the false accept and the false reject rates), the value of η should be set to 1. As pointed out earlier, the assumption of a normal distribution for the scores is generally not true for biometric match scores.

Jain et al., 2005 propose the use of the Parzen window based non-parametric density estimation method (Duda et al., 2001) to estimate the conditional density of the genuine and impostor scores. After estimating the conditional densities,

equations 4.30 through 4.34 can be applied to make a decision. Although the Parzen window density estimation technique is appropriate for estimating the conditional densities $p(s_j|genuine)$ and $p(s_j|impostor)$, especially when the densities are non-Gaussian, the resulting density estimates may still have inaccuracies due to the finite training set and the problems in choosing the optimum window width during the density estimation process.

Both Snelick et al., 2003 and Jain et al., 2005 estimate only the marginal densities of the individual matchers in a multibiometric system. The combination of these marginal densities is achieved using the framework developed by Kittler et al., 1998 based on the assumption of statistical independence of the feature vectors (or the biometric matchers). Prabhakar and Jain, 2002 argue that the assumption of statistical independence of the matchers may not be true in a multi-algorithm biometric system that uses different feature representations and different matching algorithms on the same biometric trait. Hence, they propose a scheme based on non-parametric estimation of the joint multivariate density. Using the genuine and impostor match scores from the R matchers that were available for training, they directly estimate the R-variate densities $p(s_1, \ldots, s_R|genuine)$ and $p(s_1, \ldots, s_R|impostor)$. But estimating the joint multivariate densities requires a larger number of training samples than estimating the univariate (marginal) densities. Hence, this approach is applicable only when a very large amount of training data is available to estimate the joint densities. Based on the joint densities, the posteriori probabilities can be computed using the Bayes rule as follows.

$$P(genuine|s_1, \ldots, s_R) = \frac{p(s_1, \ldots, s_R|genuine)P(genuine)}{p(s_1, \ldots, s_R)} \tag{4.35}$$

and

$$P(impostor|s_1, \ldots, s_R) = \frac{p(s_1, \ldots, s_R|impostor)P(impostor)}{p(s_1, \ldots, s_R)}, \tag{4.36}$$

where

$$\begin{aligned} p(s_1, \ldots, s_R) &= p(s_1, \ldots, s_R|genuine)P(genuine) + \\ &\quad p(s_1, \ldots, s_R|impostor)P(impostor). \end{aligned}$$

Hence, the ratio of the posteriori probabilities is given by

$$\frac{P(genuine|s_1, \ldots, s_R)}{P(impostor|s_1, \ldots, s_R)} = \frac{p(s_1, \ldots, s_R|genuine)P(genuine)}{p(s_1, \ldots, s_R|impostor)P(impostor)}. \tag{4.37}$$

When the prior probabilities of the genuine and impostor classes are equal, the ratio of the posteriori probabilities is

$$\frac{P(genuine|s_1,\ldots,s_R)}{P(impostor|s_1,\ldots,s_R)} = \frac{p(s_1,\ldots,s_R|genuine)}{p(s_1,\ldots,s_R|impostor)}. \tag{4.38}$$

The terms $p(s_1,\ldots,s_R|genuine)$ and $p(s_1,\ldots,s_R|impostor)$ are also referred to as the likelihood of the genuine and impostor class with respect to $[s_1,\ldots,s_R]$. Hence, the ratio on the right hand side in Equation 4.38 is known as the likelihood ratio. The Neyman-Pearson theorem (Lehmann and Romano, 2005) states that when the prior probabilities of the classes are equal (or not known), the *optimal* test for deciding whether a match score vector $\boldsymbol{s} = [s_1,\ldots,s_R]$ corresponds to a genuine or impostor match is the likelihood ratio test. The Neyman-Pearson decision rule is optimal in the sense that if we assume that the false accept rate (FAR) is given, the likelihood ratio test will minimize the false reject rate (FRR) for the fixed FAR and no other decision rule will give a lower FRR. The decision rule based on the likelihood ratio test can be stated as follows.

$$\text{Assign} \quad X \rightarrow genuine \quad \text{if}$$

$$\frac{p(s_1,\ldots,s_R|genuine)}{p(s_1,\ldots,s_R|impostor)} \geq \eta, \tag{4.39}$$

where η is the threshold value that achieves the specified value of FAR. The likelihood ratio test is optimal only when the underlying densities are either known or can be estimated very accurately. Hence, given a set of genuine and impostor match scores, it is important to be able to estimate the conditional densities $f_{gen}(s)$ and $f_{imp}(s)$ without incurring large errors in the estimation process.

Another important consideration is that the distribution of genuine and impostor scores of some biometric matchers may exhibit discrete components. This happens because most biometric matching algorithms apply certain thresholds at various stages in the matching process. When the required threshold conditions are not met, pre-determined match scores are output by the matcher (e.g., some fingerprint matchers produce a match score of zero if the number of extracted minutiae is less than a threshold, irrespective of how many minutiae actually match between the query and the template). This leads to discrete components in the match score distribution that cannot be modeled accurately using a continuous density function. Thus, discrete components need to be detected and the discrete and continuous portions of the density must be modeled separately to avoid large errors in estimating $f_{gen}(s)$ and $f_{imp}(s)$. To address this problem, Dass et al., 2005 propose a framework for combining the

match scores from multiple matchers based on generalized densities estimated from the genuine and impostor match scores. The generalized densities are a mixture of discrete and continuous components and a brief description of the methodology used for computing the generalized densities is presented below.

4.4.1 Generalized densities

The following methodology models a distribution based on a generic set of observed scores (the same formulation can be used for both genuine and impostor scores from any biometric matcher). Let S denote a generic match score with distribution function F and density $f(s)$, i.e.,

$$P(S \leq s) = F(s) = \int_{-\infty}^{s} f(v)dv. \tag{4.40}$$

For a fixed threshold T, the discrete values are identified as those values s_0 with $P(S = s_0) > T$, where T is a threshold, $0 \leq T \leq 1$. Since the underlying match score distribution is unknown, the probability $P(S = s_0)$ can be estimated by $N(s_0)/N$, where $N(s_0)$ is the number of observations in the data set that equals s_0 and N is the total number of observations. Let the subset of all discrete components for a match score distribution be denoted by

$$\mathcal{D} \equiv \left\{ s_0 : \frac{N(s_0)}{N} > T \right\}. \tag{4.41}$$

The discrete components constitute a proportion $p_D \equiv \sum_{s_0 \in \mathcal{D}} \frac{N(s_0)}{N}$ out of the total of N available observations. The subset $\mathcal{C}$ of observations can be obtained by removing all discrete components from the available data set. The scores in $\mathcal{C}$ constitute a proportion $p_C \equiv 1 - p_D$ of the entire data set, and they are used to estimate the continuous component of the distribution ($F_C(s)$) and the corresponding density ($f_c(s)$). A non-parametric kernel density estimate of $f_c(s)$ is obtained from $\mathcal{C}$ as follows. The empirical distribution function for the observations in $\mathcal{C}$ is computed as

$$\hat{F}_C(s) = \frac{1}{N_C} \sum_{v \in \mathcal{C}} I\{ v \leq s \}, \tag{4.42}$$

where N_C is the number of observations in $\mathcal{C}$ and

$$I\{v \leq s\} = \begin{cases} 1, & \text{if } v \leq s, \\ 0, & \text{otherwise;} \end{cases} \tag{4.43}$$

also, $N_C \equiv N\, p_C$. Note that $\hat{F}_C(s) = 0 \; \forall \; s < s_{min}$ and $\hat{F}_C(s) = 1 \; \forall \; s \geq s_{max}$, where s_{min} and s_{max}, respectively, are the minimum and maximum values of the observations in $\mathcal{C}$. For values of $s, s_{min} < s < s_{max}$, not contained in $\mathcal{C}$, $\hat{F}_C(s)$, is obtained by linear interpolation. Next, B samples are

simulated from $\hat{F}_C(s)$, and the density estimate of $f_C(s)$, $\hat{f}_C(s)$, is obtained from the simulated samples using a Gaussian kernel density estimator. The optimal bandwidth, h, of the kernel is obtained using the "solve-the-equation" bandwidth estimator (Wand and Jones, 1995), which is an automatic bandwidth selector that prevents oversmoothing and preserves important properties of the distribution of match scores. The generalized density is defined as

$$f(s) = p_C \hat{f}_C(s) + \sum_{s_0 \in \mathcal{D}} \frac{N(s_0)}{N} \cdot I\{s = s_0\}, \tag{4.44}$$

where

$$I\{s = s_0\} = \begin{cases} 1, & \text{if } s = s_0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.45}$$

The distribution function corresponding to the generalized density is defined as

$$F(s) = p_C \int_{-\infty}^{s} \hat{f}_C(v)\, dv + \sum_{s_0 \in \mathcal{D},\, s_0 \leq s} \frac{N(s_0)}{N}. \tag{4.46}$$

For a multibiometric system with R matchers, the generalized densities and distributions estimated for the genuine (impostor) scores for the j^{th} matcher will be denoted by $f_{j,gen}(s)$ and $F_{j,gen}(s)$ ($f_{j,imp}(s)$ and $F_{j,imp}(s)$), respectively, for $j = 1, 2, \ldots, R$. Figures 4.3 (a)-(f) give the plots of $f_{j,gen}(x)$ and $f_{j,imp}(x)$ ($j = 1, \ldots, R$) for the distributions of observed genuine and impostor match scores for $R = 3$ modalities of the MSU-Multimodal database (Jain et al., 2005). Figures 4.3 (a)-(f) also give the histograms of the genuine and impostor match scores for the three modalities. The "spikes" (see Figures 4.3 (d) and (e)) represent the detected discrete components whose individual heights are greater than the threshold $T = 0.02$. Note that the individual "spikes" cannot be represented by a continuous density function. Forcing a continuous density estimate for these values will result in gross density estimation errors and yield suboptimal performance of the multibiometric system.

The above procedure only estimates the marginal score distributions of each of the R matchers in the multibiometric system instead of estimating the joint distribution. The simplest approach to estimate the joint distribution is to assume statistical independence between the R matchers and estimate the joint distribution as the product of the R marginal distributions. In this case, the fused likelihood ratio (referred to as the product fusion score, $PFS(\boldsymbol{s})$) is the product of the likelihood ratios of the R matchers. Given the vector of match scores $\boldsymbol{s} = [s_1, \ldots, s_R]$, $PFS(\boldsymbol{s})$ is given by

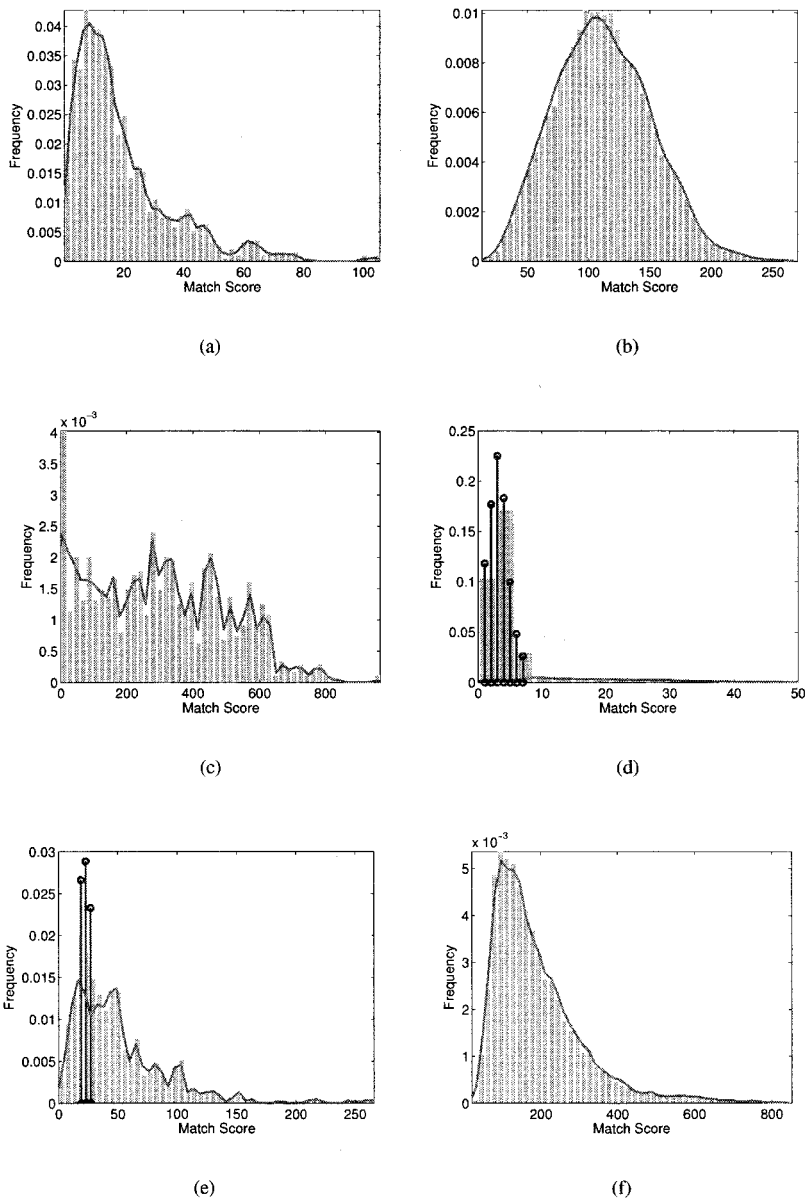


Figure 4.3. Histograms of match scores and corresponding generalized density estimates for MSU-Multimodal database. First Row: Histograms of match scores for face modality (a) genuine and (b) impostor. Second Row: Histograms of match scores for fingerprint modality (c) genuine and (d) impostor. Third Row: Histograms of match scores for hand geometry modality (e) genuine and (f) impostor. The solid line is the estimated density using the kernel density estimator, and the spikes in (d) and (e) correspond to the detected discrete components. Note that no score normalization needs to be performed before density estimation.

