2007

A statistical approach towards performance analysis of multimodal biometrics systems

Wei Gan
University of Windsor

A Statistical Approach towards Performance Analysis of Multimodal Biometrics Systems

by

Wei Gan

A Thesis
Submitted to the Faculty of Graduate Studies through Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

2007

Your file Votre référence
ISBN: 978-0-494-34942-7
Our file Notre référence
ISBN: 978-0-494-34942-7

ABSTRACT

Fueled by recent government mandates to deliver public functions by the use of biometrics, multimodal biometrics authentication has made rapid progress over the past a few years. Performance of multimodal biometrics systems plays a crucial role in government applications, including public security and forensic analysis. However, current performance analysis is conducted without considering the influence of noises, which may result in unreliable analytical results when noise levels change in practice.

This thesis investigates the application of statistical methods in performance analysis of multimodal biometric systems. It develops an efficient and systematic approach to evaluate system performance in different situations of noise influences. Using this approach, 126 experiments are conducted with the BSSR1 dataset. The proposed approach helps to examine the performance of typical fusion methods that use different normalization and data partitioning techniques.

Experiment results demonstrate that the Simple Sum fusion method working with the Min-Max normalization and Re-Substitution data partitioning yields the best overall performance in different noise conditions. In addition, further examination of the results reveals the need of systematic analysis of system performance as the performance of some fusion methods exhibits big variations when the level of noises changes and some fusion methods may produce very good performance in some application though normally unacceptable in others.

DEDICATION

To

two beloved women in my life:

my mother and my wife,

for their understanding and encouragement

of a lifetime

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mother and my wife. Without their support and encouragement at crucial periods of my life, it would not have been possible for me to pursue graduate studies and aim for greater things in my life.

I am very grateful to my advisor, Dr. Xiaobu Yuan, for his constant motivation, support and infectious enthusiasm in guiding me towards the successful completion of this thesis research. My interactions with him have been of immense help in defining my research goals and in identifying ways to achieve them.

I would like to acknowledge Dr. Bojan Cukic and Mrs. Nevena Samoska from West Virginia University for the invaluable instructions about MUBI analysis tool. I would also like to thank my thesis committee members, Dr. Chunhong Chen, Dr. Ziad Kobti and Dr. Christie Ezeife, who have been all generous and patient. Their confidence in my abilities has been unwavering, and has helped to make this thesis a solid work. Special thanks to Alison Samson and Pat Cousins for the endeavours they have kindly provided in relieving the financial burdens during my studies.

Finally I want to extend my gratitude to my parents-in-law for all their prayers and blessings and my friends for great company and inspirational conversations during my stay at University of Windsor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 MOTIVATION

Biometric authentication is a young yet fast evolving science that establishes an identity based on the physical or behavioral attributes of an individual. It has been seen the emerging technologies replacing their alphanumeric counterparts with traits that cannot be forgotten, easily stolen, or given to another person. Furthermore, multimodal biometric systems have been established (Jain, 2004a) to outperform the unimodal biometric systems.

Performance of multimodal biometrics systems plays the crucial role especially when these systems are employed in government, financial or forensic applications. Many researches on performance analysis of multimodal biometrics systems have been conducted. The first attempt of performance evaluation took place in 2000 (Blackburn, 2000) and there have been reports of several other testing on the performance of different biometric systems in specific applications afterwards (Maio, 2004), (Wilson, 2004a). Those studies are all based on the testing protocol that lacks the thorough study of the system performance in the variety of noise presences. The influence of noise on the system performance, however, may result in different analysis results when the noise varies.

1.2 CONTRIBUTIONS

This thesis is intended to study the performance of multimodal biometrics systems under the influence of various noises in a systematical manner and identify the most optimum design of a multimodal system under noise disturbances. In addition, the cost of evaluation should be reduced to the minimum despite the exponential growth of possible noise conditions and system parameters.

To achieve the above goals, following the principle of Design of Experiments (DoE), this thesis proposes a statistical approach to model noise influences on system performance, to evaluate performance under the noise disturbances efficiently and systematically, and to identify optimum configurations.

1.3 ORGANIZATION

The rest of this thesis is organized as follows. Chapter 2 introduces the basic concepts of biometrics and multimodal biometrics, as well as different normalization methods and fusion techniques. Chapter 3 examines performance analysis approaches and metrics for multimodal biometrics system. The relationship between noise factors and performance is also investigated. Chapter 4 presents the problem domain and discusses robust parameter design for performance and proposed statistical approach in detail. Chapter 5 explains the tools and databases to be used in implementation, and expected results. Finally, the conclusions are presented in Chapter 6.

CHAPTER 2 MULTIMODAL BIOMETRICS

2.1 OVERVIEW

Biometric authentication, or simply biometrics, is a young yet fast evolving science that establishes an identity based on the physical or behavioral attributes of an individual, including fingerprint, face, voice, gait, iris, signature, hand geometry and ear (Ross, 2006).

Biometrics have been seen the emerging technologies replacing their alphanumeric counterparts with traits that cannot be forgotten, easily stolen, or given to another person.

2.1.1 Biometrics traits

According to Jain et al. (Jain, 2004b), a potential biometric trait should meet the following listed requirements:

1. Universality - the trait should be possessed by each individual in the given population.
2. Distinctiveness - the trait should be unique for each person within that population.
3. Permanence - the trait should not change over a period of time with respect to the matching algorithm.
4. Collectability - the trait should be easy to collect automatically in modern biometric systems and must be measurable quantitatively.
5. Performance - the trait should lend itself to fast and accurate identification.
6. Acceptability - people should be able to accept the use of a certain biometric trait.
7. Circumvention - reflects how easily the biometric trait can be spoofed using fraudulent methods.

There are two major groups of biometric traits (see figure 2.1), physical or behavioral traits. The first group includes fingerprint, hand geometry, iris, retina, face, palmprint, ear

structure and DNA etc. The second group consists of voice, gait, signature dynamics and keystrokes dynamics.

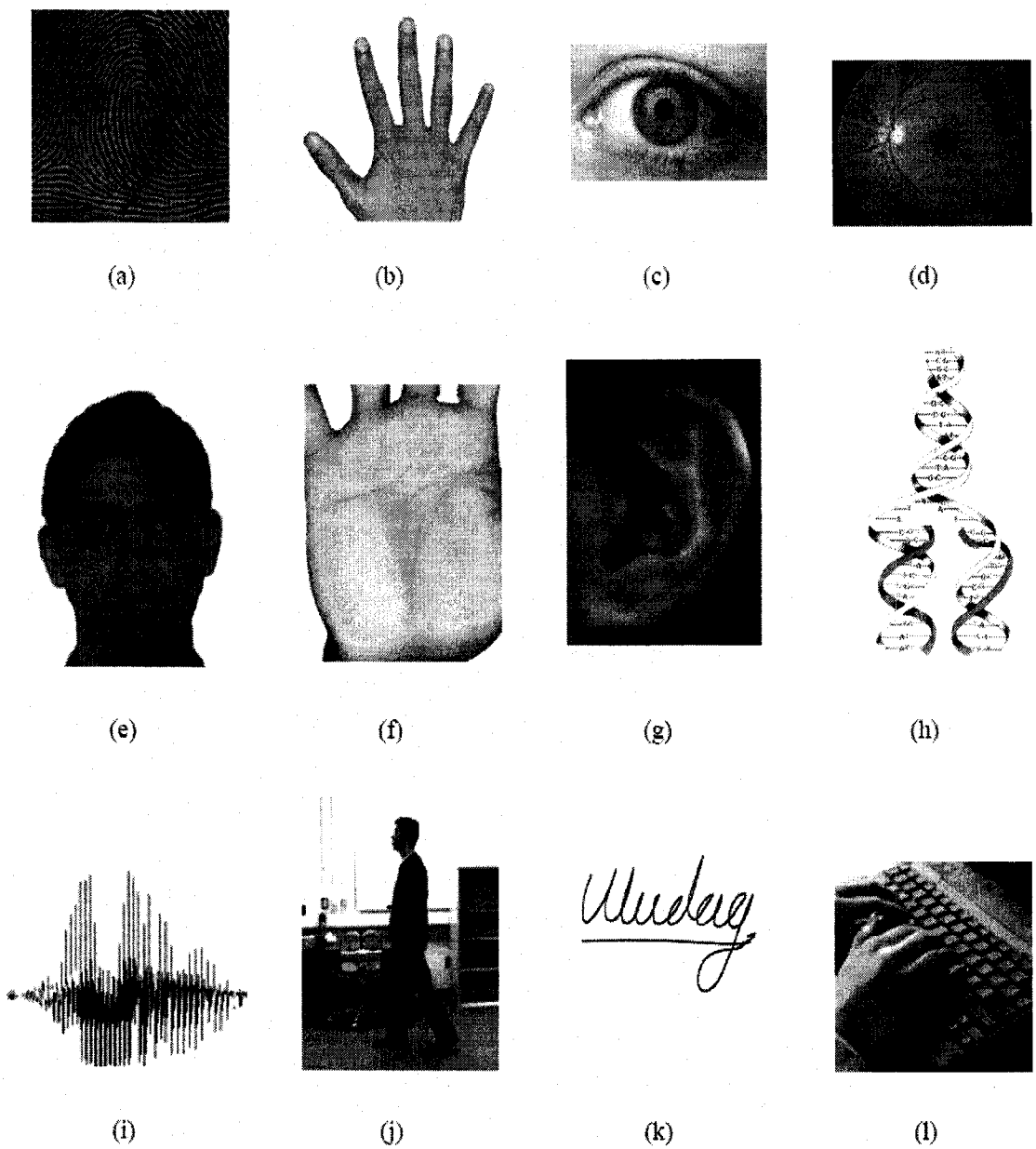


Figure 2. 1 Physical traits include (a) Fingerprint; (b) Hand-geometry; (c) Iris; (d) Retina; (e) Face; (f) Palmprint; (g) Ear structure; (h) DNA; Behavioral traits consist of (i) Voice; (j) Gait; (k) Signature and (l) Keystroke dynamics. (Nandakumar, 2005)

2.1.2 Biometrics applications

Biometrics offers a natural and reliable solution to the problem of identity determination by recognizing individuals based on their physiological and/or behavioral characteristics that are inherent to each person. Biometrics technique has made rapid progress over the past few years. Especially due to recent government mandates stipulating the use of biometrics for delivering crucial public functions.

The US-VISIT program (United States Visitor and Immigration Status Indicator Technology), for example, is a border security system that validates the travel documents of foreign visitors to the United States. Currently, fingerprint images of left- and right-index fingers of a person are being used to associate a visa with an individual entering the United States; in the future, all ten fingers may be used thereby necessitating the development of efficient data capture as well as fusion algorithms.

Figure 2.2 The US-VISIT immigration system (Wikipedia, 2007)

Generally biometrics is being increasingly incorporated in several different applications. These applications can be categorized into three main groups (Ross, 2006):

1. Commercial applications

Computer network login, electronic data security, e-commerce, Internet access, ATM or credit card use, physical access control, mobile phone, PDA, medical records management, distance learning.

2. Government applications
National ID card, managing inmates in a correctional facility, driver's license, social security, welfare-disbursement, border control, and passport control.

3. Forensic applications
Corpse identification, criminal investigation, parenthood determination, missing child.

2.1.3 Biometric system modules

Ross et al. (Ross, 2003) describes a simple biometric system with four major modules: (1) Sensor module which acquires the trait in the form of raw biometric data. An example is a fingerprint sensor that captures fingerprint impressions of a user.

(2) Feature extraction module which processes data to extract a feature set with compact representation of the trait. For example, the position and orientation of minutiae points in a fingerprint image would be extracted in the feature extraction module of a fingerprint system.

(3) Matching module which employs a classifier to compare the extracted feature set against those in the template by generating a matching score. For example, in this module, the number of matching minutiae points between the query and the template will be computed and treated as a matching score.

(4) Decision-making module in which the user's identity is established or a claimed identity is either accepted or rejected based on the matching score generated in the matching module.

2.2 MULTIMODAL BIOMETRICS

2.2.1 Why multimodal biometrics

Most biometric systems relying on the evidence of a single source of information for authentication (e.g., single fingerprint or face) have to contend with a variety of problems (Ross, 2004):

(a) Noise in sensed data: Noises presented in the acquired biometric data are mainly contributed by defective or improperly maintained sensors (e.g., accumulation of dirt on a fingerprint sensor) or unfavorable ambient conditions (e.g., poor illumination of a user's face in a face. A fingerprint image with a scar, or a voice sample altered by cold are also examples of noisy data.

(b) Intra-class variations: The biometric data acquired from a user during verification will not be identical to the data used for generating the user's template during enrollment.These variations may be due to improper interaction of the user with the sensor (e.g., incorrect facial pose), or use of different sensors during enrollment and verification, (e.g., optical versus solid-state fingerprint sensors), changes in the ambient environmental conditions (e.g., illumination changes in a face recognition system) and inherent changes in the biometric trait (e.g., appearance of wrinkles due to aging or presence of facial hair in face images, presence of scars in a fingerprint, etc.).

(c) Inter-class similarities: In a biometric system comprising of a large number of users, there may be inter-class similarities (overlap) in the feature space of multiple users. For example, currently used appearance-based facial features have limited distinguishing abilities. Because of the genetic factors, a number of specific groups (e.g., father and son, identical twins, etc.) are hard to be identified.

(d) Non-universality: The biometric system may not be able to acquire meaningful biometric data from a subset of users. In other words, not all biometric traits are strictly universal. For example, a report (NIST, 2000) by the National Institute of Standards and

Technology (NIST) to the United States Congress concluded that approximately two percent of the population does not have a legible fingerprint and therefore cannot be enrolled into a fingerprint biometrics system.

(e) Spoof attacks: This type of attack is especially relevant when behavioral traits such as signature or voice are used. However, physical traits such as fingerprints are also susceptible to spoof attacks.

Due to these limitations imposed by unimodal biometric systems, error rates are fairly high, which makes them unacceptable for deployment in security critical applications. It is estimated that if NY airports, which boast an average of more than 300,000 passengers pass through daily, deploy unimodal biometric systems like fingerprint, face or voice for identification respectively, there would be 600 falsely rejected (and inconvenienced) passengers per day for fingerprints, 30,000 for face and 45,000 for voice. Similar numbers can be computed for false accepts.

2.2.2 Multimodal biometrics

Multimodal biometrics is the usage of more than one physiological or behavioural characteristic to identify an individual. It involves the fusion of two or more technologies such as fingerprint, facial recognition, iris scanning, hand geometry, signature verification, or speech recognition.

It is must be noted that multimodal biometrics does not only refer to multiple biometric traits scenario, (Nandakumar, 2005) illustrates fusion in multi-modal biometrics systems can be implemented in the following five scenarios, among which the first four scenarios are based on the same biometric trait (see figure 2.3):

1) Multiple Instances of the same biometric may be combined (e.g., multiple face images of a person obtained under different pose/lighting conditions).

2) Multiple Sensors may be used to capture the same biometric (e.g., optical and solid state fingerprint sensors).

3) Multiple Representations and matching and/or feature extraction algorithm may be used on the same biometric reading to give separate results (e.g., multiple face matchers like PCA and LDA).

4) Multiple Units of the same biometric may be taken (*e.g.*, two different fingerprints or both irises);

5) Multiple Biometric Traits may be captured.

Some of inherent limitations of the unimodal biometric can be alleviated by fusing the information presented by multiple sources. A multimodal system demonstrates increased improvements in anti-spoofing, and the ability to deal with large user population and acceptable error rates. The difficulty to forge multiple biometric traits within a certain time frame makes spoofing attacks hard to be conducted. In addition, those people who are missing some traits like a mute person or a person without several fingers can be identified by using multimodal biometrics.

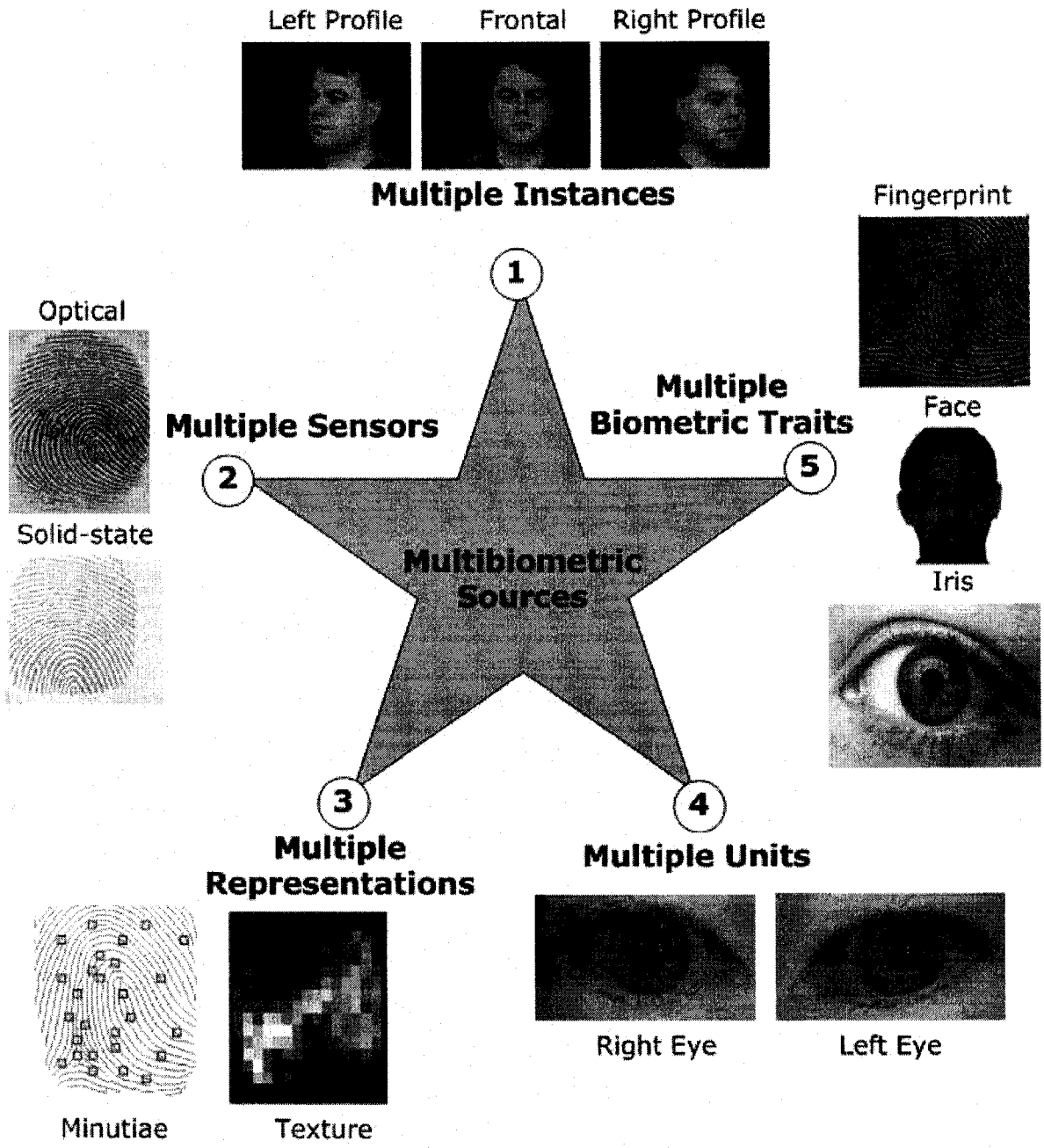


Figure 2.3 Sources of multiple evidence in multimodal biometric systems (Nandakumar, 2005)

Another significant advantage the multimodal biometrics brings over uniodal biometrics is the obvious increase of system performance. These systems with multiple and independent sources of evidence can offer more reliable and higher verification rates, and improve the accuracy greatly.