$$PFS(\boldsymbol{s}) = \frac{p(s_1, \ldots, s_R|genuine)}{p(s_1, \ldots, s_R|impostor)} = \prod_{j=1}^{R} \frac{f_{j,gen}(s_j)}{f_{j,imp}(s_j)}, \tag{4.47}$$

where $f_{j,gen}(\cdot)$ and $f_{j,imp}(\cdot)$ are the estimates of generalized densities of the genuine and impostor scores of the j^{th} biometric matcher. Hence, the decision rule can be stated as follows.

$$\text{Assign} \quad X \rightarrow genuine \quad \text{if}$$

$$PFS(\boldsymbol{s}) \geq \eta, \tag{4.48}$$

where η is the decision threshold.

However, a more appropriate procedure to estimate the joint density is to incorporate the correlation (if it exists) among the R matchers. One way to incorporate the correlation between the matchers is by using the copula models (Nelsen, 1999). Let $F_1, F_2, \ldots, F_R$ be R continuous distribution functions on the real line and F be a R-dimensional distribution function with the j^{th} marginal given by $F_j, j = 1, 2, \ldots, R$. According to Sklar's Theorem (Nelsen, 1999), there exists a unique function $C(u_1, u_2, \ldots, u_R)$ from $[0,1]^R \rightarrow [0,1]$ satisfying

$$F(s_1, s_2, \ldots, s_R) = C(F_1(s_1), F_2(s_2), \ldots, F_R(s_R)), \tag{4.49}$$

where $s_1, s_2, \ldots, s_R$ are R real numbers. The function C is known as a R-copula function that "couples" the one-dimensional distribution functions $F_1, F_2, \ldots, F_R$ to obtain the R-variate function F. Equation 4.49 can also be used to construct R-dimensional distribution function F whose marginals are the distributions $F_1, F_2, \ldots, F_R$.

Copula functions are effective in modeling a joint distribution whose marginal distributions are non-normal and do not have a parametric form (as is usually the case for biometric match scores). The family of copulas considered in Dass et al., 2005 is the R-dimensional multivariate Gaussian copulas (Cherubini et al., 2004). These functions can represent a variety of dependence structures among the R matchers using a $R \times R$ correlation matrix Σ_ρ. Note that multivariate Gaussian copulas do not assume that the joint or marginal distributions are Gaussian. They simply incorporate the second-order dependence in the form of a $R \times R$ correlation matrix. The R-dimensional Gaussian copula function with correlation matrix Σ_ρ is given by

$$C^R_{\Sigma_\rho}(u_1, \ldots, u_R) = \Phi^R_{\Sigma_\rho}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_R)), \tag{4.50}$$

where each $u_j \in [0,1]$ for $j = 1, \ldots, R$, $\Phi(\cdot)$ is the distribution function of the standard normal, $\Phi^{-1}(\cdot)$ is its inverse, and $\Phi^R_{\Sigma_\rho}$ is the R-dimensional distribution

function of a random vector $\mathcal{Z} = (Z_1, \ldots, Z_R)^T$ with component means and variances given by 0 and 1, respectively. The density of $C^R_{\Sigma_\rho}$, denoted by $c^R_{\Sigma_\rho}$, is given by

$$\begin{aligned} c^R_{\Sigma_\rho}(u_1, \ldots, u_R) &\equiv \frac{\partial C^R_{\Sigma_\rho}(u_1, \ldots, u_R)}{\partial u_1 \ldots \partial u_R} \\ &= \frac{\phi^R_{\Sigma_\rho}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_R))}{\prod_{j=1}^R \phi(\Phi^{-1}(u_j))}, \end{aligned} \tag{4.51}$$

where $\phi^R_{\Sigma_\rho}(v_1, \ldots, v_R)$ is the joint probability density function of the R-variate normal distribution with mean 0 and covariance matrix Σ_ρ, and $\phi(x)$ is the standard normal density function.

The (m, n)-th entry of Σ_ρ, ρ_{mn}, measures the degree of correlation between the scores of the m-th and n-th matchers for $m, n = 1, \ldots, R$. In practice, the correlation matrix Σ_ρ is unknown. We can estimate Σ_ρ using the product moment correlation of normal quantiles corresponding to the observed scores from the R matchers as follows. Suppose there are N training score vectors available for density estimation. Let $\boldsymbol{s}^i = [s^i_1, \ldots, s^i_R]$ denote the i^{th} score vector, $i = 1, \ldots, N$. The normal quantile of score s^i_j is given by

$$z^i_j = \Phi^{-1}\left(F_j\left(s^i_j\right)\right), \tag{4.52}$$

where $F_j(\cdot)$ denotes the j^{th} marginal distribution, $j = 1, \ldots, R$ and $i = 1, \ldots, N$. Thus, the i^{th} score vector $\boldsymbol{s}^i = [s^i_1, \ldots, s^i_R]$ is transformed to $\boldsymbol{z}^i = [z^i_1, \ldots, z^i_R]$, $i = 1, \ldots, N$. The covariance matrix of the N vectors, $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^N$, is estimated as follows.

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{z}^i - \bar{\boldsymbol{z}}\right)^T \left(\boldsymbol{z}^i - \bar{\boldsymbol{z}}\right), \tag{4.53}$$

where

$$\bar{\boldsymbol{z}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{z}^i. \tag{4.54}$$

The estimate of the (m, n)-th entry of Σ_ρ, $\hat{\rho}_{mn}$, is given by

$$\hat{\rho}_{mn} = \frac{\hat{\sigma}_{mn}}{\sqrt{\hat{\sigma}_{mm}\hat{\sigma}_{nn}}}, \tag{4.55}$$

where $\hat{\sigma}_{mn}$ is the (m, n)-th entry of $\hat{\Sigma}$.

The joint density function of genuine (impostor) match scores for R matchers, f^R_{gen} (f^R_{imp}) for some correlation matrix $\Sigma_{\rho,gen}$ ($\Sigma_{\rho,imp}$) is given by

$$f^R_{gen}(s_1,\ldots,s_R) = \left(\prod_{j=1}^{R} f_{j,gen}(s_j)\right) c^R_{\Sigma_{\rho,gen}}(F_{1,gen}(s_1),\ldots,F_{R,gen}(s_R)) \tag{4.56}$$

and

$$f^R_{imp}(s_1,\ldots,s_R) = \left(\prod_{j=1}^{R} f_{j,imp}(s_j)\right) c^R_{\Sigma_{\rho,imp}}(F_{1,imp}(s_1),\ldots,F_{R,imp}(s_R)). \tag{4.57}$$

Given the vector of match scores $\boldsymbol{s} = [s_1,\ldots,s_R]$, the likelihood ratio of the joint densities known as the copula fusion score $CFS(\boldsymbol{s})$, is given by

$$\begin{aligned} CFS(\boldsymbol{s}) &= \frac{f^R_{gen}(s_1,\ldots,s_R)}{f^R_{imp}(s_1,\ldots,s_R)} \\ &= PFS(\boldsymbol{s})\frac{c^R_{\Sigma_{\rho,gen}}(F_{1,gen}(s_1),\ldots,F_{R,gen}(s_R))}{c^R_{\Sigma_{\rho,imp}}(F_{1,imp}(s_1),\ldots,F_{R,imp}(s_R))}, \end{aligned} \tag{4.58}$$

where $F_{j,gen}(s_j)$ and $F_{j,imp}(s_j)$ are, respectively, the estimates of generalized distribution functions for the j^{th} biometric component, and $c^R_{\Sigma_\rho}$ is the density of $C^R_{\Sigma_\rho}$ as defined in Equation 4.51. The decision rule is given by

$$\text{Assign} \quad X \rightarrow genuine \quad \text{if}$$

$$CFS(\boldsymbol{s}) \geq \eta, \tag{4.59}$$

where η is the decision threshold.

Dass et al., 2005 demonstrated that fusion based on the generalized density estimates gives better performance over fusion based on continuous density estimates. The MSU-Multimodal database (Jain et al., 2005) collected from 100 users, with each user providing 5 face, fingerprint and hand geometry samples is used in this study. Fingerprint matching is done using the minutiae features (Jain et al., 1997b) and the output of the fingerprint matcher is a similarity score. Eigenface coefficients are used to represent features of the face image (Turk and Pentland, 1991). The Euclidean distance between the eigenface coefficients of the face template and that of the input face is used as the matching score. The hand geometry images are represented by a 14-dimensional feature

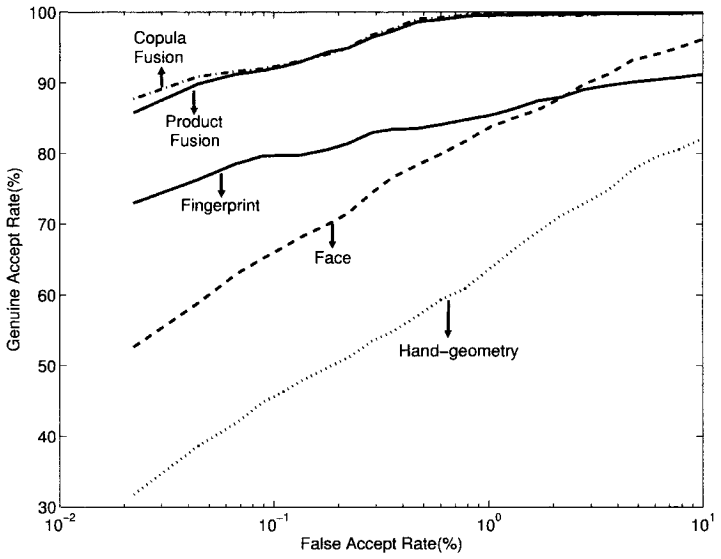
vector (Jain et al., 1999d) and the matching score is computed as the Euclidean distance between the input feature vector and the template feature vector. The histograms of the genuine and impostor scores of the three modalities in the MSU-Multimodal database are shown in Figure 4.3.

Figure 4.4 shows the ROC curves for the product and copula fusion rules (given by equations 4.48 and 4.59, respectively) and the ROC curves based on the match scores of the individual modalities for the MSU-Multimodal database. Figure 4.4(a) shows the recognition performance when the genuine and impostor score distributions of the three modalities are modeled purely by continuous densities, while Figure 4.4(b) gives the ROCs for generalized densities. Substantial performance improvement is obtained by modeling the match score distributions as a mixture of discrete and continuous components (generalized densities); for example, at a False Accept Rate (FAR) of 0.1%, the corresponding values of Genuine Accept Rate (GAR) for the continuous and generalized densities are 90.0% and 99.26%, respectively. Further, we can observe that although both the product and copula fusion rules give significantly better matching performance compared to the best individual modality, there is not much difference between the product and copula fusion rules. Dass et al., 2005 argue that this is due to the fact that the best modality in the MSU-Multimodal database is approximately independent (low correlation) of the other modalities, so the copula fusion involving more parameters than product fusion was not needed. The estimates of the correlation of the best single modality (fingerprint) with the other two modalities (face and hand geometry) were -0.01 and -0.11 for the genuine scores, and -0.05 and -0.04 for the impostor scores.