Michigan State University (Ross, 2003) has conducted a study on evaluating the ROC curves of fingerprint, facial and hand geometry systems (see Figure 2.4). The performance of any individual modality is far below the performance of the combination of all the three modalities.

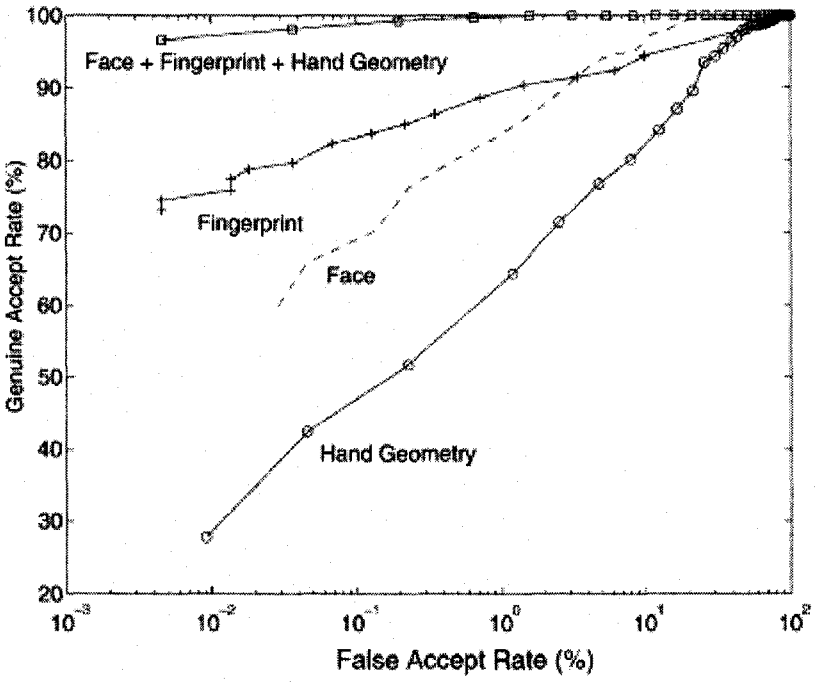


Figure 2.4 ROC Curve for a system utilizing multiple biometric traits (Ross, 2003)

2.2.3 Fusion in biometrics

Information fusion is the essential element in multimodal biometrics. Information fusion in multimodal biometrics is the integration of data pertaining to multiple independent biometric devices. Fusion in multimodal biometric systems can take place at three major levels, namely, feature level, score level and decision level. Figure 2.5 displays the fusion of a biometric system at various levels.

Feature Extraction level: Combining different feature vectors that are obtained from one of the following sources: multiple sensors for the same biometric trait, multiple instances of the same biometric trait, multiple units of the same biometric trait or multiple

biometric traits. Combining more feature vectors results in one vector with higher dimensionality and may increase the probability of correctly identifying a person. However, integration at the feature level is difficult to achieve in practice because of the 'curse of dimensionality' problem (Duda, 2001), unknown relationship between the feature spaces of different biometric systems, and inaccessible feature vectors for most commercial biometric systems

Although information fusion at an early stage results in more effective performance than performing fusion at later time, the above obstacles prevent most of the researchers from studying integration at the feature level.

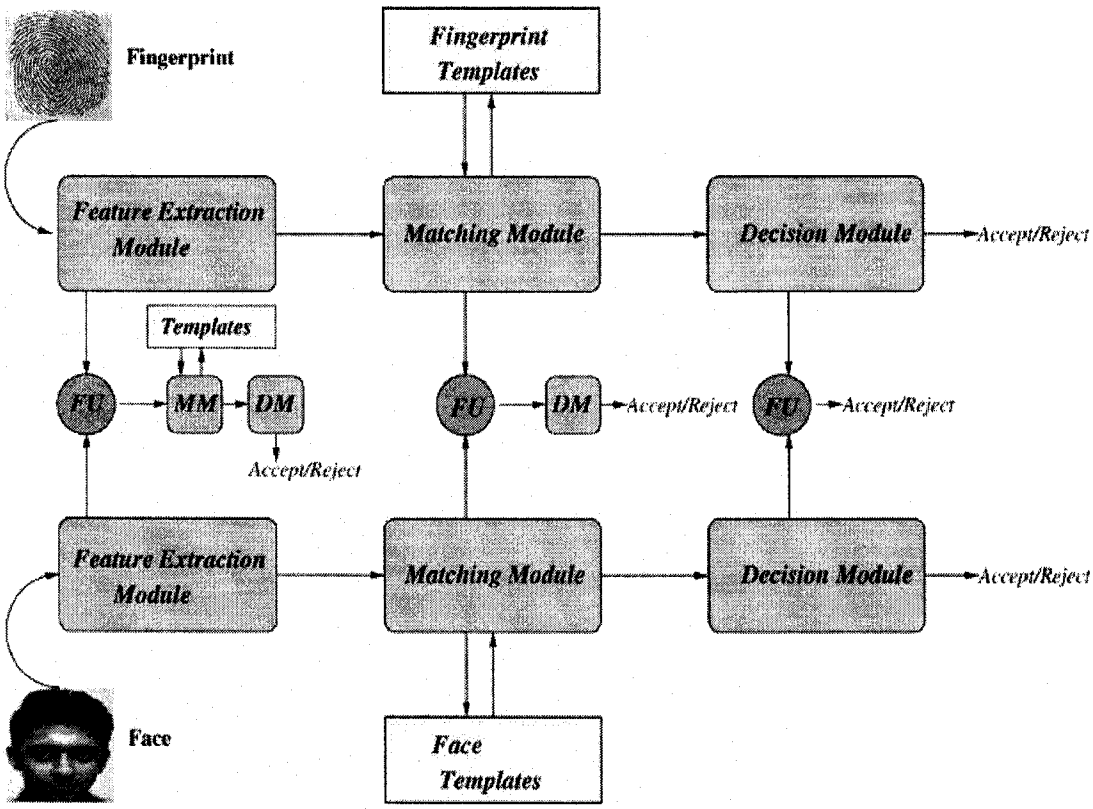


Figure 2. 5 A bimodal biometric system showing the three levels of fusion (FU: fusion module, MM: matching module, DM: decision module). (Ross, 2003)

Decision level (or abstract level)**:** Integration of information at the decision level can help to reach the final decision when single biometric matcher individually decides on the best match based on the input presented to it. Methods at this level including majority voting (Lam, 1997), behavior knowledge space (Lam, 1995), weighted voting etc.
Fusion at decision level is the least informative and least effective since it happens at the last stage of the system processing. It does not work well enough, and often gives a combined decision worse than the decision from the best individual biometric device.

Matching score level: This level is also known as confidence level or measurement level. Fusion at this level is much more effective than fusion at the decision level. Matching score is a measure of the similarity between features derived from a presented sample and a stored template. Each unimodal biometric system measures and calculates its own matching score and these matching scores are fused to reach a final match/ non match decision based on a certain decision threshold.

There are two approaches for consolidating the scores obtained from different matchers. One approach is to formulate it as a classification problem where for each biometric modality a feature vector is constructed using the matching scores. This feature vector is then classified into one of two classes: "Accept" (genuine user) or "Reject" (impostor user). In general the classifier used in this scenario has the ability to learn the decision boundary irrespective of the generation of feature vector. The output scores of the different modalities can be non-homogeneous (distance or similarity metric, different numerical ranges, etc). They are not required to be processed before being fed into the classifier.

The second approach combines the individual matching scores to generate a single scalar score, which is then used to make the final decision. Since the matching scores are heterogeneous, to ensure a meaningful combination of the scores from the different modalities, normalization is required to transform these scores into a common domain.

(Snelick, 2005) analyzed the advantages of fusion at matching score stage in several aspects. Firstly matching score fusion does not affect the existing proprietary biometric systems, allowing for a common middleware layer to handle the multimodal application but with a small amount of common information. These existing and proprietary unimodal biometric systems can be easily combined into a multimodal biometrics system given some basic information provided. Secondly the data from prior evaluations of single-modal biometric systems can be reused. This avoids live testing or re-running individual biometric algorithms.

Another advantage is that the matching scores output by the matchers contain the second richest information about the input pattern next to the feature vectors; however it is much easier to access and to combine the scores generated by the different matchers compared to fusion at the feature extraction level.

Consequently, integration of information at the matching score level is the most common approach in multimodal biometric systems nowadays.

2.3 COMBINATION APPROACH TO SCORE LEVEL FUSION

When comparing the two approaches for score level fusion, experiments indicate that the combination approach performs better than the classification approach (Ross, 2003); we will therefore discuss more about combination approach to score level fusion.

Prior to combining scores of different matchers into a single score, several issues need to be considered. First of all, the match scores generated by the individual matchers may not be compatible. For example, one matcher may output a distance (dissimilarity) measure while another may output a similarity measure. Furthermore, the outputs of the individual matchers may have different numerical scales (range).For example, one matcher may output the interval within (0, 1) while another output the interval within (0,100). Finally, the match scores may follow different probability distributions. Normalization technique is then used to address the problems.

2.3.1 Normalization methods

To address the problem of incomparable classifier output scores in different combination classification systems, normalization methods are used to change the location and scale parameters of the matching score distributions at the outputs of the individual matchers. In such a way, various matching scores of different matchers are converted into a common domain and can be combined later on (Jain, 2005).

It is highly desirable that the normalization of the location and scale parameters of the matching score distribution must be *robust* and *efficient*. Huber (Huber, 1981) defines *robustness* as insensitivity to the presence of outliers and *efficiency as* the proximity of the obtained estimate to the optimal estimate when the distribution of the data is known. Huber also argues even though many techniques can be used for score normalization, the challenging work is to identify a technique that can be both robust and efficient.

The following is a list of normalization methods that are commonly used and their robustness and efficiency have been examined. We denote a raw matching score set $\{S_k\}$ of all scores for a matcher, and the corresponding normalized score set as $\{S'_k\}$.

1) Min-max normalization

Min-max is the simplest normalization technique. It 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 shift the minimum and maximum scores to 0 and 1, respectively. Min-max normalization keeps the original distribution of scores except for a scaling factor and transforms all the scores into a common range [0, 1].

The normalized scores are given by

$$s'_k = \frac{s_k - min}{max - min}.$$

We can estimate the minimum and maximum values for a set of matching scores from the training set even if the matching scores are not bounded. But the method is not robust in that case as it is highly sensitive to outliers in the training set used for estimation.

2) Decimal scaling normalization

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;1]and the other has scores in the range [0;100], the following normalization could be applied.

$$s'_k = \frac{s_k}{10^n},$$

where $n = \log_{10} \max(s_i)$.

The problems with this approach are lack of robustness and the assumption that the scores of different matchers vary by a logarithmic factor (Jain, 2005). If the matching scores of the modalities are not distributed on a logarithmic scale, then this normalization technique cannot be applied.

3) Z-score normalization

Z-score is the most commonly used score normalization technique. The normalized score is calculated using the arithmetic mean and standard deviation of the given data. If we have known the nature of the matching algorithm, it will work well by using this scheme, otherwise we have to estimate the average score and score variations of the matcher from a given set of matching scores.

The normalized scores are given by

$$s'_k = \frac{s_k - \mu}{\sigma},$$

where μ is the arithmetic mean and σ is the standard deviation of the given data.

We can see both mean and standard deviation are sensitive to outliers and Z-score method is therefore not a robust one. Furthermore, a common numerical range of the normalized scores from the different matchers is not promised by using Z-score method. And due to the fact that mean and standard deviation are only the optimal location and scale parameters for Gaussian distribution, the output of Z-score normalization for a non-Gaussian distribution input fails to keep the original distribution.

4) Median and median absolute deviation (MAD) normalization

The *median and median absolute deviation (MAD)* is insensitive to outliers and the points in the extreme trails of the distribution. Hence, median and MAD method is robust and is given by

$$s'_k = \frac{s_k - median}{MAD},$$

where MAD =median($|S_k$ - median|).

However, the median and the MAD estimators have a low efficiency compared to Z-score method (Jain, 2005).

5) Tanh-estimators normalization

The tanh-estimators introduced by Hampel et al. (Hampel, 1986) are robust and highly efficient.

The normalization is given by

$$s'_k = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{s_k - \mu_{GH}}{\sigma_{GH}}\right)\right) + 1\right\},$$

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 used to reduce the influence of outliers in the distribution based on the influence (ψ) - function below

$$\psi(u) = \begin{cases} u & 0 \le |u| < a, \\ a * sign(u) & a \le |u| < b, \\ a * sign(u) * \left(\frac{c-|u|}{c-b}\right) & b \le |u| < c, \\ 0 & |u| \ge c. \end{cases}$$

The Hampel influence function can reduce the influence of the points at the tails of the distribution (identified by a, b, and c) during the estimation of the location and scale parameters. This method is therefore insensitive to outliers. However, tradeoff between the robustness and efficiency of this method should be decided cautiously. If too many points from the tail of the distributions are removed, estimation becomes robust but not efficient. Otherwise efficiency increases and robustness goes down when points from the tail are kept as many as possible. Practically parameters (a, b, and c) are chosen depending on the amount of noise in the training data set because it decides the extent of robustness the system requires.

2.3.2 Fusion methods

In the famous theoretical framework (Kittler, 1998) for consolidating the evidence obtained from multiple classifiers, Kittler et al. offer a number of fusion schemes including Min rule, Max rule, Sum rule and Product rule. These techniques can be applied to the system only if the output of each modality is in the form of P(genuine|X),where X is the input pattern. That is, what to be fused in the system is the posteriori probability of user being "genuine" given the input biometric sample X. However, practically most biometric systems output a matching score ***s***.

One solution is approximating P(genuine|X) by P(genuine|s) which can be calculated from the matching scores. But Jain et al. (Jain, 2004b) argue that without corresponding

confidence measure, the calculated value of P(genuine|s) is not a good estimate of P(genuine|X) and this can result in poor recognition performance. Hence, when consolidating the matching scores of individual modalities which don't offer confidence measure, it would be better to combine the matching scores directly using an appropriate method without converting them into probabilities.

The following is the fusion techniques that use the multiple normalized scores directly and combine them into a single score.

If s_i is the matching score from i^{th} modality, s represents the resulting fused score.

1) **The *Simple Product Rule*** combines the scores by multiplying all of the individual scores.

$$s = s_1 * s_2 * \ldots * s_n$$

2) **The *Simple Sum Rule*** combines the scores as a linear transformation.

$$s = (a_1 s_1 - b_1) + \ldots + (a_n s_n - b_n)$$

a_i and b_i represents the weights and biases, respectively, which can be specified by the user.

3) **The *Simple Max Rule*** is the maximum score from the different modalities.

$S = \text{Max}(s_1, s_2, \ldots\ldots s_n)$

4) **The *Simple Min Rule*** is the minimum score from the different modalities.

$S = \text{Min}(s_1, s_2, \ldots\ldots s_n)$

In addition to the above techniques, BGI/ LRGI is another fusion method that have been used in many existing biometric system.

5) Biometric Gain against Impostor (BGI) / Likelihood Ration of Genuine to Impostor (LRGI)

The BGI is a very useful concept. It is a measurement about how many times more likely we believe it that the claimant is an impostor, after having made biometric measurements, than we believed it beforehand. Its mathematical definition is the ratio of the a posteriori to the a priori probabilities of the claimant being an impostor. (Sedgwick, 2004)

$$BGI = \frac{\text{Probability of being an impostor, given the biometric evidence too}}{\text{Probability of being an impostor, given only prior knowledge}}$$

The modified BGI as the Likelihood Ratio of Genuine to Impostor (LRGI) is a very good approximation to the BGI during most of the time.

$$BGI \approx LRGI = \frac{\text{Probability of seeing the evidence from an impostor}}{\text{Probability of seeing it from the expected genuine subject}}$$

Every score that comes out of the biometric devices is transformed to the LRGI scale. This is a score normalisation process. Then the various scores are combined by multiplication or by addition of the log likelihood ratios. This characteristic of BGI/LRGI fusion method exempts itself from score normalization in the sense it can normalize and fuse the matching scores together and no normalization is needed when using this fusion method.

Due to the fact some biometric traits can not be reliably obtained in some cases (e.g. good quality faces can not be obtained from users with dry faces), Jain and Ross (Jain, 2002) have proposed the use of user specific weights for computing the weighted sum of scores from the different modalities. For the example of dry face users, a lower weight can be assigned to the face score while raising the weight to the scores of the other modalities

The same scheme can be applied to threshold. (Jain, 2002) has shown that the use of user-specific weights and thresholds can improve the performance by approximately 3% and 2%, respectively. However, this method requires learning of user-specific weights from the training scores available for each user.

CHAPTER 3 PERFORMANCE ANALYSIS ON MULTIMODAL BIOMETRICS SYSTEM

The performance of the biometric system has received increasing concerns especially when biometric systems are employed in public security, financial or forensic applications. Bad performance of a biometric system may contribute to very serious problems. However, performance analysis of a complete biometric system is a comprehensive and challenging task which involves the concerns about matching or technical performance, engineering performance, security performance and user's habituation and privacy etc (Ross, 2006). In this thesis, we mainly focus on the matching performance of a biometric system.

3.1 MATCHING PERFORMANCE METRICS

There are a number of matching performance metrics evaluating a multimodal biometrics system to a given application, including accuracy, cost and speed of the system. Hong et al. (Hong, 1999) believe that as the higher speed processors are becoming available at cheaper prices and as the cost of the biometric sensors is dramatically reduced, the accuracy performance of biometrics systems plays a much more significant role than others in its performance assessment.

3.1.1 Identification error rates

Biometric systems are designed to make binary decisions accepting the authorized enrollee and rejecting the impostors. There are two types of identification errors the system probably makes: it may either falsely accept an impostor (FA) or falsely rejects an enrollee (FR). These are also called False Match (FM) and False Non-Match (FNM) respectively.

False Acceptance is caused by the incorrect judge that an impostor has matched to an enrollee's template stored in the system's database. And false rejection is caused by the incorrect judge that an enrollee does not match his or her own enrollment template. FA is considered the most serious of biometric security error, with an unauthorized person being admitted. A FR results in convenience problems, since genuinely enrolled identities are denied access to the application, or at least will have to go through some further check to be admitted.

False Acceptance Rate (FAR) is defined as the ratio of the number of false acceptances divided by the number of identification attempts. **False Rejection Rate (FRR)** is accordingly defined as the ratio of the number of false rejection divided by the number of identification attempts. In many cases, **Genuine Acceptance Rate (GAR)** which is the complement of FRR is used to replace FRR (GAR=1-FRR).

Genuine and impostor scores are used to help calculating FAR and FRR.Genuine scores are matching sccores that result from comparing elements in the target and query sets of the same subject. Impostor scores are matching scotes resulting from comparisons of different subjects.When a large number of genuine and impostor scores is available and a matching score threshold is chosen, FAR and FRR then can be derived based on the threshold. Figure 3.1 shows two curves representing the genuine and impostor probability density functions respectively, a matching score threshold t is chosen.
Then the FRR is the area under the genuine density function to the left of the threshold and the FAR is the area under the impostor density function to the right of the threshold.

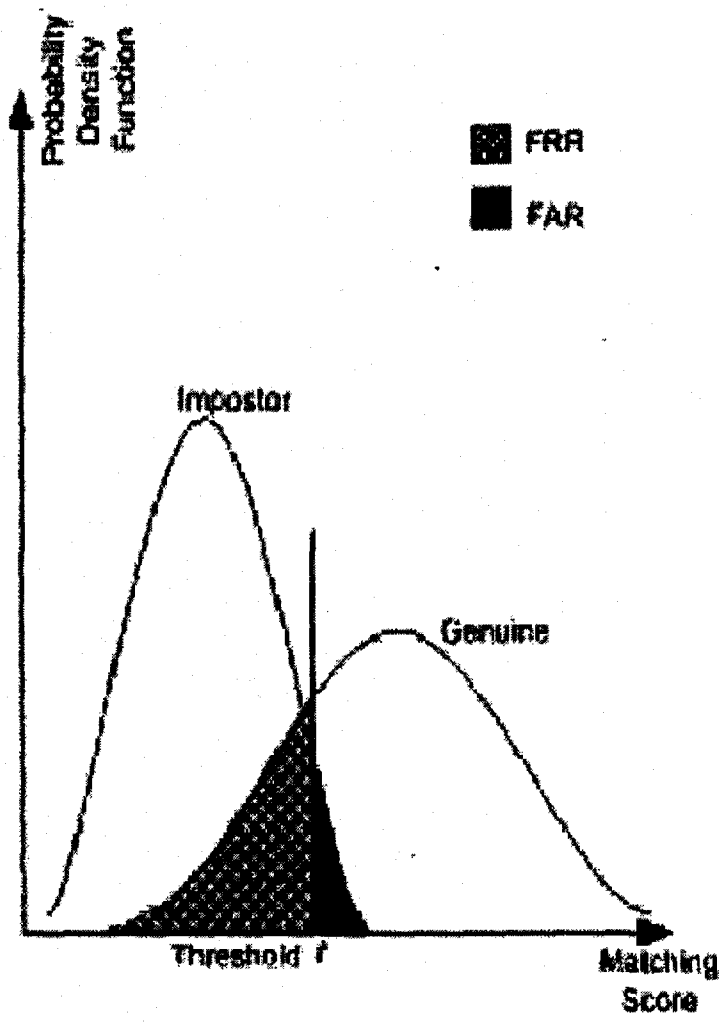


Figure 3. 1 Error rates as function of threshold(Ross, 2006)

Mathematically, let ***p(s|genuine)*** and ***p(s|impostor)*** represent the probability density functions of the score s under the genuine and impostor conditions, respectively. Then for a particular threshold t,

$$\text{FAR}(t) = \int_{t}^{\infty} p(s \mid impostor)ds$$

$$\text{FRR}(t) = \int_{-\infty}^{t} p(s \mid genuine)ds$$

If the match score represents a distance or dissimilarity value, then *FAR(t)* and *FRR(t)* may be expressed as follows:

$$FAR(t) = \int_{-\infty}^{t} p(s \mid impostor)ds$$

$$FRR(t) = \int_{t}^{\infty} p(s \mid genuine)ds$$

It is noticeable from figure 3.1 that there is no way decreasing both these errors simultaneously. The figure illustrates that changing the threshold to decrease FAR increases the FRR. Therefore, if the threshold setting is increased to make the access harder for impostors, some enrollees may find it more difficult to gain access. Determining appropriate thresholds is one of the predominant focuses in performance analysis. It requires knowledge of system scale, estimates of prior probabilities of genuine and impostor subjects, and risk/cost functions for false rejections and false acceptances.

3.1.2 Receiver operating characteristic (ROC)

An ROC is a precise complete specification of a single biometric matcher's performance. It provides a biometric system the ability to distinguish subjects previously known to the system (enrollees) from subjects not known to the system (impostors.).

(Bolle, 2003) depicts the process of generating a ROC curve as follows:
Suppose for the moment that the integrals of FAR and FRR can be evaluated for any threshold T. In a multimodal biometric system, after the calculation of every fusion score from multimodal scores, each fusion score is used as a threshold. Then the functions FAR (T) and FRR (T) give the error rates when the match decision is made at some threshold T. A mapping table of the threshold values and the corresponding error rates (FAR and FRR) are stored. And at last the error rates can be plotted against each other as a two-dimensional curve based on the previous mapping:

ROC (T) = (FAR (T), FRR (T)).

We can operate the matcher using any point (i.e. operating point) on the ROC. But practically we choose a desired operating point based on the FAR or FRR which is meaningful for a particular system. And then we can determine the corresponding threshold from the mapping table.

In many cases, the GAR and the FAR are plotted against each other to yield a ROC curve. The FAR and FRR (GAR) behavior is expressed in terms of a Receiver Operating Characteristic (ROC) curve (Germain, 1999).

Furthermore, ROC is the most common measurement for comparing two or more biometric systems. When comparing two ROCs, one may be consistently superior (its GAR is higher at every FAR). We say that one system is more accurate than another when its ROC is consistently superior. But this is rare case, in most common cases the two curves may cross over, which means the GAR of curve ***a*** is lower than that of curve ***b*** at a specified point of FAR, while GAR of curve ***a*** may be higher than GAR of curve ***b*** at another point of FAR. Hence it is important to ensure we must comparing GARs based on the same FAR which is decided in the context of different biometric applications. Figure 3.2 (Jain, 2004b) displays the typical operating points of different biometric applications. We can see from the figure high security applications focus on the low FMR (i.e. FAR) which requires a low rate of unauthorized user gaining access into the secure system. On the other hand, forensic applications (e.g corpse identification) do not want to miss any possible subject, such that they are tolerant with the false acceptance rate. Furthermore they prefer high FAR because that can bring more conveniences in biometric selection and implementation than system with low FAR.(Bolle, 2003). As (Jain, 2004b) pointed out the lack of understanding of the error rates for a specific application is a primary source of confusion in assessing system accuracy in vendor/user communities.

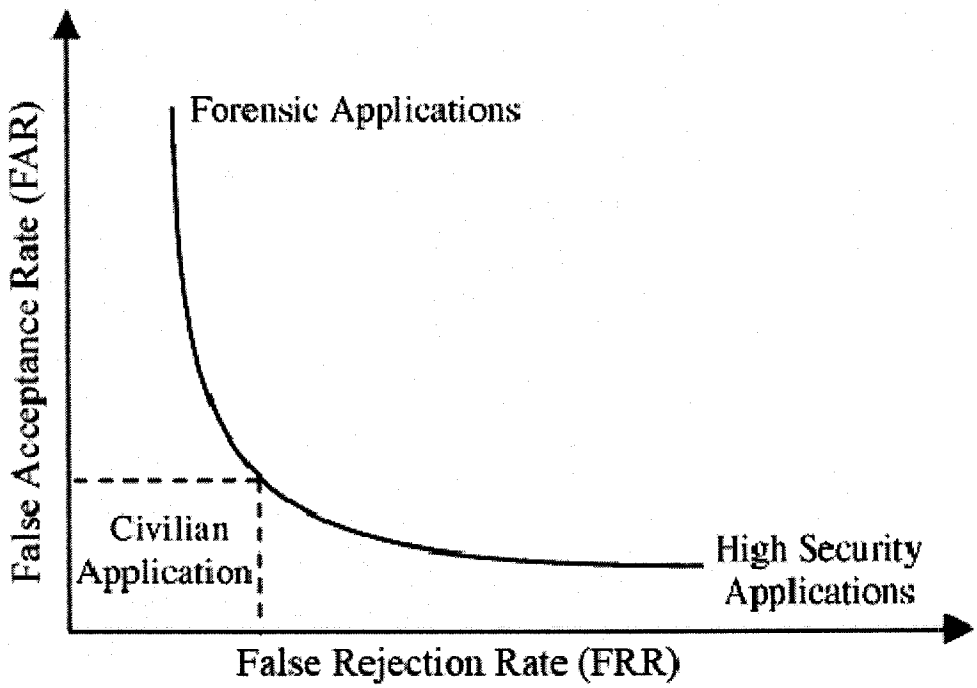


Figure 3. 2 Typical operating points of different biometric applications (Jain, 2004b)

3.1.3 Other accuracy performance metrics

The Equal Error Rate (ERR)

Other than the ROC which presupposes having a known operating point, people are always desiring a performance summary that can reduce the information in the ROC to a single number. Choosing a best matcher for one is just a matter of choosing the one with the best performance figure (Bolle, 2003).

The equal error rate is one of the metrics that have been attempted but with great limitations so far.

The EER point is the point at the intersection of the line FAR=FRR with the ROC of the matcher. The Equal Error Rate is the value of the error rates at the point

EER=FAR=FRR. In figure 3.3 (Bolle, 2003), the Equal Error Rate EER***a*** of matcher ***a*** is clearly less than the Equal Error Rate EER***b***.

The EER can tell us if one system performs better than other but only in narrow range of points FAR = (EE***a***, EE***b***) and FRR = (EE***a***, EE***b***).Beyond that range, the ROC curves may cross over each other and the EER would be invalid as displayed on Figure 3.3. That is why the EER is an unreliable summary of system accuracy.

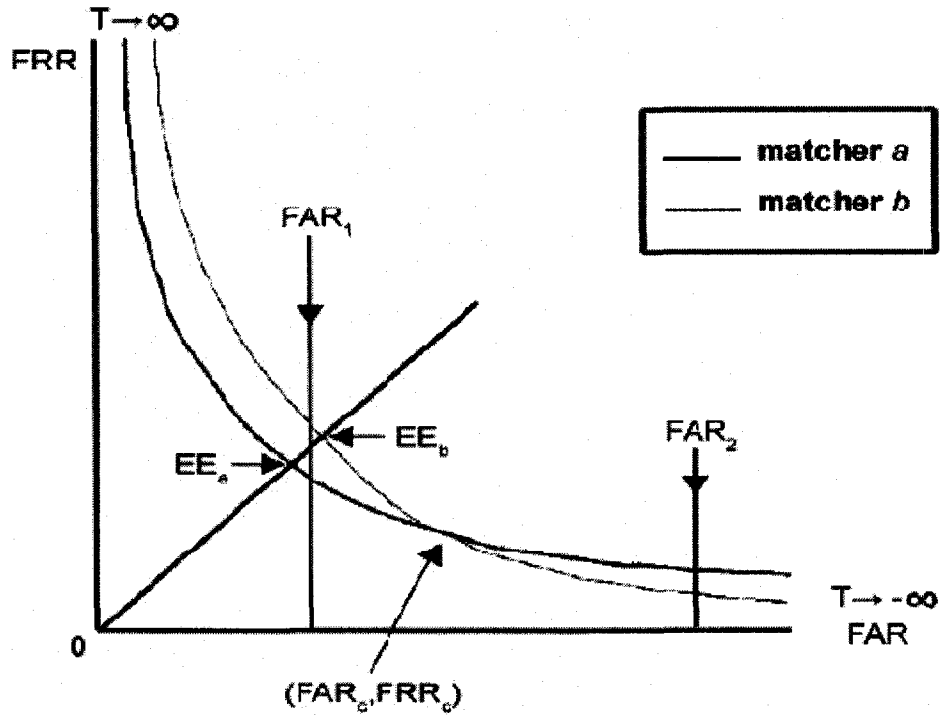


Figure 3. 3 Equal Error Rate Example

The Failure to Acquire Rate (FTA)

Besides the two types of errors (false accept and false reject) indicated in section 3.1.1, a biometric system can encounter other types of failures which will affect the accuracy performance as well. The Failure to Acquire (FTA) (also known as Failure to Capture (FTC)) rate denotes the proportion of times the biometric device fails to capture a sample when the biometric characteristic is presented to it. This type of error typically occurs

when the device is not able to locate a biometric signal of sufficiently good quality (e.g., an extremely faint fingerprint or an occluded face image). The FTA rate is also impacted by sensor wear and tear.

3.2 PERFORMANCE ANALYSIS OF MULTIMODAL BIOMETRICS SYSTEM

3.2.1 Matching performance analysis approach

The evaluation of biometrics systems can be carried out from three different perspectives: technology evaluations, scenario evaluations, and operational evaluations (Philips, 2000).

1) Technology evaluation: Technology evaluation compares competing algorithms from a single technology on a standardized database. Since the database is fixed, the technology evaluation results are repeatable. Some organization, often a government agency, releases databases at some point, and test participants submit their algorithms within some period of time after the release of the test data. The results are then compared over these pre-collected databases.

2) Scenario evaluation: Testing aims to determine the overall performance of a complete system in an environment that closely models a real-world target application. As each tested system will acquire its own biometric data, the evaluation will receive slightly different data even if we acquire samples from the same individuals. Evaluation results may be repeatable only in the carefully controlled condition.

3) Operational evaluation: An operational evaluation involves performance measurement in a real environment with real users. In general, operational evaluation results will not be repeatable.

If a database can be representative of the user population and the inevitable collection problems can be engineered away or statistically modeled better, technology evaluations

are a reasonable and relatively cheap way to compare performance of different biometric systems (Bolle, 2003).

Snelick et al. (Snelick, 2005) have summarized the general testing framework derived from previous works, (Philips, 1996) and (Philips, 2003), for evaluating the matching performance of multimodal biometrics system based on technology evaluation. There are five major steps in this framework:

1) For each modality, two sets of biometric signatures- a target and a query set are assembled respectively. The target set stores the set of signatures of enrolees, i.e. subjects known to the system. The query set contains signatures of users that are to be compared against the target set. Each comparison of query and target signatures generates a matching score and stores in the similarity matrix, whose size is query set size by target set size.

2) Normalization technique is used to transform the matching scores of different modalities and map them into a common domain. The transformed scores representing different biometric modalities are then combined using fusion method into a single fused matching score.

3) Each fused score is used as a threshold and compute the corresponding genuine acceptance rate (GAR) and false acceptance rate (FAR). Those rates (GAR and FAR) and the threshold values will then be stored in a mapping table which derives the plotting of the ROC curve for the system eventually.

4) Repeat steps 1-3 for different combinations of competing matching algorithms, normalization techniques like min–max, z-score, median and MAD, and tanh estimators, and fusion techniques like simple sum of scores, maximum score, minimum score, sum of posteriori probabilities (sum rule), and product of posteriori probabilities (product rule).

5) The ROC curves of combinations of varied factors will be compared and the desired combination of factors will be identified in the context of the applications.

This evaluation framework allows designers to evaluate the matching performance of biometric systems by varying different factors influencing the matching performance such as the biometric traits, matching algorithms, normalization schemes, and fusion methods. Systems can then be built to optimally suit a particular application based on evaluation results.

To illustrate this testing methodology, Snelick et al. evaluated the performance of a multimodal biometric system that used face and fingerprint classifiers and the database for conducting the experiment provides more than 1000 users. The results showed that the min–max normalization followed by the sum of scores fusion method generally provided better recognition performance than other schemes.

FRVT 2000 (Face Recognition and Verification Test) (Balckburn, 2000) was the first attempt to characterize performance measures for assessing commercially available face identification systems. The five participating vendors had to compute an all-against-all match of a database of 13,872 face images with varying parameters of compression, image distance, and facial expression.

There are also many other evaluations on the performance of different biometric systems have been analyzed following the general framework, but methodologies have been extended and adapted according to specific applications. For example, Indovina et al. (Indovina, 2003) carried out their experiments on the virtual multimodal database and found out the variation in matching performance among these virtual user sets is not significant. Wilson et al. (Wilson, 2004b) analyzed the matching performance in conjunction with a watch-list for the US-VISIT IDENT system. A watch-list refers to a database of people who are of some interest. For instance, the FBI may be watching criminals who are on a so-called "do not fly" list at airports. The improved methodology

highlights the potential usefulness of biometric identifiers, such as face and fingerprints, to be associated with the watch-list for better and more reliable outcomes.

3.2.2 Database data partioning

Data partitioning plays an important role in the performance analysis process. The input dataset is partitioned into two sets, the training dataset and testing dataset respectively such that the analysis is initially performed on the training subset, while the testing subset is retained for subsequent use in confirming and validating the initial analysis. In other words, the matching score distribution of the training set is examined and a suitable model is chosen to fit the distribution and the normalization parameters are determined based on the model, and the testing set which is completely separate from the training set will then be used to evaluate the performance of the system by using those parameters derived from the training set.

This statistical practice of partitioning the sample of data into subsets is also called cross validation. There are basically three ways of cross validation (Samoska, 2006):

*1) **Re-substitution validation***

All the available data is used for training as well as testing, training and test sets are the same.

2) ***Holdout validation***

Data will be divided into independent training and test sets according to the specified percentages. The testing dataset is chosen randomly from the initial sample to form the validation data, and the remaining observations are retained as the training data. After normalization parameters are estimated from the training set, the testing set is normalized using these parameters and the fusion method is executed. Normally, less than a third of the initial sample is used for validation data.

3) ***Leave one out validation***.