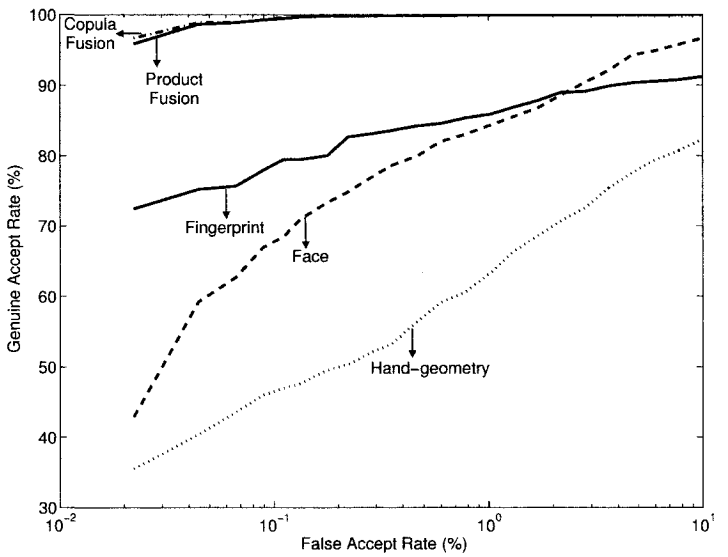
Dass et al., 2005 also applied the product and copula fusion rules on the match scores in the first partition of the Biometric Scores Set - Release I (BSSR1) released by NIST (see Appendix A.3 for more details). The ROC curves for the product and copula fusion rules on the NIST BSSR1 database are shown in Figure 4.5. In the NIST BSSR1 database, the correlation estimates of the best single modality (finger 2) with the other three modalities (face1, face2, and finger1 modalities, respectively) are -0.02, -0.06, and 0.43 for the genuine cases and 0.04, 0.02, and 0.14 for the impostor cases. Since the fusion is driven mostly by the best modality, the fact that this modality is approximately independent of the others means that the performances of product and copula fusion rules should be comparable to each other as reflected by the ROC curves in Figure 4.5.

4.5 Transformation-based score fusion

In practical multibiometric systems, the number of match scores available for training the fusion module is small due to the time, effort and cost involved in collecting multibiometric data. Due to the limited availability of training data, accurate estimation of the joint conditional densities $p(s =$



(a)



(b)

Figure 4.4. Performance of product and copula fusion on the MSU-Multimodal database based on (a) continuous and (b) generalized density estimates.

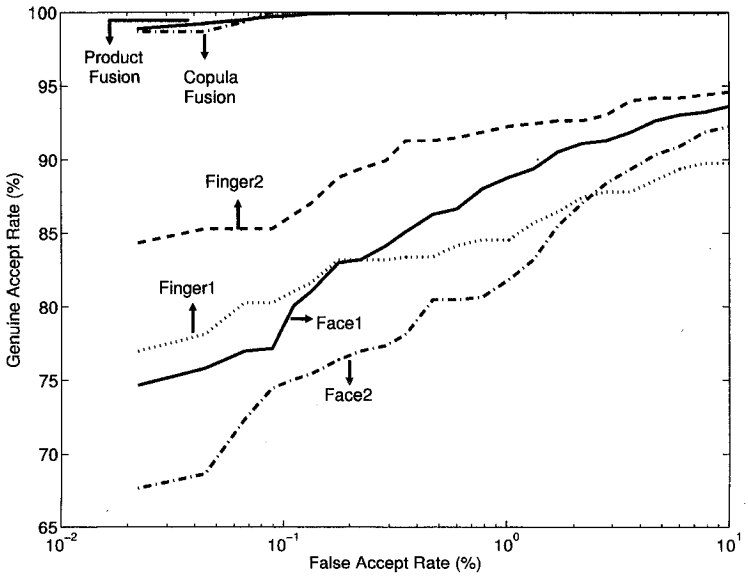


Figure 4.5. Performance of product and copula fusion on the NIST BSSR1 database.

$[s_1, \ldots, s_R]|genuine)$ and $p(\boldsymbol{s} = [s_1, \ldots, s_R]|impostor)$ is not always possible. In such situations, a more appropriate fusion method is to directly combine the match scores provided by different matchers without converting them into posteriori probabilities. However, the combination of match scores is meaningful only when the scores of the individual matchers are comparable. Hence, a transformation known as score normalization is applied to transform the match scores obtained from the different matchers into a common domain. The sum, max and min classifier combination rules developed by Kittler et al., 1998 (as discussed earlier in Section 4.2) can be applied to obtain the fused match scores from the normalized match scores. Since the normalized match scores do not have any probabilistic interpretation, the sum, max and min rules are referred to as sum of scores, max score and min score fusion rules, respectively. The max score and min score fusion rules are referred to as order statistics. The combined match score can also be computed as a weighted sum of the match scores of the individual matchers (Ross and Jain, 2003; Wang et al., 2003), which is known as the weighted sum of scores rule (or simply, weighted sum rule).

Figures 4.3(a)-(f) show the conditional distributions of the face, fingerprint and hand geometry match scores of the MSU-Multimodal database used in ex-

periments by Jain et al., 2005. The scores obtained from the face and hand geometry matchers are distance scores whereas those obtained from the fingerprint matcher are similarity scores. One can also observe the non-homogeneity (differences in the numerical scale and statistical distributions) in these scores demonstrating the need for score normalization prior to any meaningful combination.

4.5.1 Score Normalization

Score normalization refers to changing the location and scale parameters of the match score distributions at the outputs of the individual matchers, so that the match scores of different matchers are transformed into a common domain. When the parameters used for normalization are determined using a fixed training set, it is referred to as *fixed score normalization* (Brunelli and Falavigna, 1995). In such a case, the set of match scores available for training the fusion module of a multibiometric system is examined and a suitable statistical model is chosen to fit to the data. Based on the model, the score normalization parameters are determined. In *adaptive score normalization*, the normalization parameters are estimated based on the match score of the current test sample. This approach has the ability to adapt to variations in the input data such as the changes in the duration of the speech signals in speaker recognition systems.

For a good normalization scheme, the estimates of the location and scale parameters of the match score distribution must be *robust* and *efficient*. *Robustness* refers to insensitivity to the presence of outliers whereas *efficiency* refers to the proximity of the obtained estimates to the optimal estimates when the distribution of the data is known. Huber, 1981 explains the concepts of robustness and efficiency of statistical procedures and emphasizes the need for statistical procedures that have both these desirable characteristics. Although many techniques can be used for score normalization, the challenge lies in identifying a technique that is both robust and efficient.

The simplest normalization technique is the *min-max* normalization. Min-max normalization is best suited for the case where the bounds (maximum and minimum values) of the scores produced by a matcher are known. In this case, we can easily transform the minimum and maximum scores to 0 and 1, respectively. However, even if the match scores are not bounded, we can estimate the minimum and maximum values for the given set of training match scores and then apply the min-max normalization. Let s_j^i denote the i^{th} match score output by the j^{th} matcher, $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, R$ (R is the number of matchers and N is the number of match scores available in the training set). The min-max normalized score, ns_j^t, for the test score s_j^t is given by

$$ns_j^t = \frac{s_j^t - \min_{i=1}^{N} s_j^i}{\max_{i=1}^{N} s_j^i - \min_{i=1}^{N} s_j^i}. \tag{4.60}$$

When the minimum and maximum values are estimated from the given set of match scores, this method is not robust (i.e., the method is sensitive to outliers in the data used for estimation). Min-max normalization retains the original distribution of scores except for a scaling factor and transforms all the scores into a common range $[0, 1]$. Distance scores can be transformed into similarity scores by subtracting the normalized score from 1.

Decimal scaling can be applied when the scores of different matchers are on a logarithmic scale. For example, if one matcher has scores in the range $[0, 10]$ and the other has scores in the range $[0, 1000]$, the following normalization could be applied to transform the scores of both the matchers to the common $[0, 1]$ range.

$$ns_j^t = \frac{s_j^t}{10^{n_j}}, \tag{4.61}$$

where $n_j = \log_{10} \max_{i=1}^{N} s_j^i$. In the example with two matchers where the score ranges are $[0, 10]$ and $[0, 1000]$, the values of n would be 1 and 3, respectively. The problems with this approach are the lack of robustness and the implicit assumption that the scores of different matchers vary by a logarithmic factor.

The most commonly used score normalization technique is the *z-score* normalization that uses the arithmetic mean and standard deviation of the training data. This scheme can be expected to perform well if the average and the variance of the score distributions of the matchers are available. If we do not know the values of these two parameters, then we need to estimate them based on the given training set. The z-score normalized score is given by

$$ns_j^t = \frac{s_j^t - \mu_j}{\sigma_j}, \tag{4.62}$$

where μ_j is the arithmetic mean and σ_j is the standard deviation for the j^{th} matcher. However, both mean and standard deviation are sensitive to outliers and hence, this method is not robust. Z-score normalization does not guarantee a common numerical range for the normalized scores of the different matchers. If the distribution of the scores is not Gaussian, z-score normalization does not preserve the distribution of the given set of scores. This is due to the fact that mean and standard deviation are the optimal location and scale parameters only for a Gaussian distribution. While mean and standard deviation are reasonable estimates of location and scale, respectively, they are not optimal for an arbitrary match score distribution.

The *median* and *median absolute deviation* (MAD) statistics are insensitive to outliers as well as points in the extreme tails of the distribution. Hence, a normalization scheme using median and MAD would be robust and is given by

$$ns_j^t = \frac{s_j^t - med_j}{MAD_j}, \tag{4.63}$$

where $med_j = median_{i=1}^{N} s_j^i$ and $MAD_j = median_{i=1}^{N} |s_j^i - med_j|$. However, the median and the MAD estimators have a low efficiency compared to the mean and the standard deviation estimators, i.e., when the score distribution is not Gaussian, median and MAD are poor estimates of the location and scale parameters. Therefore, this normalization technique does not preserve the input score distribution and does not transform the scores into a common numerical range.

Cappelli et al., 2000 use a *double sigmoid function* for score normalization in a multibiometric system that combines different fingerprint matchers. The normalized score is given by

$$ns_j^t = \begin{cases} \frac{1}{1+\exp\left(-2\left(\frac{s_j^t - \tau}{\alpha_1}\right)\right)} & \text{if } s_j^t < \tau, \\ \frac{1}{1+\exp\left(-2\left(\frac{s_j^t - \tau}{\alpha_2}\right)\right)} & \text{otherwise,} \end{cases} \tag{4.64}$$

where τ is the reference operating point and α_1 and α_2 denote the left and right edges of the region in which the function is linear. The double sigmoid function exhibits linear characteristics in the interval $(\tau - \alpha_1,\ \tau - \alpha_2)$. Figure 4.6 shows an example of the double sigmoid normalization, where the scores in the $[0, 300]$ range are mapped to the $[0, 1]$ range using $\tau = 200$, $\alpha_1 = 20$ and $\alpha_2 = 30$.

While the double sigmoid normalization scheme transforms the scores into the $[0, 1]$ interval, it requires careful tuning of the parameters τ, α_1 and α_2 to obtain good efficiency. Generally, τ is chosen to be some value falling in the region of overlap between the genuine and impostor score distributions, and α_1 and α_2 are set so that they correspond to the extent of overlap between the two distributions toward the left and right of τ, respectively. This normalization scheme provides a linear transformation of the scores in the region of overlap, while the scores outside this region are transformed non-linearly. The double sigmoid normalization is very similar to the min-max normalization followed by the application of a two-quadrics (QQ) or a logistic (LG) function as suggested by Snelick et al., 2005. When the values of α_1 and α_2 are large, the double sigmoid normalization closely resembles the QQ-min-max normalization. On

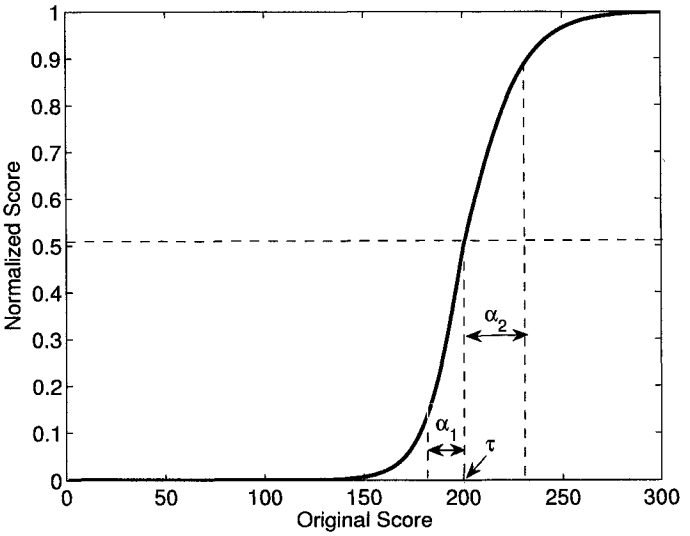


Figure 4.6. Double sigmoid normalization with $\tau = 200$, $\alpha_1 = 20$, and $\alpha_2 = 30$.

the other hand, we can make the double sigmoid normalization approach toward LG-min-max normalization by assigning small values to α_1 and α_2.