As the name suggests, leave-one-out cross-validation (**LOOCV**) involves dividing the dataset to n-1 different training samples and 1 testing sample, for N different times. The N results from the folds then can be averaged (or otherwise combined) to produce a single estimation.

Since the normalization parameters depend heavily on the selected data points in the training set and testing set, the subsequent performance metrics such as the probability densities of genuine and impostor scores, the FAR and FRR, and ROC also vary from different data partitioning methods. Consequently the evaluation analysis may be significantly different depending on the way of partitioning the dataset. Figure 3.4 displays the obviously different probability density curves of a fingerprint biometric system based on two different partitioning methods (Re-substitution and Hold-out) and same normalization method (Decimal scaling).

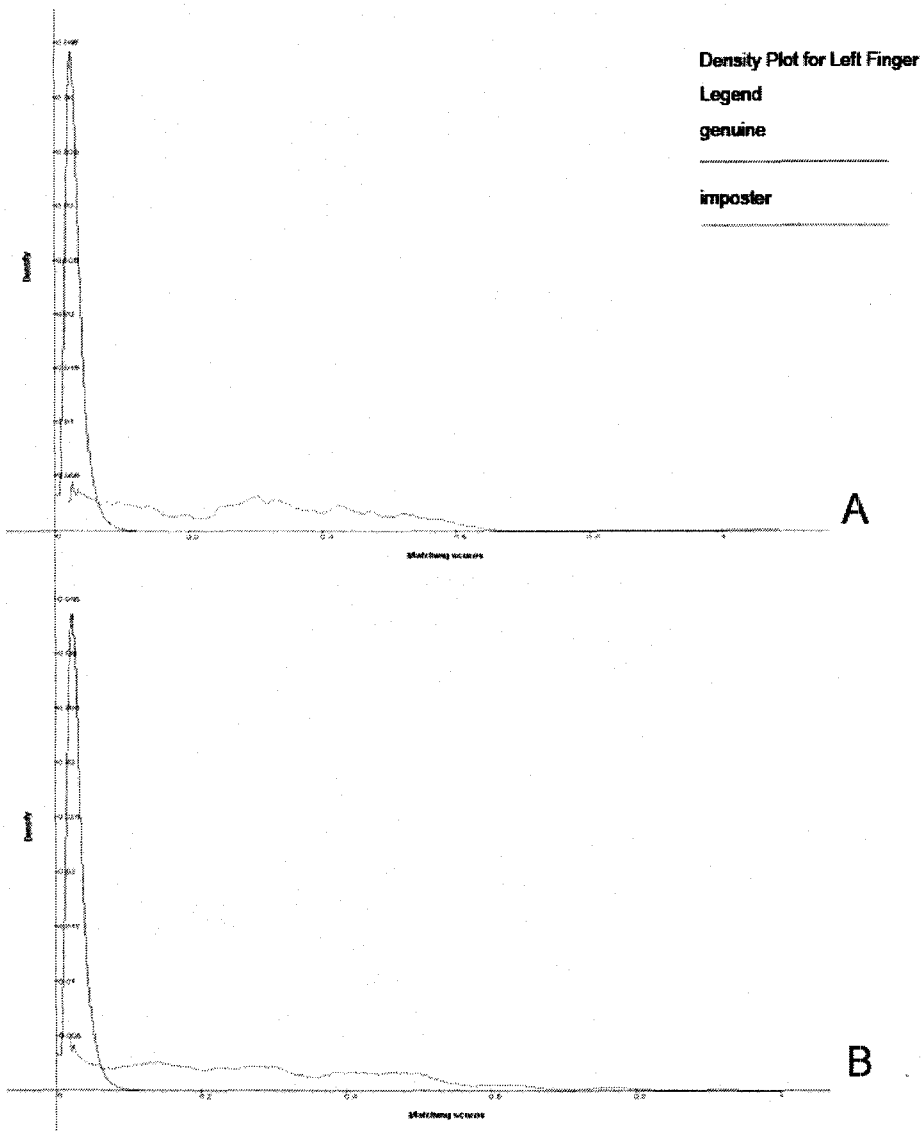


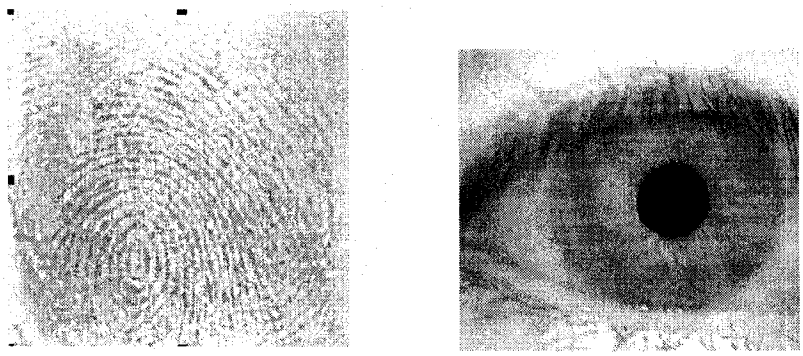
Figure 3. 4 (A) Hold-out partitioning and Decimal scaling normalization

(B) Re-substitution partitioning and Decimal scaling normalization

3.3 NOISE AND MATCHING PERFORMANCE

3.3.1 Noise sources

Noise is an inevitable factor that affects the performance of biometric systems significantly as shown in Figure 3.5.



(a) (b)

Figure 3. 5 Examples of noisy biometric data. (a) A noisy fingerprint image due to smearing deposits; (b) A blurred iris image due to loss of focus

The sources of noise are various: Noise can come from the acquired biometric data through the defective or improperly maintained sensors as we discussed in section 2.2.1 like the accumulation of dirt on a fingerprint sensor or unfavourable ambient conditions Noise can also be due to the user's physical or behavioural characteristic like a fingerprint image with a scar, or impression to impression variation.

(Mansfield, 2002) has analyzed and categorized the factors that could possibly influence the system performance, which are also the sources of noise as in the following list:

1. Population demographics: Age, Gender, Ethnic Origin etc.
2. Application: User familiarity, Time elapsed between enrolment and verification, Time of day (Behaviour and physiology can change during the day) etc.
3. User physiology: Beards& Moustaches, Disability, Height etc.
4. User behaviour: Accent, Facial expression, Movement, Pose etc.
5. User appearance: Contact lens, Hair style, Tattoo etc.
6. Environmental influences: Background (Color, noise or other voices etc), Lighting (Lighting levels, direction, reflection etc), Weather (Temperature, Humidity etc), etc)

7. Sensor and hardware: Sensor wear, Sensor quality, Sensor variations, Transmission channel etc)

8. User interface: Feedback(e.g. Did they see their submitted fingerprints?), Instruction, Supervision (e.g. User attempts due to the differences and changes in supervisors) etc)

Although noise can be summed up in eight categories, the exact list of sources of noise can never be enumerated.

3.3.2 Influence of noise on matching performance

The matching performance of biometrics system is analyzed based on the matching scores. Matching score is a measure of the similarity between features derived from a presented sample and a stored template. However, essentially we should use this metric: P (genuine| X), which is the posteriori probability of user being "genuine" given the input biometric sample X.

Verlinde et al. (Verlinde, 1999) have examined the relationship between the matching score S and the input biometric sample X. It is revealed that the matching score S is related to P (genuine| X) as follows:

$S = f(P(\text{genuine}|X)) + \eta(X)$

Where f is a monotonic function and $\eta(\mathbf{X})$ is the error made by the biometric system that depends on the input biometric sample X. This error could be due to the noise introduced in the previous section or error made by feature extraction and matching processes.

If we assume that $\eta(\mathbf{X})$ is zero, we will have $S = f(P(\text{genuine}|X))$, which means the output matching score S can accurately reflect the capacity of the system in identifying the system input – biometric sample X. However, $\eta(\mathbf{X})$ can not be zero as noise is existent everywhere and every time. Furthermore, the value of $\eta(\mathbf{X})$ is uncertain because

the noise varies greatly among different applications, points of time and environments etc.

Consequently matching scores can not exactly reflect the capacity of the system under the influences of noise, and the performance analysis results based on matching scores may change under different noise influences.

3.3.3 Related work to reduce noise

Being aware the existence of noise, many endeavors have been devoted to reduce it throughout the authentication process from the acquiring phase to decision phase. For example the feature extraction typically engages enhancement operations to suppress the inherent noise from the acquired raw data (Sanderson, 2003). However, the enhancement procedure in itself may add spurious (e.g. extraction errors) information to the original raw data, so does the matching process (Ross, 2006). Thus, noise is existent at any stage in a biometric system and it can not be eliminated completely. Consequently, more and more efforts have been attempted to analyze the performance of biometrics system under the influence of noise and to develop approaches that make the robust performance under the noise disturbance.

Some researches have been focusing on the effect of influencing factors on the performance, and have attempted to figure out the extents the performance of the system varies with variations of the particular factors because of noise. Given et al. (Givens, 2004) proposed the use of ANOVA (Analysis Of Variance) to study the statistical effects of demographic features such as age, sex, facial hair, etc on face recognition performance. Mitra et al. (Mitra, 2007) further extend the previous fixed effects models to a random effects model so that performance of the system on the potentially different databases can be predicted by using various explanatory variables incorporated with the random effects model.

If all sources of noise, such as sensor noise, feature noise and distortions between pairs of matching face templates could be modeled, the error rates could be computed analytically. However, the most difficult issue is the modeling of all sources. As we have discussed, it is clear there is no possibility to identify and model all noise sources (Bolle, 2000). Therefore those researches are only limited to the analysis of the effect on the performance by several or a few particular noise which are known in the evaluation. The analysis on the overall performance of biometrics system influenced by all possible noise can not be conducted in this way.

Other efforts, from another perspective, take the whole environment into consideration to investigate the performance of recognition systems under various environments when the noise changes. Wang and Ji (Wang, 2006) have introduced a concept of "perfect recognition" which depends on the intrinsic structure of a recognition system to model the performance of face recognition without empirical testing. System performance can be used to select system parameters offline to achieve optimal or near-optimal performance.

There are two major drawbacks of this method when applying it to evaluate the performance under various environments caused by system errors. First of all, the statistical model needs to explicitly identify each possible environment that affect the performance, which is an extremely difficult task in practical implementation. Consequently it cannot totally model the performance under all possible environments under the noise disturbance. Secondly, as the crucial metric is extracted from perfect recognition similarity scores (PRSS) and is inherently dependent on the particular device, the evaluation is difficult to achieve when the number of competing devices increases because of the heterogeneous issues.

CHAPTER 4 PROPOSED STATISTICAL METHODOLOGY FOR PERFORMANCE ANALYSIS

4.1 PROBLEM DOMAIN

As we have discussed, the traditional testing framework allows system designers to evaluate multimodal biometric systems by varying different factors like the biometric traits, matching algorithms, normalization schemes, fusion methods and sample databases. The performance of multimodal system, however, is evaluated at a point of time and the result is affected by a certain noise factor. If we conduct the same evaluation for several times, the results of repetitions may have variations because different noise will have different influences on the process. Moreover, these variations may lead to a diverse evaluation result differing from the previous one. Therefore, the evaluation of multimodal system should take the inevitable noise factors into consideration even though they are out of designer's control.

Another challenging problem confronting us thereafter is that the experiments will usually take several days or weeks with the exponential growth of combinations of variety of noise factors and the system parameters when we are intending to carry out a systematic research on them.

Little work has been done to address the above problems. In this chapter, we are proposing a statistical methodology that helps to determine optimum configurations of the multimodal system in the presence of uncontrollable noise disturbance in an efficient and systematical way.

4.2 ROBUST PAPAMETER DESIGN

The proposed methodology falls into the category of robust parameter design for system performance. The objective of robust parameter design is to select the optimum levels for the controllable system parameters so that the system will be functional, will exhibit a

high level of performance under a wide range of conditions, and will be robust against noise factors that cause variability.

Design of experiments (DoE), which is the most efficient approach for organizing experimental work, is offered as an empirical method of robust parameter design. DoE selects a diverse and representative set of experiments in which all factors are independent of each other despite being varied simultaneously. The result is a causal predictive model showing the importance of all factors and their interactions. These models can be summarized as informative contour plots highlighting the optimum combination of factor settings. It involves many experimental methods like comparison, randomization, replication and use of factorial experiments instead of the one-factor-at-a-time method etc.

4.3 DESIGN OF EXPERIMENTS

4.3.1 P-diagram

DoE begins with determining the objectives of an experiment and selecting the process factors for the study. The Parameter Diagram (P-Diagram) (Ross, 1995) has been utilized as a visualization tool for the understanding the well-defined development scope of a software system and the identification of the design specifications, control factors, and noise factors that affect the quality characteristic of a system. As shown in Figure 4.1 this schematic diagram includes control factors, noise factors, signal factors and performance metrics. Signal factors refer to the input of the system; Control factors are the parameters that can be specified by the designer; and noise factors, in the other way around, are the parameters beyond the control of the system designer.

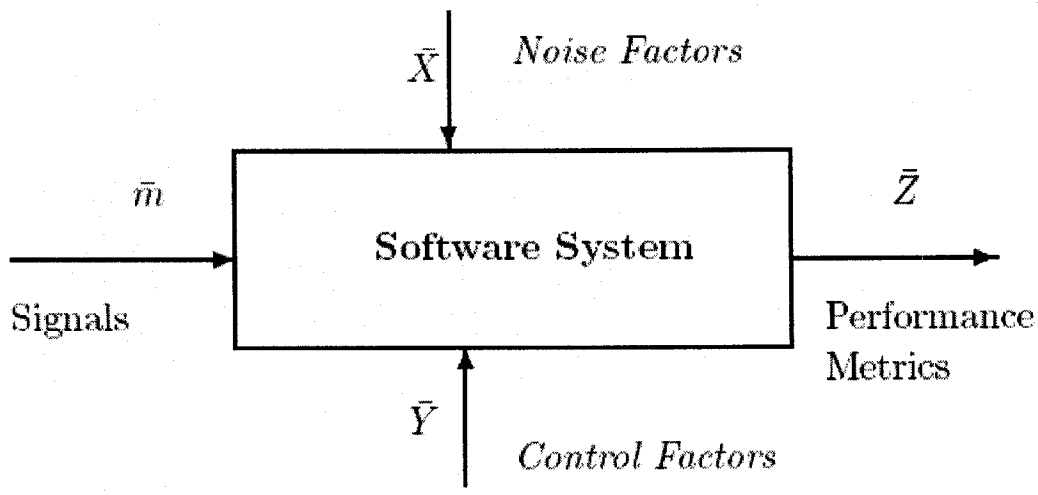


Figure 4. 1 P-Diagram of Software System

Our objective is to determine the most robust parameters configuration in the consideration of multimodal biometrics system performance. Under the general testing framework, we identified the signal factors as the extracted feature data or the matching scores of different biometric traits, the control factors could be the varied feature extraction algorithms, matching algorithms, data partitioning methods, normalization methods, and fusion methods etc, the noise factor are variety of operational imprecision or errors occurred in the process of evaluation as we have discussed in section 3.3.1. However, we may consider there are n types of noise in a multimodal system with n modalities in part because one modality can use just one noise as an aggregate of all the noise occurred in this modality. In addition, different modalities have different types of noise as they are independent from each other.

The performance metric is selected from various scenarios that could be the Genuine Acceptance Rate (GAR) or False Acceptance Rate (FAR) at a specific False Acceptance Rate (FAR), the Equal Error Rate (EER) or the Failure to Acquire Rate (FAR) etc.

4.3.2 Gaussian noise model

Since this proposed methodology is designed to study the system performance under noise disturbances, we need to investigate all the possible noise in a systematic way. The Gaussian noise model is utilized in aiding in the simulation of all the possible noise.

An ideal matcher will generate perfect scores with zero deviation for all matching pairs of the same finger/face. In practice, poor quality matching pairs and the errors occurring in extraction/matching process result in significant deviations for the matching scores. In experiment, we allow the assumption (which are valid for many applications) that the deviations of each matching score for a single modality are distributed following a zero-mean Gaussian distribution model, where the deviations can be described by its variance (σ sigma). (The 1-D Gaussian distribution has the form shown in Figure 4.2) In other words, the simulated matching score will be the sum of the matching score without noise and a random, Gaussian distributed noise value. Nevertheless, different noise factors may cause varied deviation ranges as shown in Figure 4.2, three different Gaussian distributions represent the deviations caused by three different noise. We use deviation rate which is the ratio of the maximum score deviation over the score scale of a specific modality matcher to characterize the deviation range of the noise caused by the matcher.

Please note that the matching scores without noise do not exist in the real world because of the irresistant existence of noise. Practically to conduct our experiments we adopt the matching scores acquired by domain experts as the approximated noise-free matching scores. Because domain experts, compared to other users of the system, are experienced in reducing the noise such that the deviations from the ideal matching scores can be limited to the minimum.

On the grounds of Gaussian noise model and deviation rate, the noise factors now can be investigated systematically.

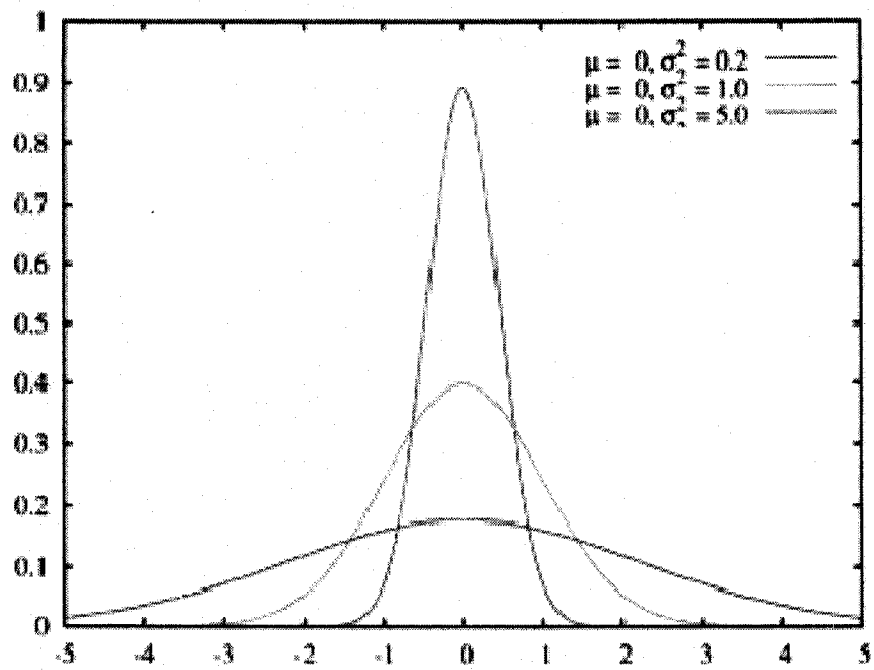


Figure 4. 2 1-D Gaussian distributions with mean 0 and different variances σ

4.3.3 Levels specification

After the parameters of control factors and noise factors have been identified, we need to decide the levels for each parameter. That is, the test values for every type of parameter in the design of experiments. It is easy to conclude that in the testing framework of multimodal biometrics system, the levels for control factors are the different partitioning methods for partitioning parameter, different normalization methods for normalization methods, different fusion methods for fusion parameter etc. These types of parameters are all discrete values. However, noise factors keep the presence in the form of continuous values. This situation can be resolved by expressing the levels as interval values (Llado, 2002). In a strategic manner we can specify several intervals for the levels of noise factors that influence the performance metrics significantly.

4.3.4 Orthogonal arrays

When the parameters and levels have been identified, the full factorial experiments can be carried out to study the influence of different combinations of factors on the system

performance in a systematic way. However, the experiments can easily become extremely costly and time-consuming with the exponential growth in the combination of parameters when the system becomes more and more complicated. Statisticians have developed more efficient test plans, which are referred to as Orthogonal Arrays (OAs). OAs use only a portion of the total possible combinations to estimate the main factor effects and interactions (Hedayat, 1999).

Orthogonal Arrays are characterized by the number of parameters and their levels. The appropriate OA can be retrieved from the references like (Ross, 1995) or online resources (e.g. the Orthogonal Array Library maintained by N.J.A. Sloane available at http://www.research.att.com/~njas/oadir/index.html).

4.3.5 Evaluation matrix

Traditionally the idea of exploring the system performance under noise disturbance is to have a full factor-effect-analysis by checking all the possible combinations of all the factors including control factors and noise factors until the combination for best performance is found. Nevertheless, there is a major drawback for this method. If the combination is finally found, can instructions be supplied with the system to tell users to apply the system only according to that combination which includes uncontrollable noise factors? The answer is impossible. We may instruct the users to use the optimum configuration of controllable factors by separating the control factors from the noise factors and to find some combination of control factors most robust to different noise combinations. Furthermore, with the power of Orthogonal Array to evaluate several factors in a minimum of tests, the experiments can be conducted in an efficient but still systematic way.

The Evaluation Matrix is applied based on the above two ideas. The template of an evaluation matrix consists of three regions (Table 1). The left region contains u control factors (cf), n combinations of the control factors (cfc), and an $n \times u$ array of control factor combination values (cfv) assigned by the orthogonal array of control factors. Similarly, the right-upper region contains v noise factors (nf), m combinations of the noise factors

(nfc), and a v×m array of noise factor combination values (nfv) assigned by the orthogonal array of noise factors. The right-lower region is an array R whose elements $r^{i,j}$, $1 \leq i \leq n$ and $1 \leq j \leq m+2$, collect experiment results and analysis values like mean and S/N. The form of evaluation matrix makes it capable to simulate the variation in the performance due to the noise parameters and to determine the optimal configuration with proper measurements

						$\mathbf{nfc}_1$	⋯	$\mathbf{nfc}_j$	⋯	$\mathbf{nfc}_m$		
						$nfv_{1,1}$	⋯	$nfv_{1,j}$	⋯	$nfv_{1,m}$	$\mathbf{nf}_1$	
						⋯	⋯	⋯	⋯	⋯	⋯	
						$nfv_{l,1}$	⋯	$nfv_{l,j}$	⋯	$nfv_{l,m}$	$\mathbf{nf}_l$	
						⋯	⋯	⋯	⋯	⋯	⋯	
	$\mathbf{cf}_1$	⋯	$\mathbf{cf}_k$	⋯	$\mathbf{cf}_u$	$nfv_{v,1}$	⋯	$nfv_{v,j}$	⋯	$nfv_{v,m}$	$\mathbf{nf}_v$ / mean	S/N
$\mathbf{cfc}_1$	$cfv_{1,1}$	⋯	$cfv_{1,k}$	⋯	$cfv_{1,u}$	$r_{1,1}$	⋯	$r_{1,j}$	⋯	$r_{1,m}$	$r_{1,m+1}$	$r_{1,m+2}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
$\mathbf{cfc}_i$	$cfv_{i,1}$	⋯	$cfv_{i,k}$	⋯	$cfv_{i,u}$	$r_{i,1}$	⋯	$r_{i,j}$	⋯	$r_{i,m}$	$r_{i,m+1}$	$r_{i,m+2}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮
$\mathbf{cfc}_n$	$cfv_{n,1}$	⋯	$cfv_{n,k}$	⋯	$cfv_{n,u}$	$r_{n,1}$	⋯	$r_{n,j}$	⋯	$r_{n,m}$	$r_{n,m+1}$	$r_{n,m+2}$

Table 1 Template of an Evaluation Matrix

4.3.6 Signal-to-noise ratio (S/N)

After the experiments have been conducted, the final step of DoE is to identify the optimal control parameters configuration within the system under evaluation.

Instead of many analyses just addressing which factors might affect the average response (i.e. mean value of performance) the Signal-to-Noise Ratio method is used to take both the mean and variation into account. The method consolidates all the repetitions for the same control parameter combination to reflect the amount of variation present and transform them into another value for measuring the variations, namely signal-to-noise ratio (S/N) (Ross, 1995). The measurement of the ratio is stated as the ratio of signal level to noise level, normally expressed in decibels (dB).

There are three typical types of S/N ratios available depending on the characteristics of the system under evaluation. They are called lower is better (LB), nominal is the best (NB), and higher is better (HB).

We can choose the appropriate formula for different performance metrics. For Genuine Acceptance Rate (GAR), HB is chosen and for other metrics like Failure to Acquire Rate (FAR), LB should be applied. NB is not applicable in the practice of multimodal biometrics system evaluation.

Finally, elements $r_{i,m+2}$ for $1 \leq i \leq n$, in the last column of array R are calculated by the formulas given in Eq.4.1 for HB or Eq.4.2 for LB (Ross, 1995).

$$r_{i,m+2} = -10 \log \left(\frac{1}{m} \sum_{t=1}^{m} \left(\frac{1}{r_{i,t}} \right)^2 \right) \qquad \text{Eq. 4.1}$$

$$r_{i,m+2} = -10 \log \left(\frac{1}{m} \sum_{t=1}^{m} r_{i,t}^2 \right) \qquad \text{Eq. 4.2}$$

CHAPTER 5 EXPERIMENTS AND DISCUSSION

5.1 EXPERIMENTAL ENVIRONMENT

5.1.1 NIST BSSR1 database

5.1.1.1 Why choose NIST BSSR1

According to the performance testing approach we have discussed in section 3.1.1, we need a database to conduct our proposed methodology. There are a number of multimodal biometrics databases for performance analysis released by different organizations or government agencies publicly (Ulery, 2006), we chose the NIST (National Institute of Standards and Technology) BSSR1 (Biometric Scores Set - Release I) dataset based on the following two reasons:

1. NIST BSSR1 is a true multimodal database

   Multimodal database can be either true or virtual. However, Poh and Bengio argue that using virtual versus true multimodal databases to evaluate the performance needs further investigation (Poh, 2005); we choose a true multimodal database for the reliable performance analysis.

2. NIST BSSR1 is the largest true multimodal biometric database among all the public domain.

   The performance metrics of a biometric system such as accuracy, throughput, and scalability can be estimated with a high degree of confidence only when the system is tested on a large representative database (Ross, 2006).

5.1.1.2 NIST BSSR1 Overview

The NIST BSSR1 (NIST, 2004) is a multimodal biometric match score database. There is no face and fingerprint images of the subject available in the dataset but matching scores.

BSSR1 is comprised of face and fingerprint matching scores from the same set of 517 individuals. For each individual, the set contains one score from the comparison of two right index fingerprints, one score from the comparison of two left index fingerprints, and two scores (from two separate matchers referred to as 'C' and 'G') from the comparison of two frontal faces. So, there are four match scores for each subject (one for each modality) as shown in figure 5.1.

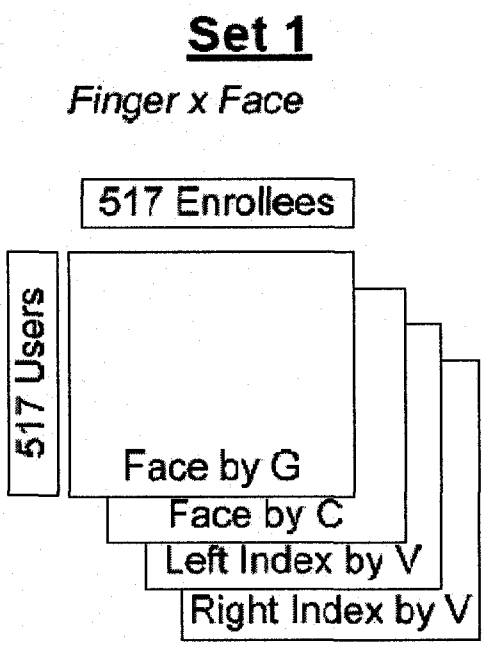


Figure 5. 1NIST (BSSR1) dataset (NIST, 2004)

5.1.1.3 Data structure

A matching score results from the comparison of two images, representing the comparison of an enrolled user's image (gallery set) with a subsequent image of either the same or another user (probe set).The gallery set has 517 subjects which also comprise the probe set. Whenever a comparison for any modality happens, a similarity file is generated which contains a genuine score and the full cross-comparison non-matching scores, i.e.516 genuine scores. Therefore, for each modality, there are 517 similarity files which have 517 genuine scores and 266,772 (516×517) impostor scores.

Along with the similarity files, there is a user.xml file with entries for every similarity file. Every entry records the similarity file location and name in the tree, as well as a unique subject_id for the similarity file. By figuring the k-th entry of the similarity file in the users.xml file, we can recover the genuine score from the corresponding similarity file which is also the k-th score. The other 516 scores in the similarity file are certainly impostor scores.

Since each modality is independent from each other, they have adopted four different score scales. Table 2 shows the score scale and Min/Max values for each modality.

Modality	Score Scale	Minimum Value	Maximum Value
Face C	1	-1	0.898
Face G	50	54.835	83.494
Left Index Finger	250	0	246
Right Index Finger	250	0	257

Table 2 Match Score In Four Modalities

5.1.2 BSSR PROCESSOR

In order to process the data in BSSR1 dataset in accordance with the proposed performance analysis methodology, the BSSR Processor is implemented in Java 5 which benefits from its portability to any operating system containing JAVA installation.

The BSSR Processor includes two components, the score extractor and Gaussian noise generator respectively. The score extractor generate two comma delimited files for each modality, one file containing the genuine scores for the modality and the other file

containing all the impostor scores. These two types of matching scores are extracted and reorganized by the extractor component based on the BSSR1 dataset mechanism.
The Gaussian noise generator is designed to simulate the matching scores for any modality in the noise factor condition by using Gaussian noise model. The scores can be generated on the basis of original dataset when specifying the score scale for a specific modality and the deviation ratio caused by the Gaussian noise. The deviation ratio may be set according to the experiment designing. In this way, the Gaussian noise generator permits the capacity to generate matching scores for a modality with any score scale and in any assumed Gaussian noise condition.

5.1.3 MUBI off-line analysis tool

Our experiments have been carried out using the Multimodal Biometric (MUBI) analyser which can be downloaded from the Center for Identification Technology Research (CITeR) web site at department of Computer Science and Electrical Engineering in West Virginia University (http://www.citer.wvu.edu/downloads/software.php). MUBI was developed as an independent multimodal biometrics system analysis tool in a great effort to empower biometric system designers to evaluate different normalization and fusion methods and to choose the "the best" integration techniques in the context of their application (Samoska, 2006).

The inputs of MUBI analyser are the genuine and the impostor scores for each modality. Several modalities can be added to make up a multimodal biometrics system so as to evaluate the performance of this hypothetical system as shown in figure 5.2.

After the modalities have been added to a system, the densities of genuine and impostor scores for each modality can be plotted, the data partitioning of a chosen method can be created and a number of normalization and fusion methods then be employed. A ROC curve will eventually be plotted for the system designer to study the performance of the selected combination of techniques as shown in Figure 5.3.

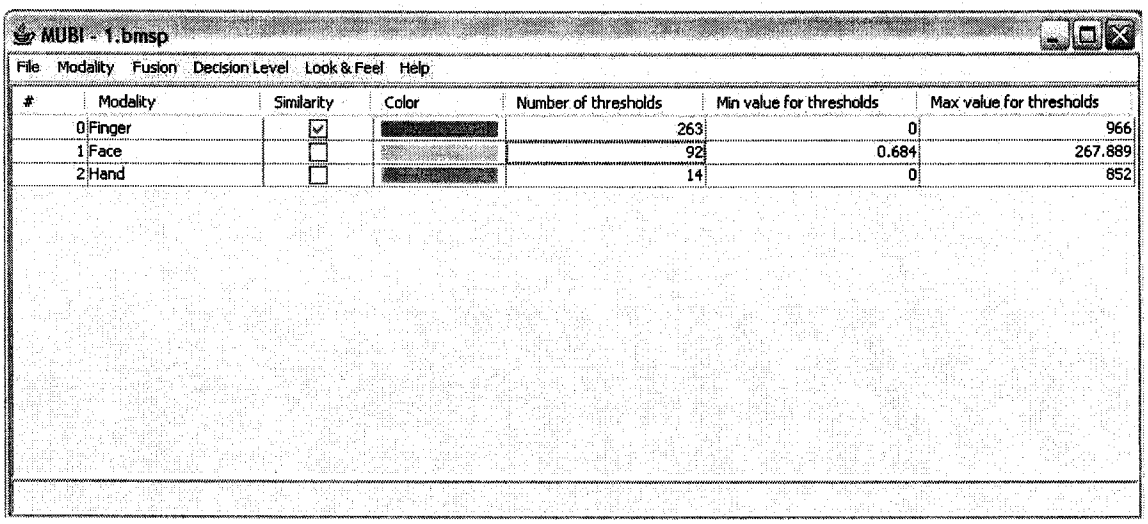


Figure 5. 2 MUBI Tool

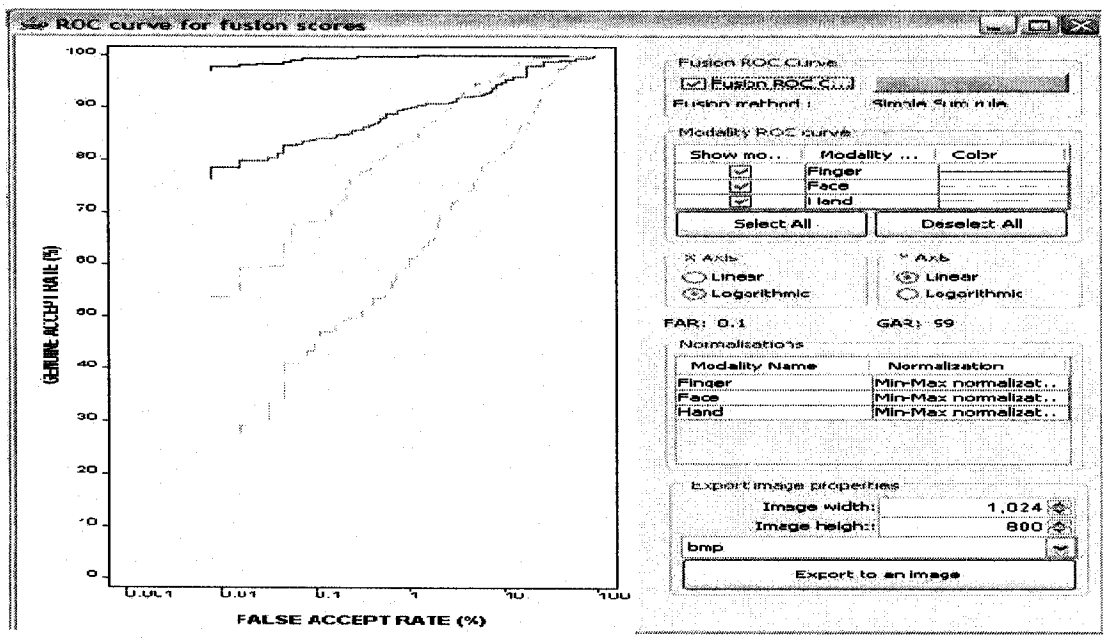


Figure 5. 3 ROC curve plot from MUBI tool (Samoska, 2006)

5.2 EXPERIMENTS

5.2.1 Identifying P-diagram parameters and levels

By examining the performance analysis framework for the multimodal biometrics system, we identified three major parameters for control factors (in capital letters) and corresponding levels of each parameter (in numbers) that affect the performance outcome significantly as following:

"A"--- Partitioning methods 1).Re-substitution 2).Hold-out 3).Leave-one-out;
"B"--- Normalization methods 1).Min-Max 2).Decimal Scaling 3).Z-Score 4).Median and MAD 5).Tanh-Estimators;
"C"--- Fusion methods 1).Simple Sum 2).Simple Product 3).Simple Minimum 4).Simple Maximum 5).BGI;

Meanwhile, the four modality matchers whose information are corrupted by the noise independently can be identified as four different parameters for noise factors (in lowercases) and each of the four parameters may have three different noise deviation levels (in numbers) as following:

"a"---Face C modality matcher 1) within1% deviation 2) within 5% deviation 3) within 10% deviation);
"b"---Face G modality matcher 1) within 1% deviation 2) within 5% deviation 3) within 10% deviation);
"c"---Left Index Finger modality matcher 1) within 1% deviation 2) within 5% deviation 3) within 10% deviation);
"d"---Right Index Finger modality matcher 1) within 1% deviation 2) within 5% deviation 3) within 10% deviation);

In the context of our application, we identified the performance metric for our experiments as the Genuine Acceptance Rate (GAR) (%) at 0.1% False Acceptance Rate (FAR) of different multimodal biometrics system.

The P-Diagram encompassing the above parameters and levels is shown in Figure 5.4.