The *tanh-estimators* introduced by Hampel et al., 1986 are robust and highly efficient. The tanh normalization is given by

$$ns_j^t = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{s_j^t - \mu_{GH}}{\sigma_{GH}}\right)\right) + 1\right\}, \tag{4.65}$$

where μ_{GH} and σ_{GH} are the mean and standard deviation estimates, respectively, of the genuine score distribution as given by Hampel estimators. Hampel estimators are based on the following influence (ψ)-function:

$$\psi(u) = \begin{cases} u & 0 \leq |u| < a, \\ a * sign(u) & a \leq |u| < b, \\ a * sign(u) * \left(\frac{c-|u|}{c-b}\right) & b \leq |u| < c, \\ 0 & |u| \geq c, \end{cases} \tag{4.66}$$

where

$$sign\{u\} = \begin{cases} +1, & \text{if } u \geq 0, \\ -1, & \text{otherwise.} \end{cases} \tag{4.67}$$

A plot of the Hampel influence function is shown in Figure 4.7. The Hampel influence function reduces the influence of the scores at the tails of the distribution (identified by a, b, and c) during the estimation of the location and scale parameters. Hence, this method is not sensitive to outliers. If many of the points that constitute the tail of the distributions are discarded, the estimate is robust but not efficient (optimal). On the other hand, if all the points that constitute the tail of the distributions are considered, the estimate is not robust but its efficiency increases. Therefore, the parameters a, b, and c must be carefully chosen depending on the amount of robustness required which in turn depends on the amount of noise in the available training data.

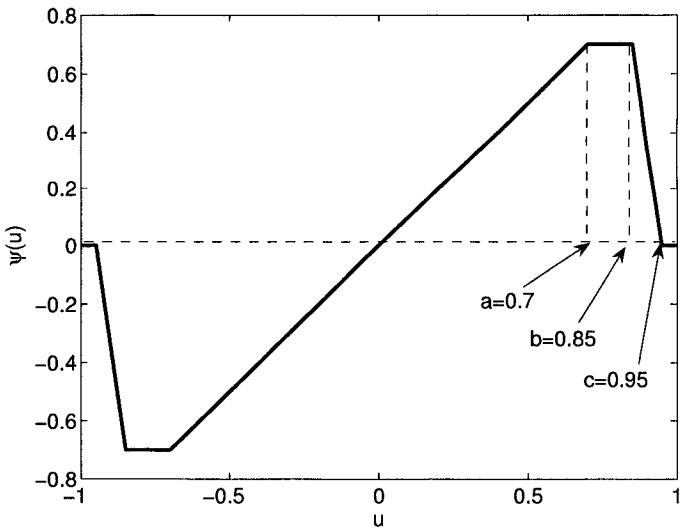


Figure 4.7. Hampel influence function with $a = 0.7$, $b = 0.85$, and $c = 0.95$.

Mosteller and Tukey, 1977 introduce the biweight location and scale estimators that are robust and efficient. But, the *biweight estimators* are iterative in nature (initial estimates of the biweight location and scale parameters are chosen, and these estimates are updated based on the training scores), and are applicable only for Gaussian data. A summary of the characteristics of the different normalization techniques discussed here is shown in Table 4.1. The min-max, decimal scaling and z-score normalization schemes are efficient, but are not robust to outliers. On the other hand, the median normalization scheme is robust but inefficient. Only the double sigmoid and tanh-estimators have both the desired characteristics, namely, robustness and efficiency.

Table 4.1. Summary of score normalization techniques.

Normalization Technique	Robustness	Efficiency
Min-max	No	High
Decimal scaling	No	High
Z-score	No	High
Median and MAD	Yes	Moderate
Double sigmoid	Yes	High
Tanh-estimators	Yes	High

4.5.2 Evaluation of normalization techniques

It must be noted that no normalization scheme has been shown to be optimal for all kinds of match score data. Hence, a number of score normalization techniques are admissible, i.e., they may work better than other normalization techniques depending on the fusion problem at hand. In practice, it is recommended that a number of normalization techniques be evaluated to determine the one that gives the best performance on the given data. Jain et al., 2005 studied the performance of a multimodal biometric system comprising of face, fingerprint and hand geometry modalities under different normalization and fusion techniques. They used the MSU-Multimodal database for these experiments. The simple sum of scores, the max-score, and the min-score fusion methods were applied on the normalized scores. The normalized scores were obtained by using the following techniques: simple distance-to-similarity transformation with no change in scale (STrans), min-max normalization (Minmax), z-score normalization (ZScore), median-MAD normalization (Median), double sigmoid normalization (Sigmoid), tanh normalization (Tanh), and Parzen normalization (Parzen). Note that the conversion of match scores into posteriori probabilities by the Parzen window density estimation method is really not a normalization technique. It actually falls under the density-based match score fusion approach. However, Jain et al., 2005 treat the ratio of the posteriori probabilities of the genuine and impostor classes as a normalized match score and hence, they refer to this method as Parzen normalization.

The recognition performance of the face, fingerprint, and hand geometry modalities in the MSU-Multimodal database is shown in Figure 4.8. We observe that the fingerprint modality gives the best performance followed by the face and hand geometry modalities in that order. At a False Accept Rate (FAR) of 0.1%, the Genuine Accept Rates (GAR) are 83.6%, 67.7% and 46.8% for the fingerprint, face and hand geometry modalities, respectively.

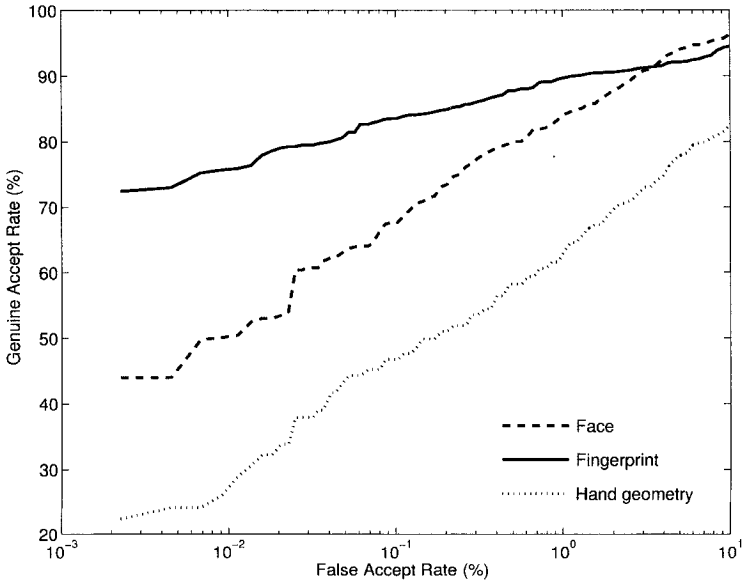


Figure 4.8. ROC curves for the individual modalities in the MSU-Multimodal database.

To evaluate the performance after fusion, the set of match scores obtained from the MSU-Multimodal database was randomly partitioned into training (60% of the scores were used for estimating the normalization parameters) and test (the remaining 40% of the scores were used for evaluating the performance of the multibiometric system) sets. This random splitting of the database into training and test sets was repeated 40 times resulting in 40 trials. Table 4.2 summarizes the average (over the 40 trials) Genuine Accept Rate (GAR) of the multimodal system along with the standard deviation of the GAR (shown in parentheses) for different normalization and fusion schemes, at a False Accept Rate (FAR) of 0.1%. From Table 4.2, it is apparent that the sum of scores method provides better recognition performance than the max-score and min-score methods. Hence, we compare the different normalization techniques only for the sum of scores fusion method.

Figure 4.9 shows the recognition performance of the multimodal system when the scores that are normalized using various techniques described above, are combined using the sum of scores method. We observe that a multimodal system employing the sum of scores method provides better performance than the best unimodal system (fingerprint in this case) for all normalization techniques except median-MAD normalization. For example, at a FAR of 0.1%, the GAR of the fingerprint module is about 83.6%, while that of the multimodal

Table 4.2. Genuine Accept Rate (GAR) (%) of different normalization and fusion techniques at the 0.1% False Accept Rate (FAR) for the MSU-Multimodal database. At 0.1% FAR, the GAR of the unimodal systems are 83.6%, 67.7% and 46.8% for the fingerprint, face and hand geometry modalities, respectively. Note that the values in the table represent average GAR, and the values indicated in parentheses correspond to the standard deviation of GAR computed over the 40 trials of randomly splitting the available data into training and test sets.

Normalization Technique	**Fusion Technique**		
	Sum of scores	**Max-score**	**Min-score**
STrans	98.3 (0.4)	46.7 (2.3)	83.9 (1.6)
Minmax	97.8 (0.6)	67.0 (2.5)	83.9 (1.6)
Zscore	98.6 (0.4)	92.1 (1.1)	84.8 (1.6)
Median	84.5 (1.3)	83.7 (1.6)	68.8 (2.2)
Sigmoid	96.5 (1.3)	83.7 (1.6)	83.1 (1.8)
Tanh	98.5 (0.4)	86.9 (1.8)	85.6 (1.5)
Parzen	95.7 (0.9)	93.6 (2.0)	83.9 (1.9)

system is 98.6% when z-score normalization is used. The performance of the multimodal biometric system is a significant improvement over the best unimodal system and it underscores the benefit of deploying multimodal systems.

Among the various normalization techniques on this dataset, we observe that the tanh and min-max normalization techniques outperform other techniques at low FARs. At higher FARs, z-score normalization provides slightly better performance than tanh and min-max normalization. In the multimodal system based on the MSU-Multimodal database, the combined score of test pattern, s^t_{fus}, after sum of scores fusion is just a linear transformation of the score vector $\boldsymbol{s}^t = [s^t_1, s^t_2, s^t_3]$, i.e., $s^t_{fus} = (a_1 s^t_1 - b_1) + (a_2 s^t_2 - b_2) + (a_3 s^t_3 - b_3)$, where s^t_1, s^t_2, and s^t_3 correspond to the match scores for the test pattern obtained from the face, fingerprint and hand geometry matchers, respectively. The effect of different normalization techniques is to determine the weights a_1, a_2, and a_3, and the biases b_1, b_2, and b_3. Since the value of the MAD statistic for the fingerprint scores is very small compared to that of face and hand geometry scores, the median-MAD normalization assigns a much larger weight to the fingerprint score ($a_2 >> a_1$ and $a_2 >> a_3$). This is a direct consequence of the moderate efficiency of the median-MAD estimator. Since the distribution of the fingerprint scores (see Figures 4.3(c) and (d)) deviates drastically from the Gaussian assumption, the median and MAD statistics are not the correct measures of location and scale, respectively. In this case, the combined score is approximately equal to the fingerprint score and the performance of the multimodal system is close to that of the fingerprint module. On the

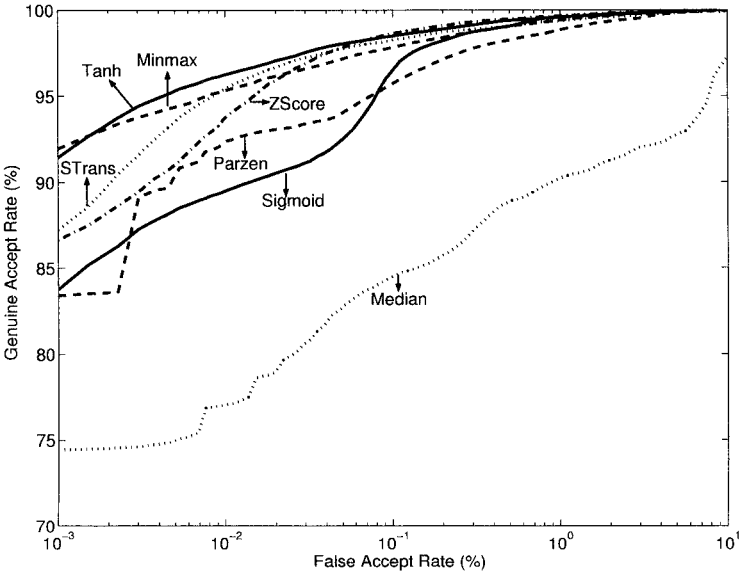


Figure 4.9. ROC curves for sum of scores fusion method under different normalization schemes on the MSU-Multimodal database.

other hand, min-max normalization, z-score normalization, tanh and distance-to-similarity transformation assign more reasonable weights to the scores of the three modalities. Therefore, the recognition performance of the multimodal system applying one of these four normalization techniques (min-max, z-score, tanh and distance-to-similarity transformation) along with the sum of scores fusion method is significantly better than that of the fingerprint matcher. The difference in performance between the min-max, z-score, tanh and distance-to-similarity transformation is relatively small. However, it should be noted that the raw scores of the three modalities used in the experiments are comparable and, hence, a simple distance-to-similarity conversion works reasonably well here. If the match scores of the three modalities were significantly different, then the distance-to-similarity transformation method would not work as well.