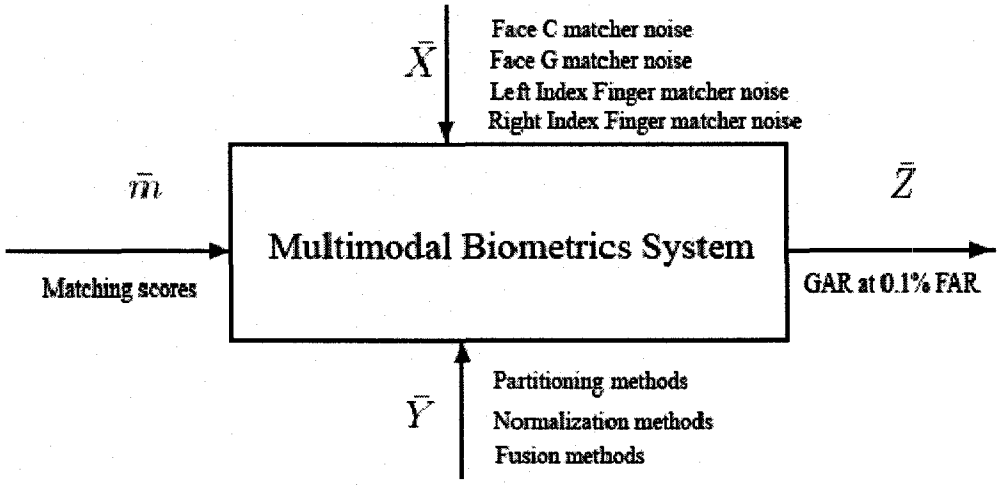


Figure 5. 4 P-Diagram of Multimodal Biometrics System

5.2.2 Gaussian noise model for matching scores

According to the instructions in section 4.3.2 we take the NIST dataset as the roughly noise free matching scores. Gaussian noise model and deviation rate will be applied based on the NIST dataset.

Since Multimodal biometrics system has different score scales for each modality, (In BSSR1 dataset, Face C scores are within [0, 1], Face G scores are within [50, 100], Left Index Finger and Right Index Finger scores are both within [0, 250]), we characterize the possible spread range of noise values by three levels of deviation rates, within 1%, within 5%, within 10% of the original matching score respectively. Because the deviation extents influenced by most common noise generally fall within these three levels.

Each modality out of the four modalities is independent and their noise values may fall in any level of deviation extent when conducting experiments every time. We therefore generate all the possible combinations of matching scores of different modalities at three noise levels.

5.2.3 Orthogonal arrays (OA) and evaluation matrix

As we have discussed previously, the number of experiments would be large if we carry out our experiments by using full factorial experiment. In that scenario, the combinations for uncontrollable noise factors alone are $3\times3\times3\times3=81$. And this number has to multiply the number of control factors combinations for the final evaluation matrix.

Equipped with the power of Orthogonal Array to evaluate several factors in a minimum of tests, therefore, we can derive an orthogonal array L9 for the noise factors which slashed the number of combinations to 9 and it is around 11% of the original. Table 3 is the OA for noise factors.

a	1	1	1	2	2	2	3	3	3
b	1	2	3	1	2	3	1	2	3
c	1	2	3	2	3	1	3	1	2
d	1	2	3	3	1	2	2	3	1

Table 3 Orthogonal Array for noise factors

Compared to noise factors OA, the OA for control factors is a little complicated due to the facts that BGI fusion method is capable of both normalizing and fusing the matching scores, it does not require separate normalization before using BGI fusion method and BGI can not be done with leave-one-out partition method.

We came up with a way to split BGI out the other four fusions for a full factorial experiment combined with partitioning methods. We then have two combinations when BGI is present,

A 1 2

B * *

C 5 5

And the remaining combinations can still be reduced by using orthogonal array. Now we have partitioning parameter with 3 levels, normalization parameter with 5 levels and fusion parameter with 4 levels. We should modify basic Orthogonal Arrays to accommodate a mixture of three-, five-, and four-level factors. The following L12 array (Table 4) is the solution for mixed level design.

A	1	1	1	1	2	2	2	2	3	3	3	3
B	1	2	3	4	1	2	3	5	1	2	4	5
C	1	2	3	4	2	3	4	1	3	4	1	2

Table 4 Orthogonal Array for control factors

Consequently, we can build a 14 ×9 evaluation matrix for this multimodal biometrics system as table 5 according to the proposed methodology.

5.2.4 Signal to noise ratio (S/N)

The signal to noise ratio (S/N) is chosen accordingly following "the higher the better" (HB) rule in the proposed methodology, the ratio equation is given as following:

$$r_{i,m+2} = -10\log\left(\frac{1}{\mathrm{m}}\sum_{t=1}^{m}\left(\frac{1}{r_{i,t}}\right)^2\right)$$

Where $r_{i,j}$ is the GAR value at 0.1% FAR and m is the number of noise combinatioins (size of noise array).

				1	2	3	4	5	6	7	8	9	No.	S/N Ratio (dB)
				1	1	1	2	2	2	3	3	3	a	
				1	2	3	1	2	3	1	2	3	b	
				1	2	3	2	3	1	3	1	2	c	
No.	A	B	C	1	2	3	3	1	2	2	3	1	d/Mean	
1	1	1	1											
2	1	2	2											
3	1	3	3											
4	1	4	4											
5	2	1	2											
6	2	2	3											
7	2	3	4											
8	2	5	1											
9	3	1	3											
10	3	2	4											
11	3	4	1											
12	3	5	2											
13	1	*	5											
14	2	*	5											

Table 5 Evaluation Matrix for NIST BSSR1

5.3 EXPERIMENTAL RESULTS

5.3.1 Matching scores disrtibution and gaussian noise model

After the extraction of genuine and impostor scores for each modality, we plotted the probability density functions (pdfs) of genuine and impostor scores for each modality from the original dataset to analyse the distribution of two types of scores as Figure 5.5, Figure 5.6, Figure 5.7, and Figure 5.8 respectively.

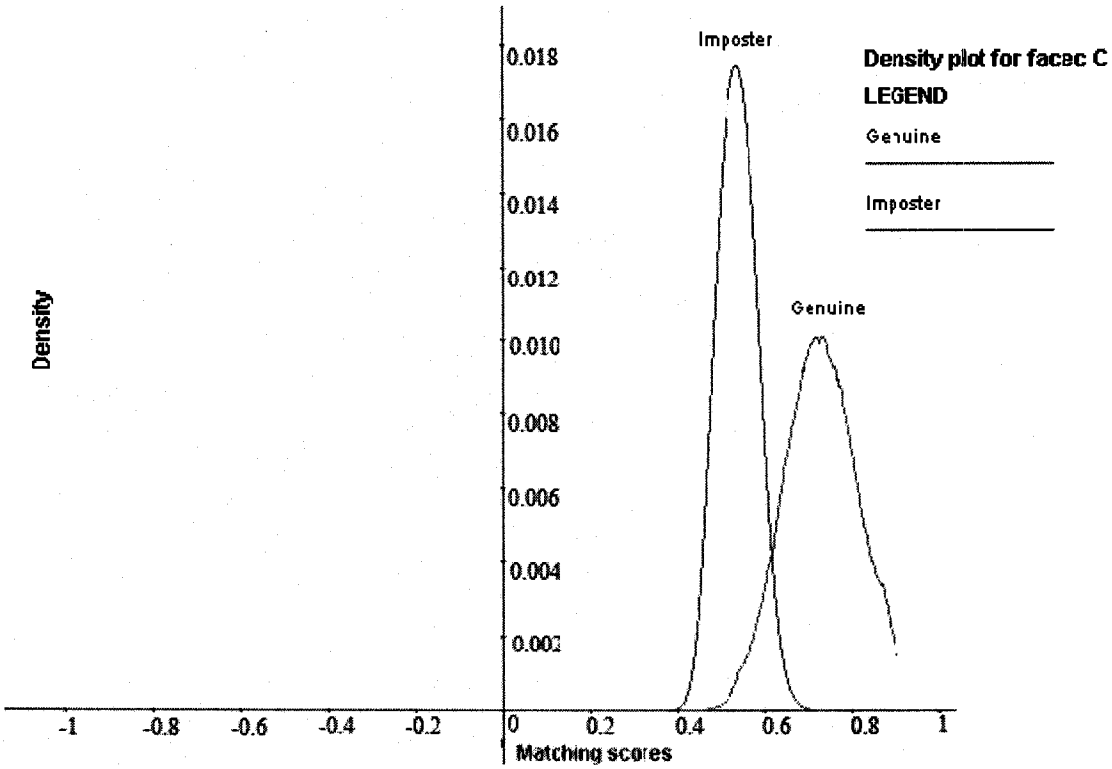


Figure 5. 5 Probability Density Function plot for Face C

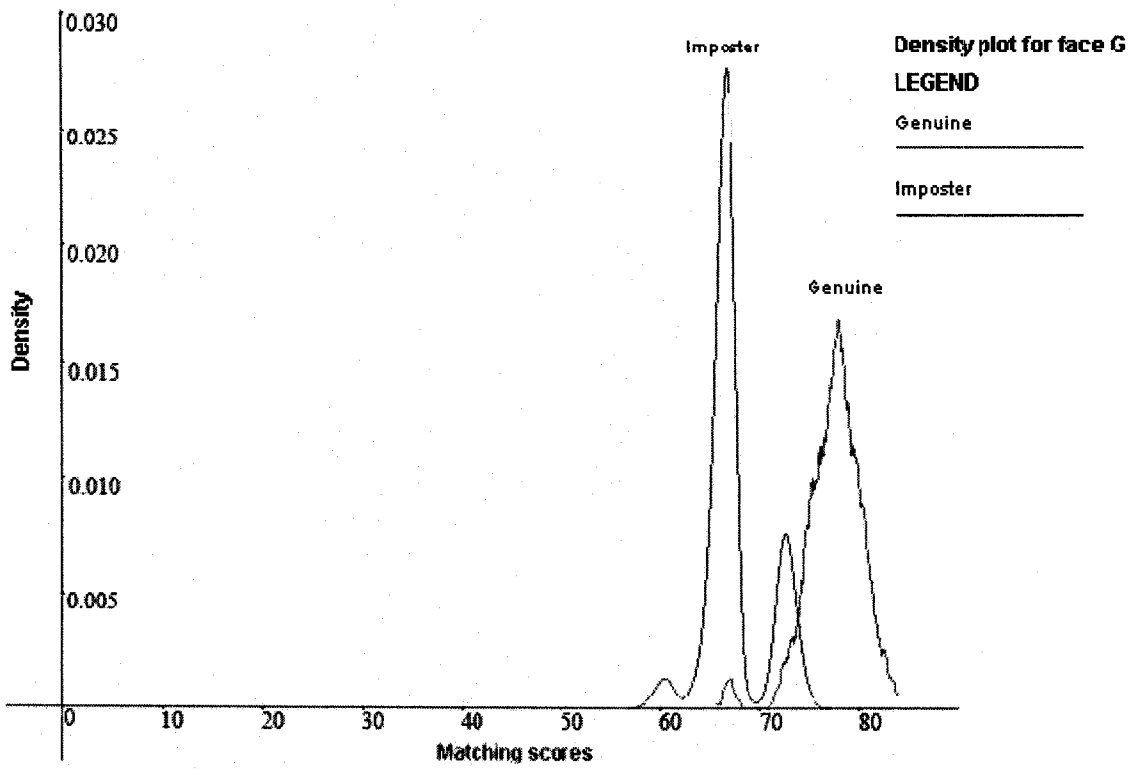


Figure 5. 6 Probability Density Function plot for Face G

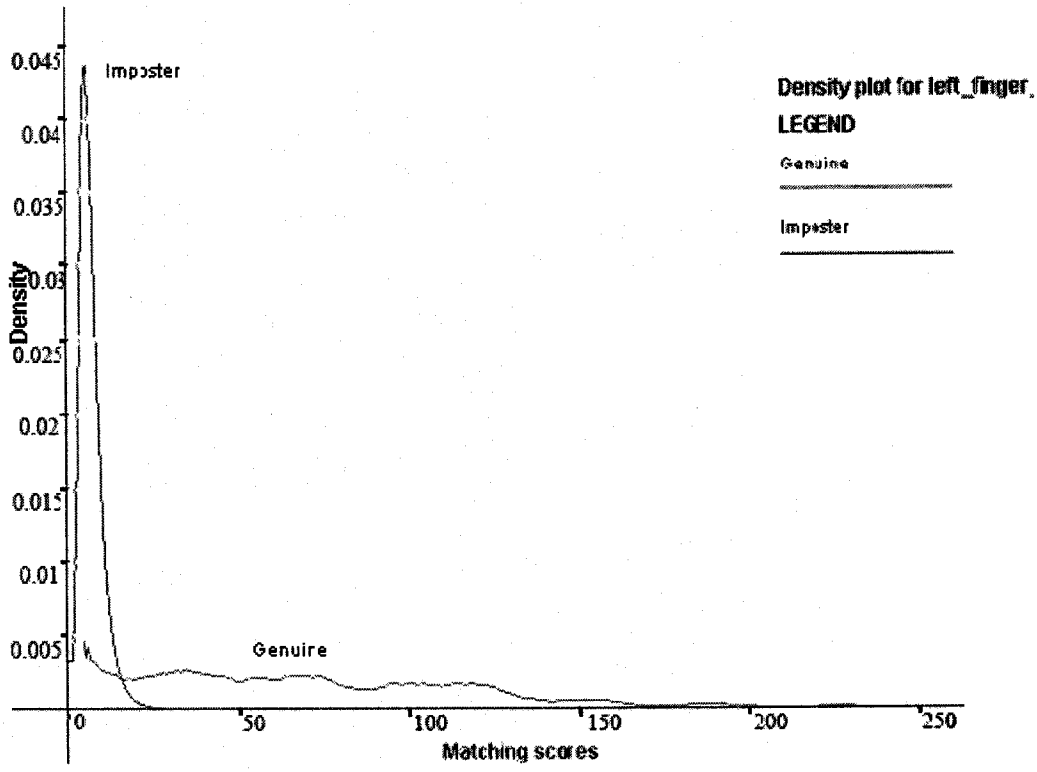


Figure 5. 7 Probability Density Function plot for Left Index Finger

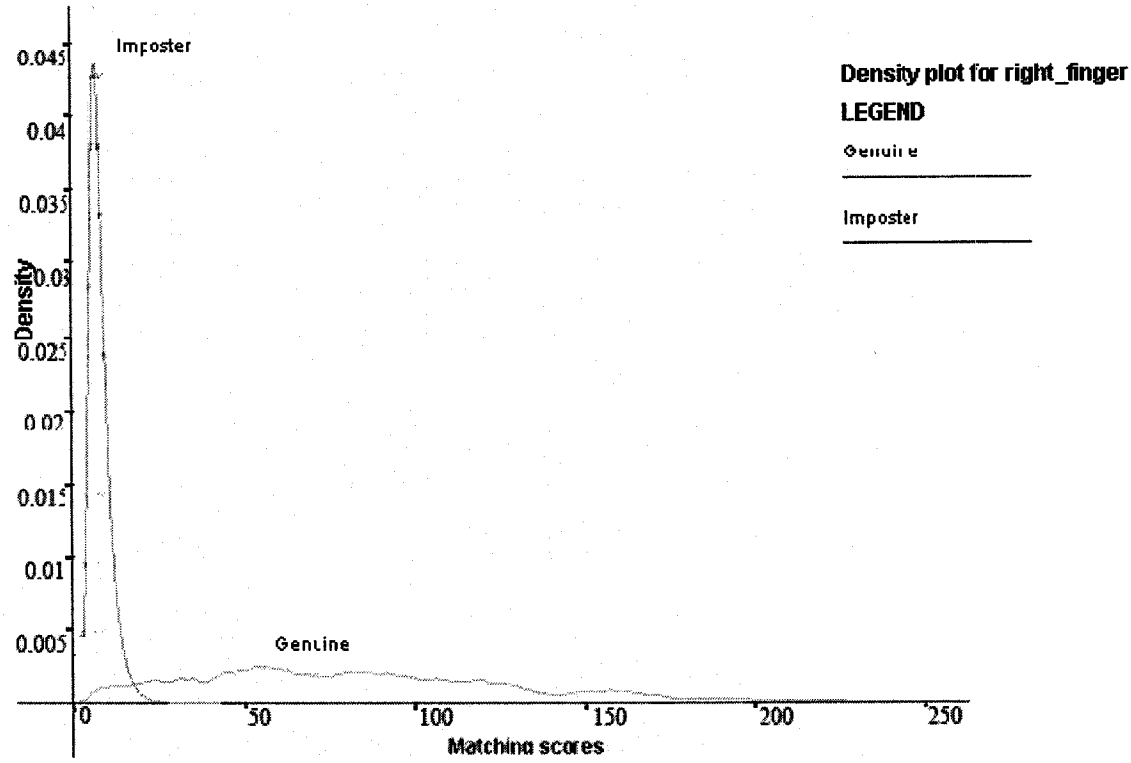


Figure 5. 8 Probability Density Function plot for Right Index Finger

The scores for each modality, however, have been deviated after the employment of Gaussian noise model. The ranges of deviation vary depending on the specified deviation rates. Table 6 presents a comparison about the changes of genuine score range, impostor score range, Minimum and Maximum values of matching score for a modality at different rates of Gaussian deviation.

Modality	Gaussian Deviation	Genuine Score Range	Impostor Score Range	Minimum Value	Maximum Value
Face C	Original	[-1, 0.898]	[-1, 0.732]	-1	0.898
	Within 1%	[-0.996, 0.893]	[-1.005, 0.732]	-1.005	0.894
	Within 5%	[-1.009, 0.912]	[-1.022, 0.739]	-1.022	0.912
	Within 10%	[-0.974, 0.897]	[-1.044, 0.729]	-1.044	0.897
Face G	Original	[64.806, 83.494]	[54.835, 76.482]	54.835	83.494
	Within 1%	[64.867, 83.594]	[54.998, 76.581]	54.998	83.594
	Within 5%	[64.911, 83.571]	[54.954, 76.896]	54.954	83.571
	Within 10%	[64.337, 84.863]	[53.724, 77.868]	53.724	84.863
Left Index Finger	Original	[4.0, 246.0]	[0, 45.0]	0	246
	Within 1%	[3.968, 246.515]	[-1.197, 44.481]	-1.197	246.515
	Within 5%	[2.240, 247.052]	[-5.665, 45.010]	-5.665	247.052
	Within 10%	[0.210, 244.051]	[-10.622, 49.342]	-10.622	244.051
Right Index Finger	Original	[0, 257.0]	[0, 43.0]	0	257
	Within 1%	[0.073, 256.994]	[-1.180, 42.535]	-1.18	256.994
	Within 5%	[1.995, 256.991]	[-6.074, 44.055]	-6.074	256.991
	Within 10%	[0.278, 253.294]	[-12.357, 40.204]	-12.357	253.294

Table 6 Comparison of matching scores of four modalities at different deviation rates

The distributions of probability densities of genuine scores and impostor scores also have been changed accordingly. Figure 5.9 illustrates the deviation of curves of probability densities for genuine scores and impostor scores for modality Right Index Finger at 10% deviation rate compared to the original.