For sum of scores fusion, we observe that the performance of a robust normalization technique like tanh is almost the same as that of the non-robust techniques like min-max and z-score normalization. However, the performance of such non-robust techniques is highly dependent on the accuracy of the estimates of the location and scale parameters. The scores produced by the matchers used by Jain et al., 2005 are unbounded and, hence, can theoretically produce any value in the interval $(0, \infty)$. Also, the statistics of the scores (e.g., aver-

age or deviation from the average) produced by these three matchers will not be known. Therefore, parameters like the average and standard deviation of scores (needed for z-score normalization) have to be estimated from the available data. The particular data set used in these experiments did not contain any outliers and, hence, the performance resulting from the use of non-robust normalization techniques was not degraded.

In order to demonstrate the sensitivity of the min-max and z-score normalization techniques in the presence of outliers, Jain et al., 2005 artificially introduced outliers in the fingerprint scores. For min-max normalization, a single outlier whose value is 75%, 125%, 150%, 175% or 200% of the maximum score (in the training set) is introduced. Figure 4.10 shows the recognition performance of the multimodal system after the introduction of the outlier. We observe that the performance is sensitive to the maximum score. A single outlier that is twice the original maximum score can reduce the recognition rate of the multimodal system by 3-5% depending on the operating point of the system. The performance degradation is more severe at lower values of FAR.

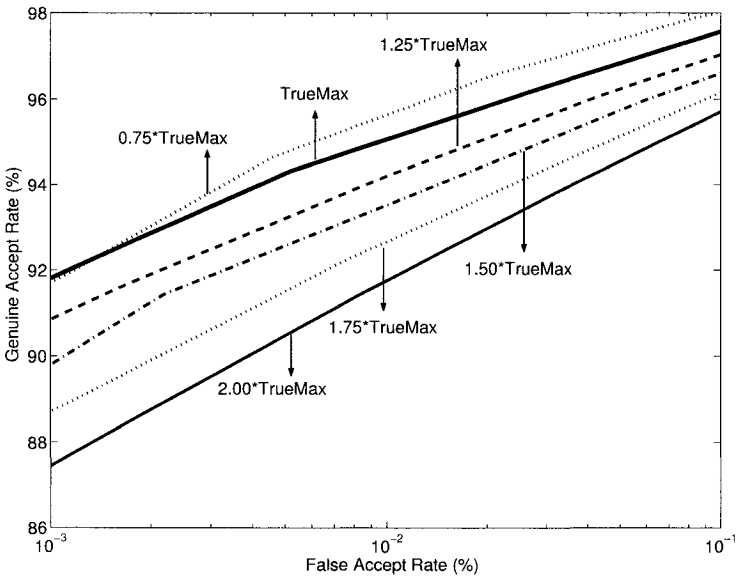


Figure 4.10. Robustness analysis of min-max normalization. Note that TrueMax represents the maximum fingerprint match score in the training set. The different ROC curves are obtained by replacing the maximum fingerprint score in the training set with an outlier score whose value is 75%, 125%, 150%, 175% or 200% of TrueMax.

In the case of z-score normalization, several outliers were introduced in the fingerprint match scores so that the standard deviation of the fingerprint score is increased by 125%, 150%, 175% or 200% of the original standard deviation. In one trial, some large match scores were reduced to decrease the standard deviation of the scores to 75% of the original value. In the case of an increase in standard deviation, the performance improves after the introduction of outliers as indicated in Figure 4.11. Since the original standard deviation was small, fingerprint scores were assigned a higher weight compared to the other modalities. As the standard deviation is increased, the dominance of the fingerprint modality was reduced and this resulted in improved recognition rates. However, the goal of this experiment was to show the sensitivity of the system to those estimated parameters that can be easily affected by outliers. A similar experiment was done for tanh normalization technique and, as shown in Figure 4.12, there is no significant variation in the performance of the tanh normalization method after the introduction of outliers. This result highlights the robustness of the tanh normalization method.

In many cases, the maximum and minimum score output by a matcher will be known in advance. Therefore, the min-max normalization scheme *may* not require an explicit estimation procedure based on the training data.

4.6 Classifier-based score fusion

In classifier-based score fusion, a pattern classifier (Duda et al., 2001) is used to indirectly learn the relationship between the vector of match scores $[s_1, s_2, \ldots, s_R]$ provided by the R biometric matchers and the posteriori probabilities of the genuine and impostor classes, namely, $P(genuine|s_1, s_2, \ldots, s_R)$ and $P(impostor|s_1, s_2, \ldots, s_R)$. In this approach, the vector of match scores $[s_1, s_2, \ldots, s_R]$ is treated as a feature vector which is then classified into one of two classes: "genuine user" or "impostor". Based on the training set of match scores from the genuine and impostor classes, the classifier learns a decision boundary between the two classes. Figure 4.13 shows an example of a linear decision boundary learned by a classifier based on the genuine and impostor match scores from two different matchers. During verification, any match score vector that falls in the genuine region (to the right of the decision boundary in Figure 4.13) is classified as "genuine". In general, the decision boundary can be quite complex depending on the nature of the classifier. However, the classifier is capable of learning the decision boundary irrespective of how the feature vectors are generated. Hence, the output scores of the different matchers can be non-homogeneous (distance or similarity metric, different numerical ranges, etc.) and no processing is required prior to designing the classifier. A limitation of the classifier-based score fusion approach is that it is not easy to fix one type of error (say FAR) and then compute the FRR at the specified FAR.

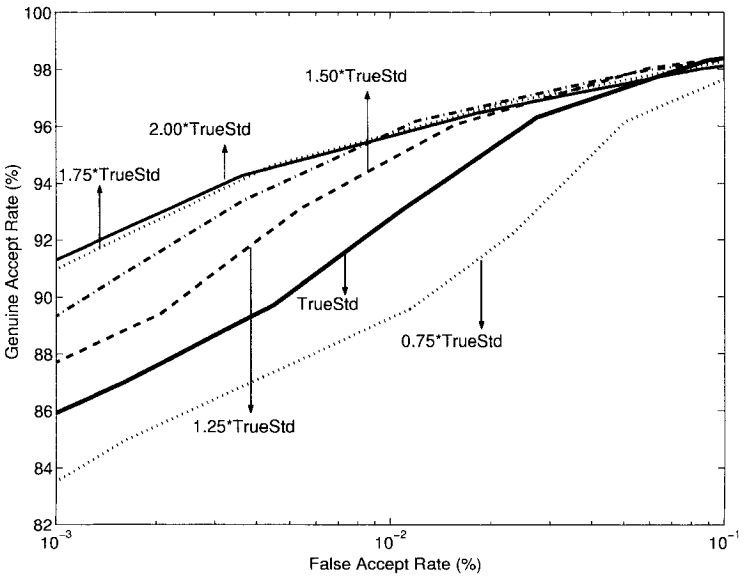


Figure 4.11. Robustness analysis of z-score normalization. Note that TrueStd represents the standard deviation of the fingerprint match scores in the training set. The different ROC curves are obtained by introducing outlier scores in the training set so that the standard deviation of the fingerprint match scores is changed to 75%, 125%, 150%, 175% or 200% of TrueStd.

Several classifiers have been used to consolidate the match scores of multiple matchers and arrive at a decision. Brunelli and Falavigna, 1995 use a HyperBF network to combine matchers based on voice and face features. The speaker recognition subsystem was based on vector quantization of the acoustic parameter space and included an adaptation phase of the codebooks to the test environment. Face identification was achieved by analyzing three facial components, namely, eyes, nose, and mouth. The basic template matching technique was applied for face matching. While the rank-one recognition rates of the voice and face matchers were 88% and 91%, respectively, the fusion of these two matchers achieved a rank-one recognition rate of 98%.

Verlinde and Cholet, 1999 compare the relative performance of three different classifiers, namely, the k-Nearest Neighbor classifier using vector quantization, the decision tree classifier, and the classifier based on logistic regression model when used for the fusion of match scores from three biometric matchers. The three matchers were based on profile face image, frontal face image, and voice. Experiments by Verlinde and Cholet, 1999 on the multimodal M2VTS database (Pigeon and Vandendrope, 1996) show that the total error rate (sum of the

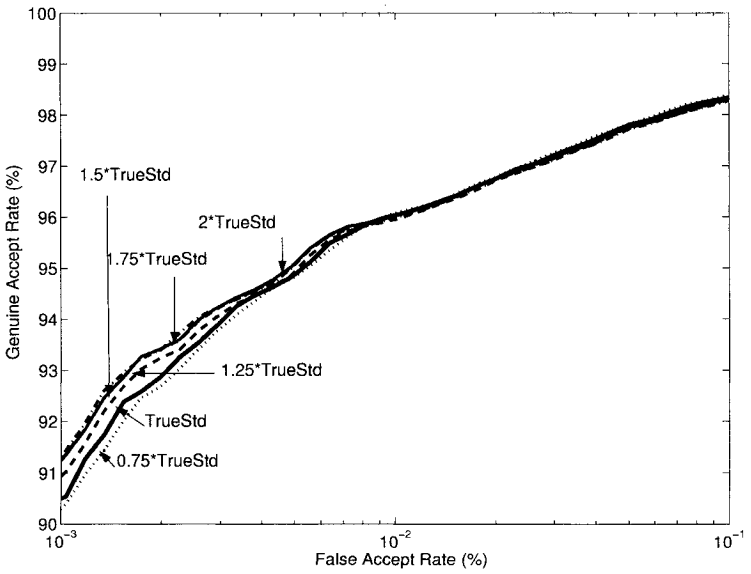


Figure 4.12. Robustness analysis of tanh normalization. Note that TrueStd represents the standard deviation of the fingerprint match scores in the training set. The different ROC curves are obtained by introducing outlier scores in the training set so that the standard deviation of the fingerprint match scores is changed to 75%, 125%, 150%, 175% or 200% of TrueStd.

false accept and false reject rates) of the multimodal system was an order of magnitude less than that of the individual modalities. While the total error rates of the individual modalities were 8.9% for profile face, 8.7% for frontal face, and 3.7% for speaker verification, the total error rate of the multimodal system was found to be 0.1% when the classifier based on logistic regression model was employed.

Chatzis et al., 1999 use classical k-means clustering, fuzzy clustering and median radial basis function (MRBF) algorithms for fusion at the match score level. Five biometric matchers that were based on the grey-level and shape information of face image and voice features were employed. Each matcher provided a match score and a quality metric that measures the reliability of the match score, and these values were concatenated to form a ten-dimensional vector. Clustering algorithms were applied on this ten-dimensional feature vector to form two clusters, namely, genuine and impostor.

Ben-Yacoub et al., 1999 evaluate a number of classification schemes for fusion of match scores from multiple modalities, including support vector machine (SVM) with polynomial kernels, SVM with Gaussian kernels, C4.5 decision

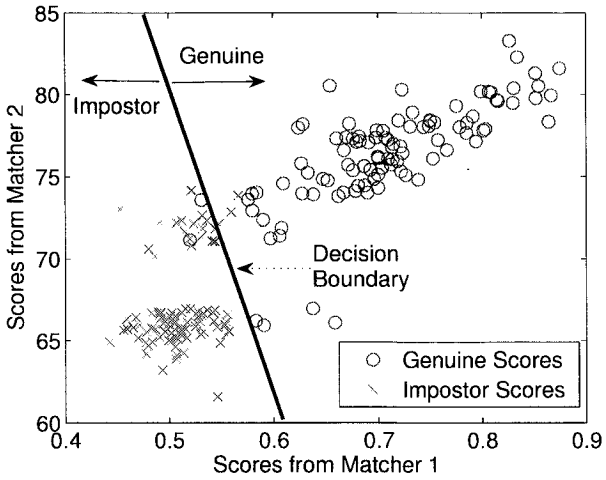


Figure 4.13. Example of a linear decision boundary learned by a classifier in a 2-dimensional ($R = 2$) feature space. During verification, any match score vector that falls in the region marked as 'Genuine' (to the right of the decision boundary) is classified as "genuine user". On the other hand, any match score vector that falls in the region marked as 'Impostor' (to the left of the decision boundary) is classified as "impostor".

trees, multilayer perceptron, Fisher linear discriminant, and Bayesian classifier. This evaluation is conducted on the XM2VTS database (Messer et al., 1999) consisting of 295 subjects. The database includes four recordings of each person obtained at one month intervals. During each session, two recordings were made: a speech shot and a head rotation shot. The speech shot was composed of the frontal face recording of each subject during the dialogue. Face recognition was performed by using elastic graph matching (EGM) (Lades et al., 1993). Two different approaches were used for speaker verification. A sphericity measure (Bimbot et al., 1995) was used for text-independent speaker verification. Hidden Markov models (HMM) were used for text-dependent speaker verification. The total error rate of 0.6% achieved by the Bayesian classifier was significantly lower than the total error rate of 1.48% achieved by the HMM based speaker recognition system, which was the best individual modality in terms of total error rate.

Bigun et al., 1997 propose a new algorithm based on the Bayesian classifier for fusion in a multibiometric system. Their model takes into account the estimated accuracy of the individual classifiers during the fusion process. Sanderson and Paliwal, 2002 use a support vector machine (SVM) to combine the scores of face and speech experts. They show that the performance of such a classifier deteriorates under noisy input conditions. To overcome this

problem, they implement structurally noise-resistant classifiers like piece-wise linear classifier and modified Bayesian classifier. Wang et al., 2003 consider the match scores resulting from face and iris recognition modules as a two-dimensional feature vector and use Fisher's discriminant analysis and a neural network classifier with radial basis function to classify the 2-dimensional match score vector into "genuine" and "impostor" classes. Ross and Jain, 2003 use decision tree and linear discriminant classifiers for combining the match scores of face, fingerprint and hand geometry modalities. Random forest algorithm was used by Ma et al., 2005 for the classification of 3-dimensional match score vectors described in Ross and Jain, 2003 into "genuine" and "impostor" classes.

4.7 Comparison of score fusion techniques

The existence of a large number of score fusion techniques makes it difficult for the designer of a multibiometric system to select an appropriate fusion method for the problem at hand. Most of these score fusion techniques have not been tested on benchmark databases. Recently, the National Institute of Standards and Technology (NIST) released a true multimodal match score database known as the Biometric Score Set Release-1 (National Institute of Standards and Technology, 2004) containing the face and fingerprint matching scores of 517 individuals (see Appendix for more details). Also, a benchmark match score database based on the XM2VTS multimodal dataset (face and voice modalities) has been released by IDIAP (Poh and Bengio, 2005a). The emergence of these benchmark databases is likely to result in a more careful and thorough evaluation of score fusion techniques. In this section, we briefly compare some of the score fusion techniques based on the NIST BSSR1 database. In order to clearly illustrate the differences in recognition performance of the various fusion schemes, we consider the match scores corresponding to only two biometric matchers in this database: the `Face-G` matcher and the fingerprint matcher corresponding to the right index fingerprint. Henceforth, we refer to these two matchers simply as face and fingerprint matchers, respectively.

Firstly, we compare the recognition performance of the various classifier combination rules (Kittler et al., 1998) described in Section 4.2. As pointed out earlier in Section 4.3, we assume that the feature representations of the two biometric matchers $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not available. Hence, the posteriori probabilities $P(genuine|\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $P(impostor|\boldsymbol{x}_1, \boldsymbol{x}_2)$ are estimated from the vector of match scores $\boldsymbol{s} = [s_1, s_2]$, where s_1 and s_2 are the match scores provided by the face and fingerprint matchers, respectively. Further, we adopt the density-based score fusion approach and estimate the conditional densities $p(s_1|genuine)$, $p(s_1|impostor)$, $p(s_2|genuine)$ and $p(s_2|impostor)$ using a non-parametric density estimation technique. Specifically, we use a Gaussian kernel density estimator and the bandwidth of the kernel is obtained using the "solve-the-equation" bandwidth estimator (Wand and Jones, 1995). We also

assume that the prior probabilities of the genuine and impostor classes are equal. 80% of the genuine and impostor scores are used for estimating the conditional densities and the remaining 20% are used for evaluating the fusion performance. Five-fold cross validation is performed and the reported results correspond to the average over the five trials.

Figure 4.14 shows the ROC curves for the face and fingerprint unibiometric systems as well as the ROC curves for the multibiometric system, when fusion is performed using the product, sum, max and min rules. We can clearly see that the fingerprint matcher is more accurate than the face matcher. For the selected face and fingerprint matchers in the NIST BSSR1 database, we also observe that fusion using the product rule gives the best recognition performance. It must be noted that the only assumption in deriving the product rule is the independence between the biometric matchers. The sum, max and min rules are derived by introducing other constraints to the product rule. This explains the relatively good performance of the product rule compared to the other fusion rules. In general, the sum rule has been shown to perform well because it is less sensitive to errors in the probability estimates (Kittler et al., 1998). However, the sum rule is derived based on the strong assumption that genuine and impostor classes are highly ambiguous, and that the observed match scores enhance the prior class probabilities marginally. This assumption is not valid for the match scores in the NIST BSSR1 database because the genuine and impostor score distributions have only a small region of overlap. Therefore, the sum rule does not provide any improvement in recognition performance over the best unimodal biometric system (fingerprint in this case).

Secondly, we compare some of the transformation-based score fusion techniques discussed in Section 4.5. Specifically, the match scores provided by the face and fingerprint matchers are first normalized using five different normalization techniques, viz., min-max, z-score, median-MAD, double sigmoid and tanh techniques. The normalized match scores are then combined using the sum of scores method. Again, five-fold cross validation is performed using 80% of the genuine and impostor scores for estimating the normalization parameters, and the remaining 20% for evaluating the performance of the fusion technique. Figure 4.15 summarizes the matching performance of the multimodal biometric system when the normalized scores are combined using the sum of scores method. We observe that the multimodal system results in better performance than the best unimodal system (fingerprint in this case). Among the various normalization schemes, the min-max and tanh normalization techniques result in the best performance on the NIST BSSR1 database.

Finally, we evaluate the performance of a classifier-based score fusion scheme on the same data set. As indicated earlier, the classifier-based approach assigns a two-dimensional match score vector to one of the two classes, namely, genuine and impostor. A Support Vector Machine (SVM) with a radial basis

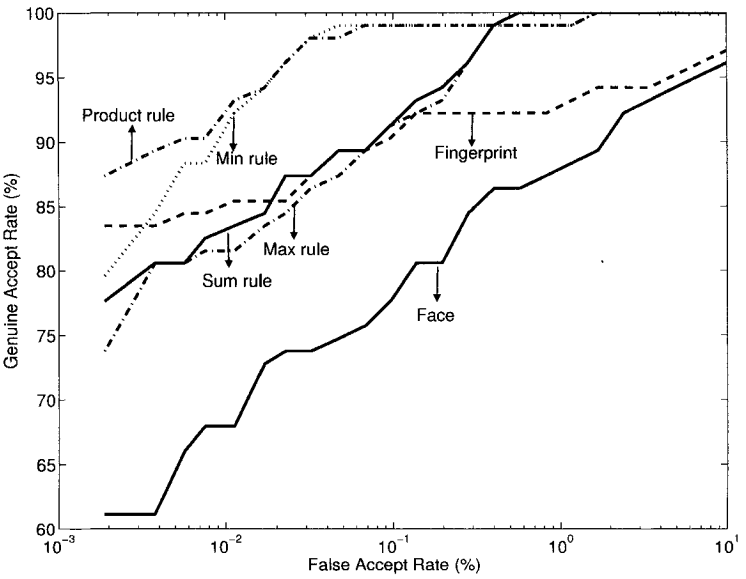


Figure 4.14. Comparison of recognition performance of the classifier combination rules proposed by Kittler et al., 1998 on the NIST BSSR1 database. In this experiment, the match scores are converted into probabilities using a non-parametric density estimation technique.

function kernel is used for the classification and a five-fold cross-validation is performed. The SVM classifier has an average FAR of 0.01% (standard deviation of 0.008%) and an average GAR of 95.1% (standard deviation of 3.4%). The performance of the SVM classifier is shown in Figure 4.16. Note that the operating point (FAR and the corresponding GAR) of a classifier-based score fusion scheme can be changed by tuning the parameters of the classifier. However, it is not always possible to fix the FAR and then compute the corresponding GAR in classifier-based score fusion.

Figure 4.16 compares the performance of the best transformation-based score fusion technique (tanh normalization followed by sum of scores fusion) and the best density-based score fusion approach (product rule). We see that fusion based on the product rule has a lower recognition rate than the sum of scores fusion method (after tanh normalization) at small values of False Accept Rate (FAR). For example, at a FAR of 0.01%, the average Genuine Accept Rate (GAR) of the product rule is 92.4% while the sum of scores fusion has an average GAR of 96.5%. This may be due to the limited availability of genuine match scores for estimating the conditional densities $p(s_1|genuine)$ and $p(s_2|genuine)$, where s_1 and s_2 are the match scores provided by the face

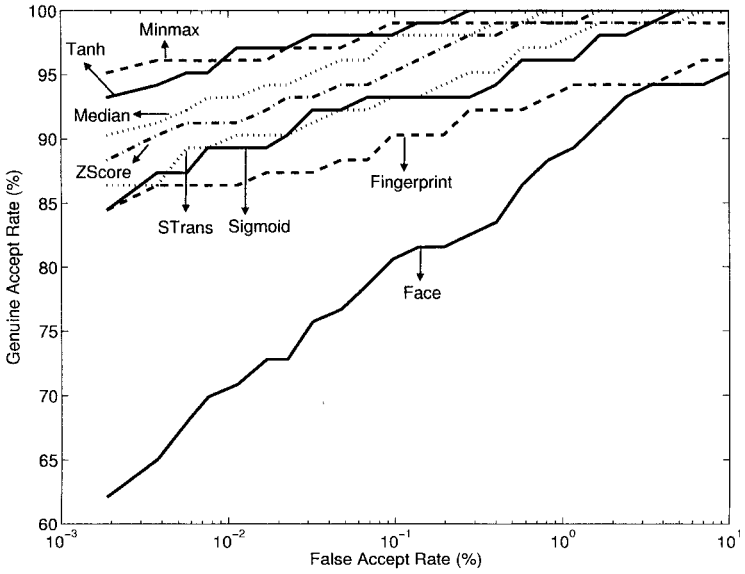


Figure 4.15. ROC curves for sum of scores fusion method under different normalization schemes on NIST BSSR1 dataset.

and fingerprint matchers, respectively. Since there are only 517 genuine match scores for each matcher in the NIST BSSR1 database, the density estimates of the genuine scores are not very reliable.

4.8 User-specific score fusion

It is possible to further enhance the performance of multibiometric systems by adopting user-specific matching thresholds and user-specific weights. Matching thresholds are used by biometric matchers to classify a certain match score as being genuine or impostor. Weights, on the other hand, are used to indicate the importance of individual biometric matchers in a multibiometric framework. In the multibiometric systems described so far, we implicitly assumed that each biometric sub-system provides the same discriminatory information across all users. In practice, the performance of a particular sub-system will vary across users.