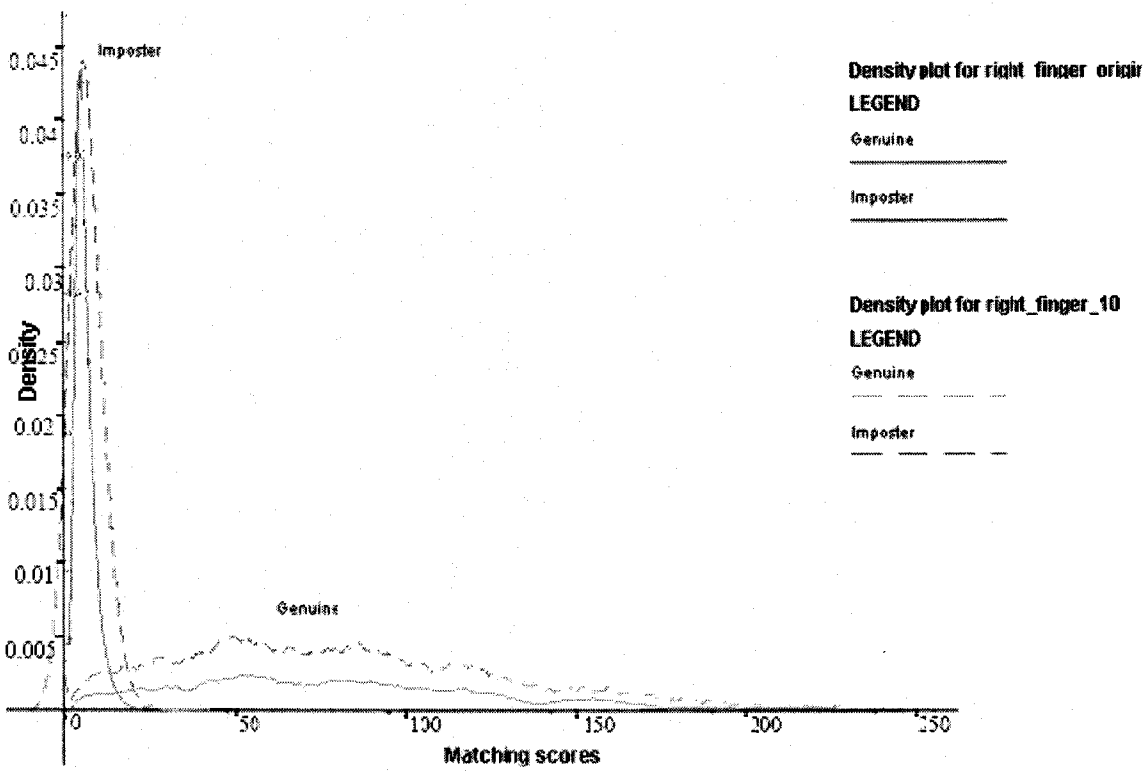


Figure 5. 9Deviation of density distribution of Right Index Finger

5.3.2 Performance analysis

After the generation of different matching scores datasets based on Gaussian noise model, a series of performance analysis have been conducted in different aspects.

1) Performance of unimodal and multimodal biometrics system.

The multimodal biometrics system can achieve much better performance than any unimodal biometric system. Figure 5.10 shows the ROC curves of four single modalities and the ROC curve after the fusion of them (Simple Product).

2) Performance of the same normalization and fusion methods based on different partitioning methods.

Using different partitioning methods makes performance variation as shown in Figure 5.11 which displays two ROC curves by the same normalization and fusion (BGI) based on different partitioning methods (Re-substitution vs. Hold-out).

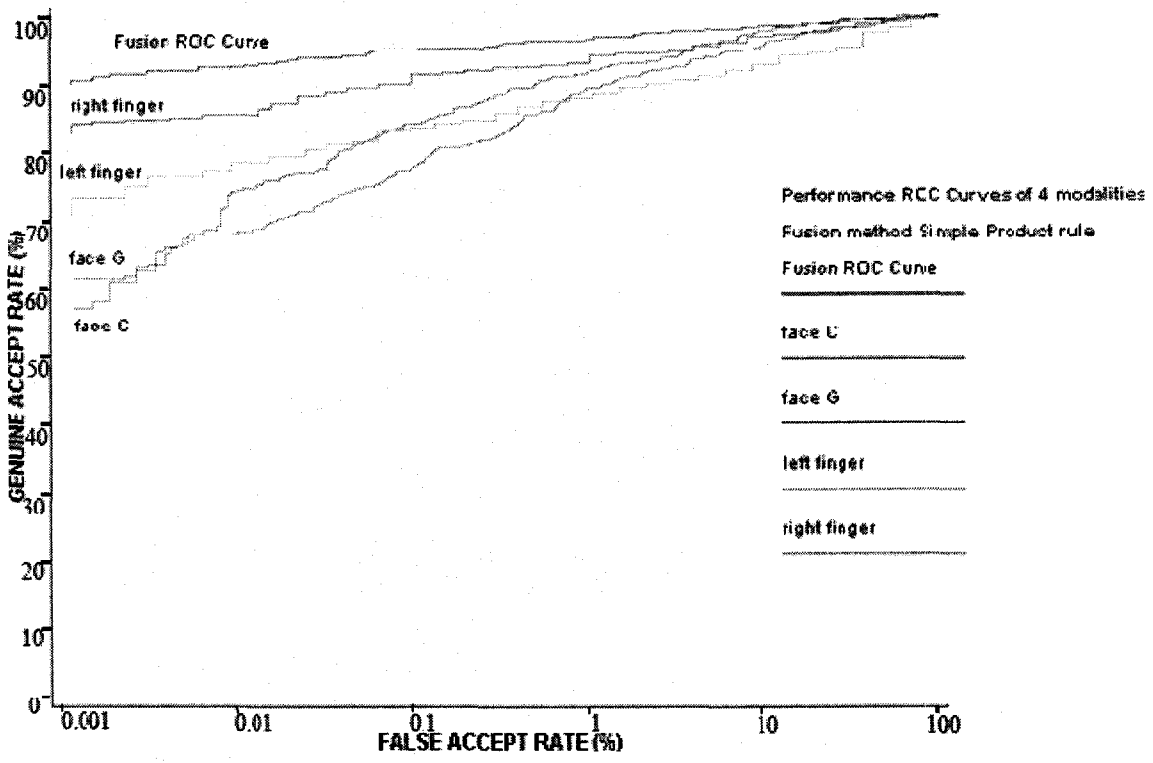


Figure 5. 10 Performance of multimodal and unimodal biometrics systems

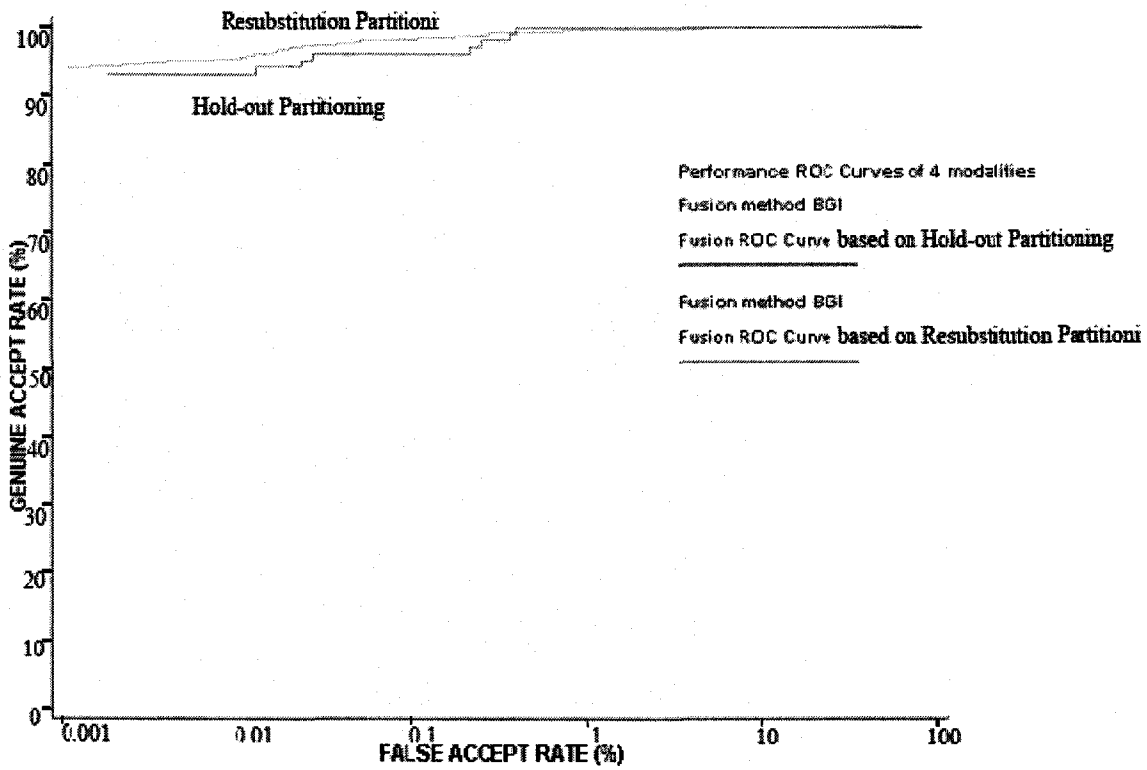


Figure 5. 11Performance of different partitioning methods

3) Performance of different normalization methods followed by the same fusion method and based on the same partitioning method.

Using different normalization methods makes performance variation as shown in Figure 5.12 which displays two ROC curves by Simple Sum fusion method and different normalizations (Min-Max vs. Zscore) based on resubstitution partitioning.

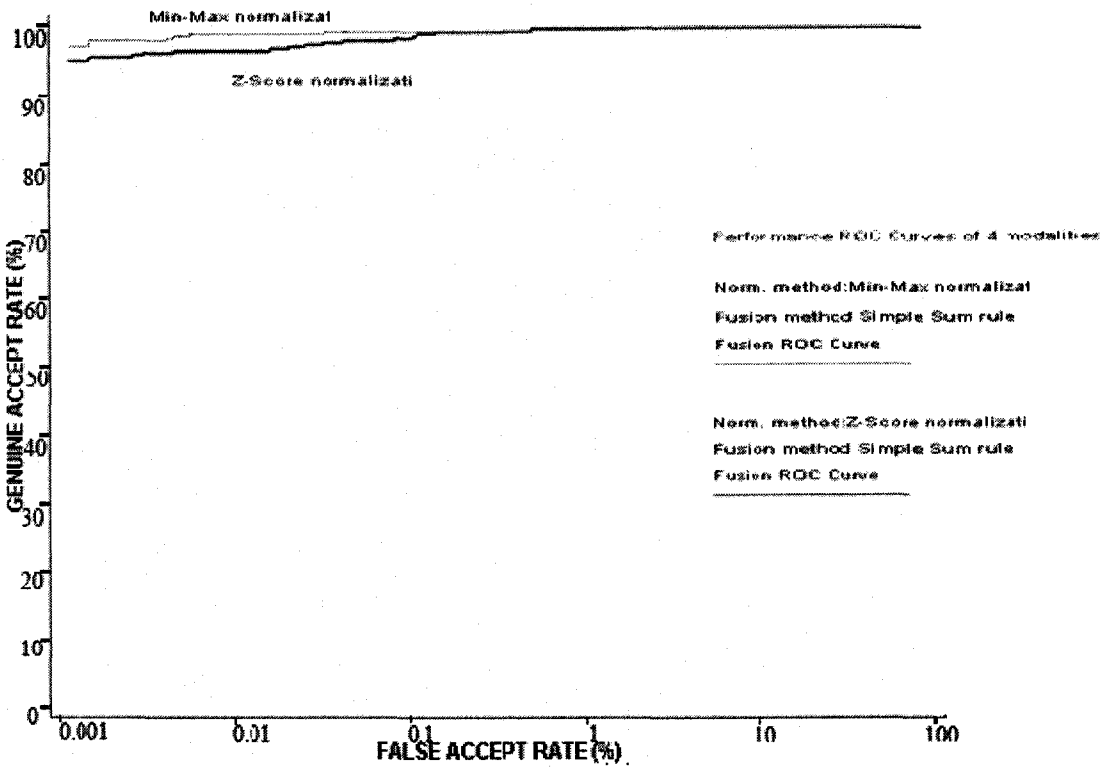


Figure 5. 12 Performance of different normalization methods

4) Performance of different fusion methods after the same normalization and partitioning methods.

Using different fusion methods makes performance variation as shown in Figure 5.13 which displays two ROC curves by different fusions (Simple Minimum vs. Simple Product) after Min-Max normalization and based on Hold-out partitioning.

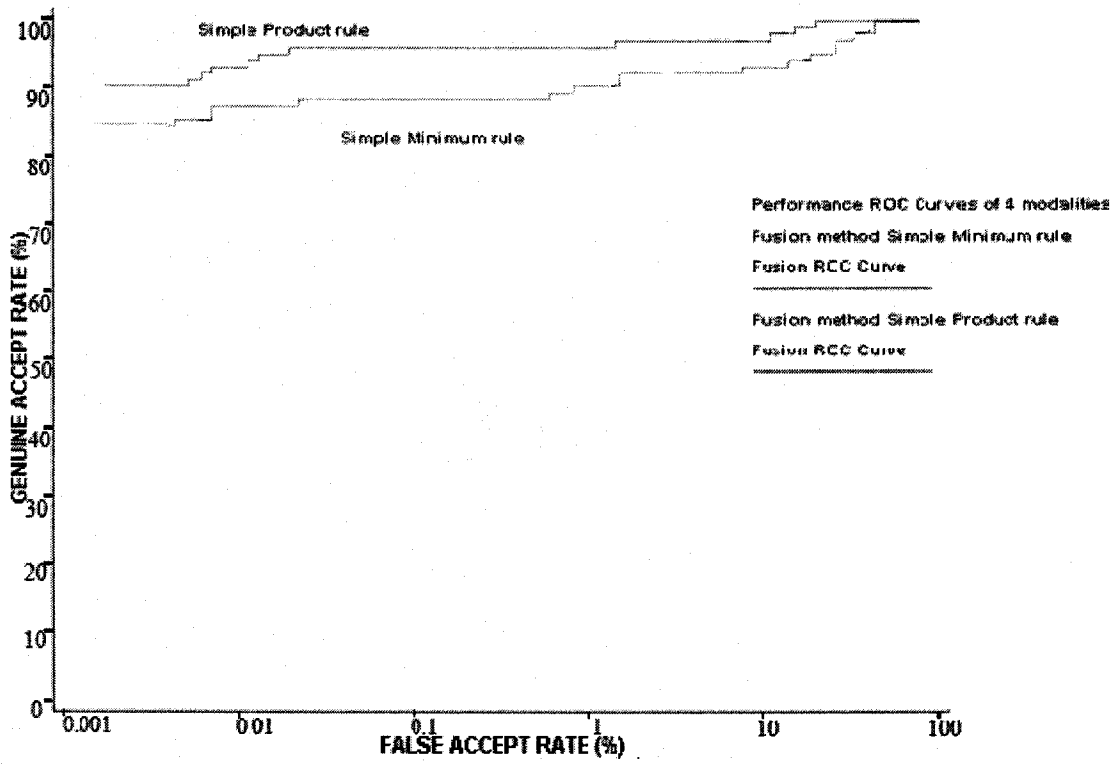


Figure 5. 13 Performance of different fusion methods

5.3.3 Evluation matrix and Robust design

After conducting 126 experiments of different combinations of control factors and noise factors, we collected the experiments results and now have the evaluation matrix filled up as shown in Table 7. The mean value of each row and related Signal to Noise (S/N) ratio were also computed.

By investigating the results, it is concluded that the first combination of control factors is the overall winner for the most robust design. In other words, the application of re-substitution partitioning, Min-Max normalization and Simple Sum fusion techniques on the multimodal biometrics system with the four specific modalities has been proved to have the best GAR performance at 0.1% FAR under the inevitable noise disturbances.

				1	2	3	4	5	6	7	8	9	No.	S/N Ratio (dB)
				1	1	1	2	2	2	3	3	3	a	
				1	2	3	1	2	3	1	2	3	b	
				1	2	3	2	3	1	3	1	2	c	
No.	A	B	C	1	2	3	3	1	2	2	3	1	d/Mean	
1	1	1	1	99.424	99.135	99.28	99.247	99.28	99.247	99.28	99.548	99.28	99.302	39.939
2	1	2	2	95.101	94.236	92.807	93.524	92.939	94.428	93.516	94.277	93.948	93.864	39.449
3	1	3	3	84.15	84.582	82.709	84.036	84.006	84.639	84.15	82.229	83.573	83.786	38.462
4	1	4	4	97.983	99.28	99.135	98.795	96.974	96.235	98.847	97.44	95.101	97.754	39.800
5	2	1	2	97.152	97.152	97.152	97.152	98.101	98.101	98.101	92.247	97.152	96.923	39.724
6	2	2	3	88.449	89.399	91.297	84.494	87.5	90.348	86.551	81.646	85.443	87.236	38.799
7	2	3	4	99.051	99.051	99.051	95.253	98.101	99.051	99.051	95.253	99.051	98.101	39.830
8	2	5	1	99.842	98.101	99.051	98.101	98.101	99.842	97.152	98.101	99.051	98.594	39.876
9	3	1	3	86.167	85.443	84.652	84.652	86.551	84.968	86.709	83.386	86.551	85.453	38.633
10	3	2	4	78.481	76.741	71.994	78.165	77.215	71.519	79.114	76.108	71.044	75.598	37.549
11	3	4	1	99.367	99.525	98.892	98.576	99.367	99.367	99.367	99.367	98.892	99.191	39.929
12	3	5	2	99.051	98.559	98.271	98.271	98.559	98.703	98.559	98.559	97.983	98.502	39.869
13	1	*	5	98.559	99.135	98.559	98.795	98.271	98.795	98.559	98.494	99.28	98.716	39.888
14	2	*	5	98.101	99.841	98.101	99.051	99.842	99.051	99.842	98.101	97.152	98.787	39.893

Table 7 Evaluation Matrix

A1--A3: Resubsitution / Hold-out/ Leave-one-out
B1--B5: Min-Max / Decimal Scaling / Z-Score / Median and MAD / Tanh-Estimators
C1--C5: Simple Sum/ Simple Product/ Simple Minimum / Simple Maximum / BGI
a1--a3: 1% / 5% /10% (so does b1--b3, c1--c3, d1--d3)

5.4 DISCUSSION

By analyzing the experimental data in the evaluation matrix (Table 7) we found some control factors combinations have consistent performances in different noise conditions. For example, the performance of the combination No.1 varies between 99.135% and 99.548%; the performance of the combination No.11 varies between 98.576% and 99.525%. However, other control factors combinations have significant changes when noise factors differ. For example, the minimum performance value for combination No. 6 is 81.646% while maximum value is 91.297%. The performance of combination No. 10 also jumps from 71.044% to 79.144%. It is therefore that the performance analysis of multimodal biometrics system can not be made just upon a certain condition; the most robust selection should be chosen after investigating all possible conditions by using the proposed systematic approach.