Users of a biometric system are prone to different types of errors. The False Reject Rate (FRR) of users with large intra-class variations will be high. Similarly, the False Accept Rate (FAR) amongst users having small inter-class variations will be high. Thus, a "strict" threshold will be appropriate to distinguish users exhibiting a high FAR, while a "loose" threshold may be necessary for

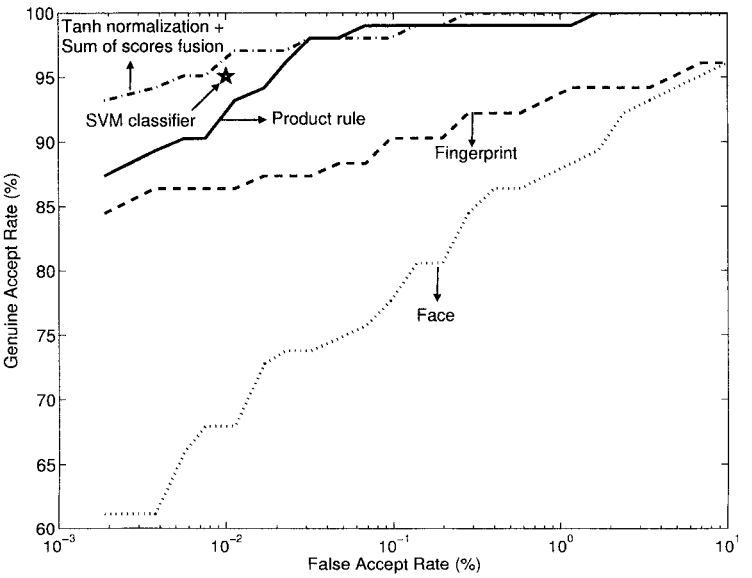


Figure 4.16. Comparison of recognition performance of the density-based, transformation-based and classifier-based score fusion approaches on the NIST BSSR1 database.

users having a high FRR. Furthermore, in a multimodal system, it is instructive to assign different degrees of importance to the various traits on a user-by-user basis. This is especially significant when the biometric traits of a user cannot be reliably acquired. For example, users with persistently dry fingers may not be able to provide good quality fingerprints. Such users can experience higher false rejects when interacting with a fingerprint system. By reducing the weight of the fingerprint trait of such users and increasing the weights associated with the other traits, the FRR of these users can be reduced.

A multibiometric system can be trained to invoke a specific set of threshold and weight parameters based on the claimed identity, I. Automatic learning and update of user-specific thresholds and weights can help reduce the error rates associated with a specific user, thereby improving the overall recognition accuracy of the system (Jain and Ross, 2002b). This will appeal to that segment of the population averse to interacting with a system that constantly requests them to provide multiple readings of the same biometric due to the poor quality of the acquired data.

Toh et al., 2004 identify the following four paradigms in the context of learning user-specific weights and thresholds:

1 Learn globally, decide globally (GG): In this scheme, the multibiometric system learns a common set of weights for the different matchers irrespective of the user and uses a common (global) decision threshold for all users.

2 Learn globally, decide locally (GL): The system learns a common set of weights for the different matchers irrespective of the user but uses user-specific decision thresholds.

3 Learn locally, decide globally (LG): The system learns user-specific weights and a common decision threshold

4 Learn locally, decide locally (LL): The system learns user-specific weights and uses user-specific thresholds.

In these four paradigms, the system either makes use of user-dependent parameters (thresholds and weights) for all users or the parameters are independent of the user. Fierrez-Aguilar et al., 2005a propose a new adaptive learning strategy that offers a trade-off between the user-specific and user-independent approaches.

4.8.1 User-specific matching thresholds

Jain and Ross, 2002b compute the matching thresholds for each user using the cumulative histogram of impostor scores corresponding to that user. Since a sufficient number of user-specific genuine scores would not be available when the user begins to use the system, only the impostor scores are used initially to learn the user-specific thresholds. The impostor scores are generated by comparing the feature sets of a user with feature sets of other users or with feature sets available in a predetermined impostor database. Suppose that the match scores have been quantized into 100 bins. The cumulative histogram at a value x_i, $i = 1, 2, \ldots 100$, is the sum of all those impostor scores less than or equal to x_i. The user-specific matching thresholds are computed as follows.

1 For the i^{th} user in the database, let $t_i(\gamma)$ correspond to the threshold in the cumulative histogram that retains γ fraction of scores, $0 \leq \gamma \leq 1$.

2 Using $\{t_i(\gamma)\}$ as the matching threshold, compute $\{FAR_i(\gamma), GAR_i(\gamma)\}$, where GAR is the Genuine Accept Rate.

3 Compute the total FAR and GAR as

$$\begin{aligned} FAR(\gamma) &= \sum_i FAR_i(\gamma) \\ GAR(\gamma) &= \sum_i GAR_i(\gamma). \end{aligned} \tag{4.68}$$

4 Use $\{FAR(\gamma), GAR(\gamma)\}$ to generate the ROC curve.

Figure 4.17 shows that the choice of the threshold relies on the distribution of impostor scores for each user. This is in contrast to traditional methods where the threshold is established by pooling together the impostor scores associated with all the users. When the multibiometric system is deployed, the γ corresponding to a specified FAR is used to invoke the set of user-specific thresholds, $\{t_i(\gamma)\}$. Table 4.3 shows the user-specific thresholds (corresponding to a FAR of 1%) associated with the 10 users whose data was collected over a period of two months. The ROC curves indicating the improved performance as a result of using user-specific thresholds are shown in Figure 4.18.

User-specific thresholds may also be derived via user-specific score normalization schemes that have been widely used in the speaker recognition community (see Poh and Bengio, 2005b and the references therein). The primary idea here is to shift and scale the genuine and/or impostor score distributions for each user so that their location coincides with a predetermined value. The amount of shift that is necessary determines the user-specific matching threshold.

Table 4.3. User-specific thresholds for the biometric traits of 10 users at a FAR of 1%.

User #	Fingerprint	Face	Hand Geometry
1	14	91	94
2	17	91	95
3	15	92	95
4	12	94	95
5	11	91	90
6	11	90	92
7	16	95	94
8	19	92	97
9	11	90	96
10	19	94	93

4.8.2 User-specific weights

Each biometric matcher provides a match score based on the input feature set and the template against which it is compared. These scores can be weighted according to the biometric trait used, in order to reduce the importance of less reliable biometric traits (and increase the influence of more reliable traits). The weights can be determined in several ways, three of which are described below.

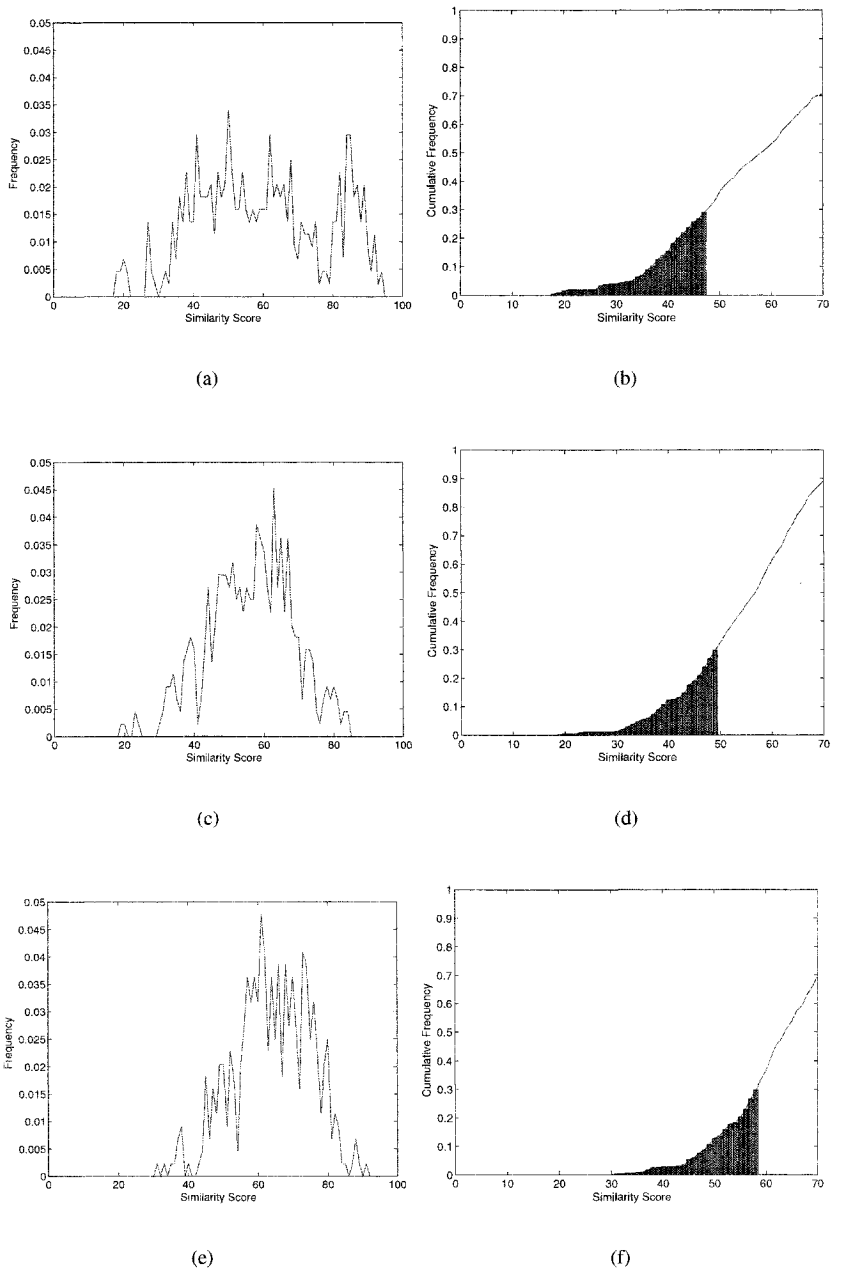
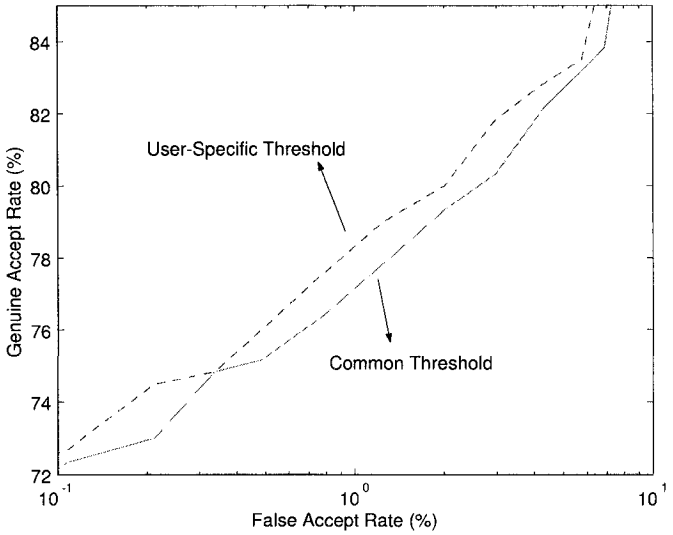
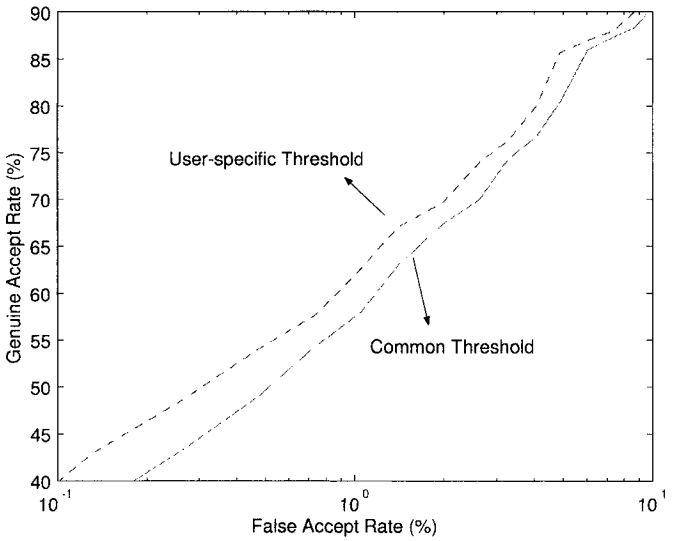


Figure 4.17. The impostor distributions of the face biometric of three different users. (a), (c) and (e) are the histograms of impostor scores associated with the three users. (b), (d) and (f) are the corresponding cumulative histograms. For $\gamma = 0.3$, it is observed that the thresholds for each of the three users are different.



(a)



(b)

Figure 4.18. ROC curves exhibiting performance improvement when user-specific thresholds are utilized to verify a claimed identity. (a) Fingerprint and (b) Face.

1 Equal weights may be assigned to all the modalities, and the fused score obtained as

$$s_{fus} = \frac{1}{n}\sum_{j=1}^{n} s_j, \tag{4.69}$$

where n represents the number of modalities considered. This technique assumes that the performance of the component classifiers are comparable (i.e., balanced classifiers), and that there is no reason to favor one modality over another.

2 Different weights may be assigned to each modality based on their individual performance as summarized by the ROC curve or the EER. Wang et al., 2003 use the following expression to compute the weights in a bimodal system utilizing the face and iris traits.

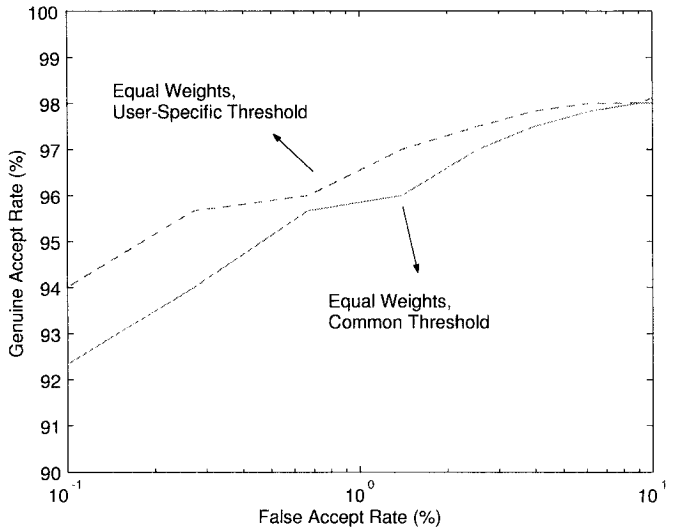
$$w_i = \frac{1 - (FAR_i + FRR_i)}{2 - (FAR_j + FRR_j + FAR_i + FRR_i)}, \tag{4.70}$$

where $i = 1, 2, j = 1, 2$ and $i \neq j$. The values for FAR and FRR in the above equation are threshold dependent. Thus, when the threshold is changed, the weights assigned to the individual modalities will be suitably modified. This technique is useful when the participating classifiers are imbalanced, i.e., when there is significant performance disparity between them. However, it must be noted that the use of order statistics (such as the min score, max score and median score fusion schemes) has been recommended when the classifiers are imbalanced and the optimal set of weights cannot be reliably estimated (Roli and Fumera, 2002; Tumer and Ghosh, 1999).

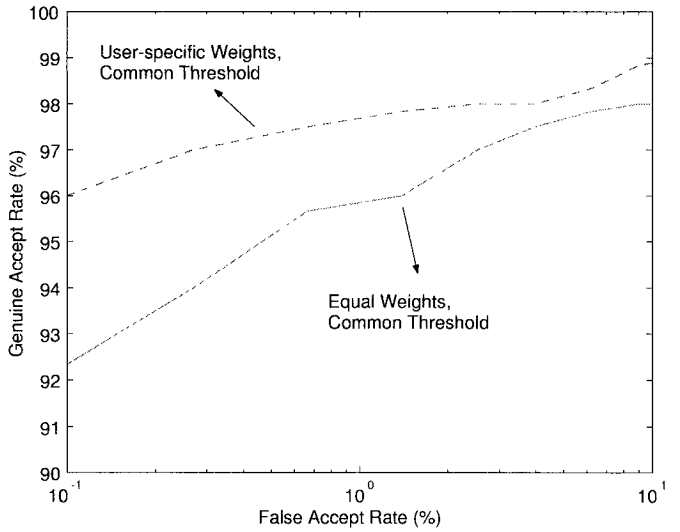
3 The set of weights can also be determined on a user-by-user basis. This process entails searching the space of weights $(w_{k,1}, w_{k,2}, \ldots, w_{k,n})$ for a user, k, such that the total error rate on a training set of fused scores corresponding to that user is minimized. The fused score is computed as

$$s_{fus} = \sum_{j=1}^{n} w_{k,j} s_j. \tag{4.71}$$

Typically, the constraints $\sum_{j=1}^{n} w_{k,j} = 1$ and $w_{k,j} \geq 0$ are applied when searching for the optimal set of weights for user k. The total error rate is the region of overlap of the genuine and impostor score distributions (s_{fus}) corresponding to that user. This approach is beneficial when the performance of individual sub-systems varies significantly across users.



(a)



(b)

Figure 4.19. ROC curves when using (a) equal weights for the three traits and a user-specific matching threshold; and (b) user-specific weights for all the three traits and a common matching threshold (Jain and Ross, 2002b).

Table 4.4. Weights of different biometric modalities for 10 users (Jain and Ross, 2002b).

User #	Fingerprint (w_1)	Face (w_2)	Hand Geometry (w_3)
1	0.5	0.3	0.2
2	0.6	0.2	0.2
3	0.4	0.1	0.5
4	0.2	0.4	0.4
5	0.5	0.2	0.3
6	0.6	0.1	0.3
7	0.6	0.1	0.3
8	0.4	0.2	0.4
9	0.5	0.1	0.4
10	0.6	0.2	0.2

Jain and Ross, 2002b explore the use of a common matching threshold with user-specific weights. Table 4.4 lists the optimal weights (w_1 for fingerprint, w_2 for face and w_3 for hand geometry) computed for the set of 10 users listed in Table 4.3. From this table we observe that for user number 4, the weight assigned to the fingerprint modality is small ($w_1 = 0.2$). Upon examining the fingerprint images corresponding to this user (Figures 4.20(a) and 4.20(b)), it is apparent that the quality of the ridges is rather poor, therefore confounding both the minutiae extractor and matcher. Thus, the match scores in this instance will be unreliable. This demonstrates the importance of assigning user-specific weights to the individual biometric traits. Similarly, user number 3 has a small weight assigned to the face biometric, possibly due to changes in the pose of the face and ambient lighting during data acquisition (Figures 4.20(c), 4.20(d) and 4.20(e)). User number 2 has a small weight attached to hand geometry due to (repeated) incorrect placement of the hand and a slight curvature of the little finger (Figures 4.20(f) and 4.20(g)). The improvement in matching performance of the user-specific system is indicated by the ROC curves in Figure 4.19(b).

The utilization of user-specific matching thresholds and weights presupposes the availability of a large number of genuine and impostor scores pertaining to an individual. Since the number of biometric samples obtained from an individual during enrollment is very limited, user-specific schemes cannot be invoked at the time of system deployment. As more and more biometric samples of a user are made available over a period of time, user-specific parameters can be utilized to enhance recognition accuracy.

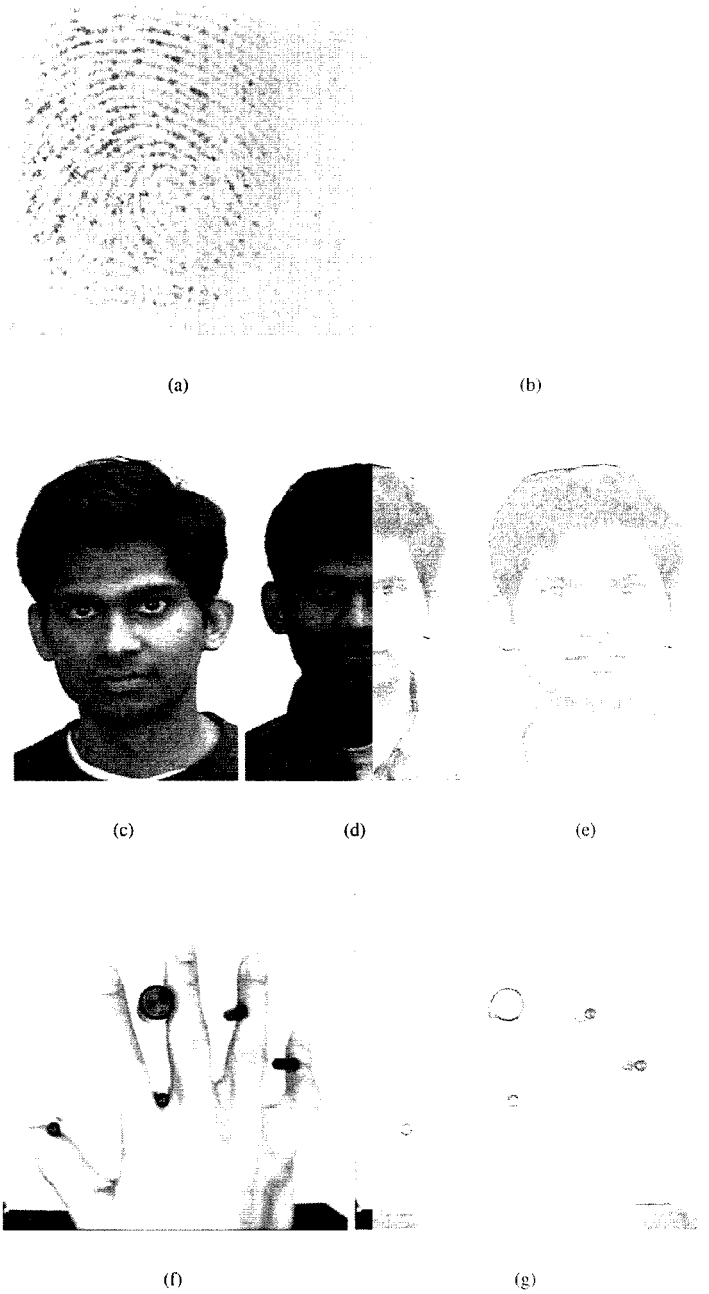


Figure 4.20. Examples of users with varying weights for the different modalities. (a) and (b) Fingerprint images of user number 4 whose ridge details are not very clear ($w_1 = 0.2$). (c), (d) and (e) Varying face poses of user number 3 ($w_2 = 0.1$). (f) and (g) Incorrect placement of hand and the curved finger of user number 2 ($w_3 = 0.2$).

4.9 Summary

In a multibiometric system, fusion at the score level offers the best tradeoff between information content and ease of fusion. Hence, score level fusion is typically adopted by most multibiometric systems. Although a wide variety of score level fusion techniques have been proposed in the literature, these can be grouped into three main categories, viz., density-based, transformation-based and classifier-based schemes. The performance of each scheme depends on the amount and quality of the available training data. If a large number of match scores is available for training the fusion module, then density-based approaches such as the likelihood ratio test can be used. Estimating the genuine and impostor distributions may not always be feasible due to the limited number of training samples that are available. In such cases, transformation-based schemes are a viable alternative. The non-homogeneity of the match scores presented by the different matchers raises a number of challenges. Suitable score normalization schemes are essential in order to transform these match scores into a comparable domain. The sum of scores fusion method with simple score normalization (such as z-score) represents a commonly used transformation-based scheme. Classification-based fusion schemes consolidate the outputs of different matchers into a single vector of scores which is then fed into a trained classifier. The classifier determines if this vector belongs to the "genuine" or "impostor" class.

User-specific fusion schemes can be invoked if sufficient training data is accumulated over a period of time for individual users. However, the following issues will have to be considered when deploying biometric systems with user-specific fusion schemes: (i) A malicious user may deliberately provide poor quality biometric data constantly (e.g., by touching the fingerprint sensor lightly), thereby forcing the system to reduce the weights associated with a specific biometric. The user may then claim that the biometric data belongs to someone else. Thus, the user can access a privilege and deny using it later. (ii) An intruder attempting to circumvent a biometric system might target enrolled users with known problems with their biometric data (e.g., users with calloused fingers or arthritis of the hand). Such users may have low weights associated with certain biometric traits and, therefore, the intruder will need to spoof only those traits with higher weights. Therefore, appropriate safeguards should be incorporated into multibiometric systems that employ user-specific fusion schemes.

As stated in the previous chapter, the gain in the matching performance of a multibiometric system is affected by the correlation between the match scores emitted by the different biometric matchers (Kuncheva et al., 2000; Prabhakar and Jain, 2002; Poh and Bengio, 2005d). In general, if the match scores are uncorrelated or negatively correlated, then the improvement in performance can be expected to be significant. Thus, combining two weak biometric matchers

that are uncorrelated may result in a significant improvement in performance than combining two strong biometric matchers that are positively correlated (this means, when more than two biometric matchers are available, combining the two best matchers will not always result in the best matcher pair). In view of this, it is imperative that system integrators do not discard biometric matchers whose individual performances are poor without attempting to fuse them with other matchers.