In addition, the experimental results do not mean the losers can not perform as good as the chosen winner all the time. They may even perform better when the operating point changes in other applications. Because the performance metric GAR is related to a specific FAR greatly. When the requirement for the FAR of the application changes (i.e. operating point changes), the GAR value will change accordingly. Take control factors combination No. 3 for example, the mean value of GAR at 0.1% FAR is only 83.786%. But when the FAR value rises, the GAR hikes rapidly as we can see in Figure 5.14. This combination may be adopted by the applications (e.g forensic application) which prefer high FAR when taking other elements into account as we have discussed in section 3.2.2.

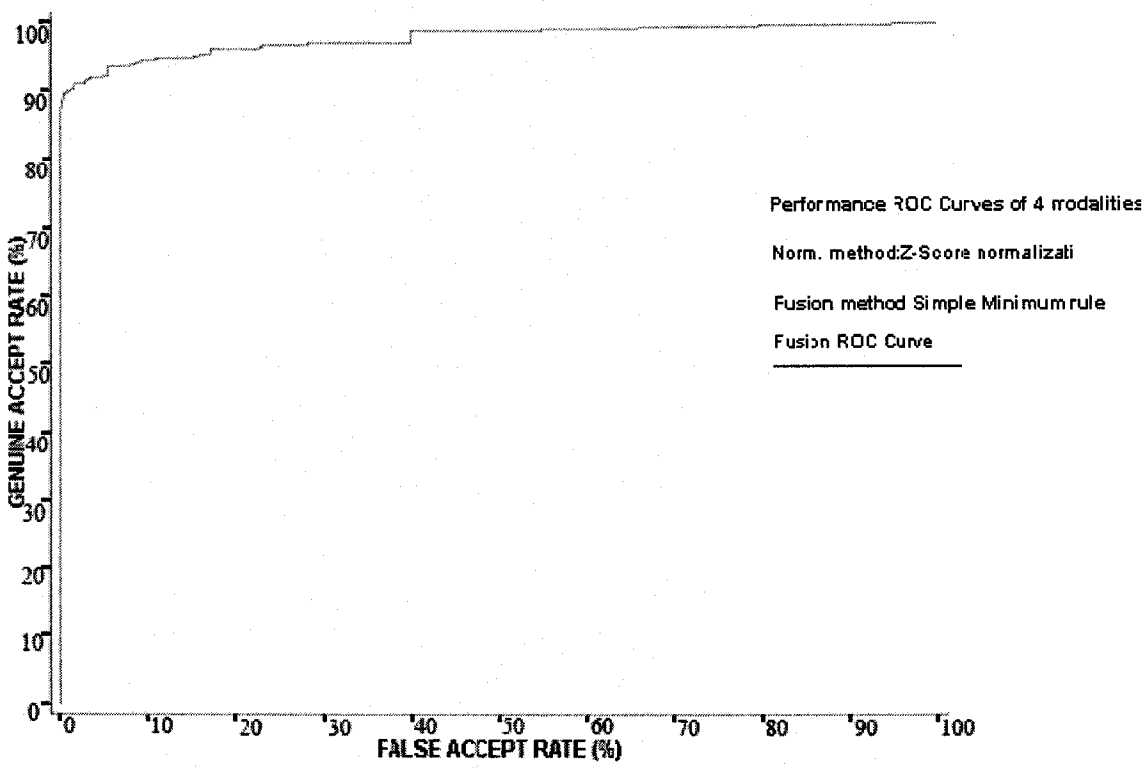


Figure 5. 14 GAR changes with the FAR value

Another issue worth noting is that equipped with Orthogonal Array (OA) technique, the proposed methodology can achieve more efficiency when the number of parameters or levels increases. Table 8 has illustrated the numbers of full factorial experiments needed for different parameters (either control factor or noise factor) and levels among the common scenarios for multimodal biometrics systems. Table 9 has derived the corresponding numbers of experiments needed by using Orthogonal Array technique. We can see from Table 10 that the percentage of experiments by OA divided by corresponding full factorial experiments slashes when the number of parameters or levels increases.

			3	3	3	4	4	4	5	5	5	noise factor parameters
			3	4	5	3	4	5	3	4	5	levels
control factor parameters	levels	full factorial	27	64	125	81	256	625	243	1024	3125	full factorial
3	3	27	729	1728	3375	2187	6912	16875	6561	27648	84375	
3	4	64	1728	4096	8000	5184	16384	40000	15552	65536	200000	
3	5	125	3375	8000	15625	10125	32000	78125	30375	128000	390625	
4	3	81	2187	5184	10125	6561	20736	50625	19683	82944	253125	
4	4	256	6912	16384	32000	20736	65536	160000	62208	262144	800000	
4	5	625	16875	40000	78125	50625	160000	390625	151875	640000	1953125	
5	3	243	6561	15552	30375	19683	62208	151875	59049	248832	759375	
5	4	1024	27648	65536	128000	82944	262144	640000	248832	1048576	3200000	
5	5	3125	84375	200000	390625	253125	800000	1953125	759375	3200000	9765625	

Table 8 Number of experiments based on full factorial experiment

			3	3	3	4	4	4	5	5	5	noise factor parameters
			3	4	5	3	4	5	3	4	5	levels
control factor parameters	levels	orthogonal arrays	9	16	25	27	16	25	27	16	25	orthogonal arrays
3	3	9	81	144	225	243	144	225	243	144	225	
3	4	16	144	256	400	432	256	400	432	256	400	
3	5	25	225	400	625	675	400	625	675	400	625	
4	3	27	243	432	675	729	432	675	729	432	675	
4	4	16	144	256	400	432	256	400	432	256	400	
4	5	25	225	400	625	675	400	625	675	400	625	
5	3	27	243	432	675	729	432	675	729	432	675	
5	4	16	144	256	400	432	256	400	432	256	400	
5	5	25	225	400	625	675	400	625	675	400	625	

Table 9 Number of experiments based on orthogonal array technique

		3	3	3	4	4	4	5	5	5	noise factor parameters
control factor parameters	levels	3	4	5	3	4	5	3	4	5	levels
3	3	11.111%	8.333%	6.667%	11.111%	2.083%	1.333%	3.704%	0.521%	0.267%	
3	4	8.333%	6.250%	5.000%	8.333%	1.563%	1.000%	2.778%	0.391%	0.200%	
3	5	6.667%	5.000%	4.000%	6.667%	1.250%	0.800%	2.222%	0.313%	0.160%	
4	3	11.111%	8.333%	6.667%	11.111%	2.083%	1.333%	3.704%	0.521%	0.267%	
4	4	2.083%	1.563%	1.250%	2.083%	0.391%	0.250%	0.694%	0.098%	0.050%	
4	5	1.333%	1.000%	0.800%	1.333%	0.250%	0.160%	0.444%	0.063%	0.032%	
5	3	3.704%	2.778%	2.222%	3.704%	0.694%	0.444%	1.235%	0.174%	0.089%	
5	4	0.521%	0.391%	0.313%	0.521%	0.098%	0.063%	0.174%	0.024%	0.013%	
5	5	0.267%	0.200%	0.160%	0.267%	0.050%	0.032%	0.089%	0.013%	0.006%	

Table 10 OA efficiency

CHAPTER 6 CONCLUSIONS AND FUTURE WORK

6.1 CONTRIBUTIOINS OF THE RESEARCH

This thesis developed a statistical approach for performance analysis of multimodal biometric systems, and presented experiment results with the four modality BSSR1 dataset. The statistical approach has made it possible to systematically study the performance of different fusion methods and normalization techniques in the presence of noise. In addition, observations made from this study not only identified the fusion methods best in performance, but also produced useful guidance for the practice of system performance analysis. Although only four out of many types of possible biometric information are used in experiment, the method is general and can be applied to applications that require large volume, high-dimensional experiments.

6.2 DIRECTIONS OF FUTURE WORK

There are two directions of future work we may pursue to work towards:

- Expanding the application of the proposed methodology in a broader spectrum

The factors we have studied by using the proposed methodology in this thesis is only a subset of factors influencing the performance of multimodal biometrics system. (Mansfield, 2002) has analyzed and summed up the factors that could possibly influence the system performance as we have discussed in section 3.3.1:

1. Population demographics (e.g. Age, Gender, Ethnic Origin etc)
2. Application (e.g. User familiarity, User Motivation etc)
3. User physiology (e.g. Beards& Moustaches, Disability, Height etc)
4. User behaviour (e.g. Facial expression, Movement, Pose etc)
5. User appearance (e.g. Contact lens, Hair style, Tattoo etc)
6. Environmental influences (e.g. Background, Lighting, Weather etc)
7. Sensor and hardware (e.g. Sensor quality, Transmission channel etc)
8. User interface (e.g. Feedback, Instruction, Supervision etc)

In the future, in a more elaborate and effective manner, we may identify the above factors as control factors and noise factors respectively and select the influential factors in the context of different biometrics applications to construct the analysis models for different performance metrics based on variety of scenarios.

- Conducting our experiments on a larger database

As we have discussed, the performance metrics of a biometric system such as accuracy, throughput, and scalability can be estimated with a high degree of confidence only when the system is tested on a large representative database. Due to the limited conditions we have employed the largest database in public for our experiments; however, it is still small compared to other proprietary databases. For example, face (Phillips, 2003) and fingerprint (Wilson, 2004) recognition systems have been evaluated on large databases (including samples from more than 25,000 subjects) acquired from a diverse population under the changing environmental conditions.

Our future experiments may employ a larger database which can be better representative of the population and each biometric trait can preferably exhibit realistic intra-class variations by collecting data over multiple sessions spread over a period of time and in different environmental conditions).

REFERENCES

(Blackburn, 2000)

D.M.Blackburn, M.Bone, and P.J.Philips. FRVT 2000: *Facial recognition vendor test. Technical report*. Dod Counterdrug Technology Development Office, Defence Advance Research Project Agency, National Institute of Justice, Dahlgren, VA; Crane, IN; Arlington, VA, December 2000.

(Bolle, 2000)

R.M. Bolle, S. Pankanti, and N.K. Ratha. *Evaluation techniques for biometrics-based authentication systems (FRR)*. Pattern Recognitioin proceedings, pages: 831-837, 2000.

(Bolle, 2003)

R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, *Guide to Biometrics*, Springer, 2003

(Duda, 2001)

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, 2001.

(Garcia-Salicetti, 2005)

Garcia-Salicetti, S., Mellakh, M.A., Allano, L., and Dorizzi, B. A Generic. *Protocol for Multibiometric Systems Evaluation on Virtual and Real Subjects*. In Fifth International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA), pages 494-502, Rye Brook, USA. 2005

(Germain, 1999)

L.Germain. *Large scale systems.* In A.K.Jain, R.M.Bolle, and S.Pankanti, editors. Biometrics: Personal Identification in Networked Society, pages 311-326. Kluwer Academic Press, Boston, MA, 1999.

(Givens, 2004)
G. Givens, J.R. Beveridge, B.A. Draper, and D. Bolme: *A Statistical Assessment of Subject Factors in the PCA Recognition of Human Subjects.* In: Proceedings of CVPR Workshop: Statistical Analysis in Computer Vision. Vol. 8, pages: 96-104, 2003

(Hampel, 1986)
F. R. Hampel, P. J. Rousseeuw, E. M. Ronchetti, and W. A. Stahel, *The Approach Based on Influence Functions*, Robust Statistics. John Wiley & Sons, 1986.

(Hedayat, 1999)
A. Hedayat, John Stufken, and N.J.A.Sloane. *Orthogonal Arrays:Theory and Applications.* Springer Verlag, 1999.

(Hong, 1999)
L. Hong, A. Jain and S. Pankanti. *Can Multibiometrics Improve performance,* Proceedings AutoID'99, Pages.59-64.Summit, NJ, Oct 1999,

(Huber, 1981)
P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

(Indovina, 2003)
M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. K. Jain. *Multimodal Biometric Authentication Methods: A COTS Approach.* In Proceedings of Workshop on Multimodal User Authentication (MMUA), pages 99-106, Santa Barbara, USA, 2003.

(iSixSigma, 2002)
iSixSigma website, "The introduction to Robust Design", 2002. Available at http://www.isixsigma.com/library/content/c020311a.asp

(Jain, 2002)

A. K. Jain and A. Ross. *Learning User-specific Parameters in a Multibiometric System.* In Proceedings of International Conference on Image Processing, pages 57–60, New York, USA, September 2002.

(Jain, 2004a)
A. K. Jain and A. Ross, *Multibiometric Systems*, Communications of the ACM, Special Issue on Multimodal Interfaces, Vol. 47, No. 1, pages. 34-40, January 2004.

(Jain, 2004b)
A. K. Jain, A. Ross, and S. Prabhakar, *An Introduction to Biometric Recognition*, IEEE Transactions On Circuits And Systems For Video Technology, vol. 14, no. 1, pages. 4–21, January 2004.

(Jain, 2005)
A. Jain, K. Nandakumar, and A. Ross, *Score normalization in Multimodal biometric systems.* Proc. of Audio- and Video-based Biometric Person Authentication (AVBPA) 2005, pages. 1049-1058, Rye Brook, NY, July 2005.

(Kittler, 1998)
J. Kittler, M. Hatef, R. P. Duin, and J. G. Matas. *On Combining Classifiers.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3):226–239, March 1998.

(Lam, 1995)
L. Lam and C. Y. Suen. *Optimal Combination of Pattern Classifiers.* Pattern Recognition Letters, Vol.16, pages: 945–954, 1995.

(Lam, 1997)
L. Lam and C. Y. Suen. *Application of Majority Voting to Pattern Recognition*: An Analysis of Its Behavior and Performance. *IEEE* Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, Vol.27, No.5, pages: 553–568, 1997.

(Llado, 2002)
C.M. Llado, J. Luthi, *Studying sensitivities of an EJB performance model*, Modeling, Analysis and Simulation of Computer and Telecommunications Systems (MASCOTS 2002), Proceedings. 10th IEEE International Symposium, Pages: 277 – 280, Oct. 2002.

(Maio, 2004)
D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman, and A. K. Jain. *FVC2004: Third Fingerprint Verification Competition.* In Proceedings of International Conference on Biometric Authentication (ICBA), pages: 1-7, Hong Kong, China, 2004

(Mansfield, 2002)
A. J. Mansfield, and J. L. Wayman. *Best Practices in Testing and Reporting Performance of Biometric Devices,* Version 2.01. Technical Report NPL Report CMSC 14/02, National Physical Laboratory, 2002.

(Mitra, 2007)
S. Mitra, M. Savvides, and A. Brockwell. *Statistical Performance Evaluation of Biometric Authentication Systems Using Random Effects Models.* Pattern Analysis and Machine Intelligence, IEEE Transactions Vo. 29, No.4, pages: 517-530. April, 2007

(Nandakumar, 2005)
Karthik Nandakumar, *Integration of Multiple Cues in Biometric Systems*, Master's Thesis, Michigan State University, 2005.

(NIST, 2000)
NIST report to the United State Congress. Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability, November 13, 2000.

(NIST, 2004)

National Institute of Standards and Technology: *NIST Biometric Scores Set Release I. (2004)*. Available at http://www.itl.nist.gov/iad/894.03/biometricscores

(Philips, 1996)
P.J. Phillips, P.J. Rauss, and S. Der. *FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results*, Army Research Laboratory technical report, ARL-TR-995. 1996 Available at http://www.frvt.org.

(Philips, 2000)
P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki. *An Introduction Evaluating Biometric Systems*. IEEE Computer, Vol.33, No.2, pages: 56-63, 2000

(Philips, 2003)
J. Phillips, et al. *Face Recognition Vendor Test 2002*: Evaluation Report. NISTIR 6965, March 2003. Available at http://www.frvt.org.

(Poh, 2005)
N. Poh, and S. Bengio. *Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments?* In Second International Machine Learning and Multimodal Interaction Workshop (MLMI), Edinburgh, UK. 2005.

(Ross, 1995)
P. Ross. Taguchi Techniques for Quality Engineering. McGraw-Hill, 1995.

(Ross, 2003)
A. Ross and A. K. Jain, *Information Fusion in Biometrics,* Pattern Recognition Letters, Vol. 24, Issue 13, pages: 2115-2125, September 2003.

(Ross, 2004)

A. Ross, A. K. Jain, *Multimodal Biometrics: An Overview*, Proceedings of 12th European Signal Processing Conference, pages: 1221-1224, 2004

(Ross, 2006)
A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*, Springer; 1 edition (May 24, 2006).

(Samoska, 2006)
N. Samoska, Evaluation and performance prediction of multimodal biometric systems, Master's Thesis, West Virginia University, 2006

(Sanderson, 2002)
C. Sanderson and K. K. Paliwal. *Information Fusion and Person Verification using speech and face information.* Research Paper IDIAP-RR 02-33, IDIAP, September 2002.

(Sanderson, 2003)
C. Sanderson, and K. K. Paliwal. *Noise compensation in a person verification system using face and multiple speech features*. Pattern Recognition. Vol. 36, pages: 293-302. 2003

(Sedgwick, 2004)
N. Sedgwick, *The Need for Standardization of Multi-Modal Biometric Combination,* Business/technical presentation,Cambridge Algorithmica Limited, 2004 Available at http://www.camalg.co.uk/s03017_pr0/pr0_040216a.pdf

(Snelick, 2005)
R. Snelick, M. Indovina, J. Yen, A. Mink, *Multimodal Biometrics: Issues in Design and Testing*, National Institute of Standards and Technology, ICMI'03, Pages: 68 - 72 Vancouver, British Columbia, Canada. November 5–7, 2003.

(Ulery, 2006)

B. Ulery, A. Hicklin, C. Watson, W. Fellner, and P. Hallinan. *Studies of Biometric Fusion.* Technical Report 7346, National Institute of Standards and Technology (NIST). 2006.

(Verlinde, 1999)
P. Verlinde, P. Druyts, G. Cholet, and M. Acheroy. *Applying Bayes based Classifiers for Decision Fusion in a Multi-modal Identity Verification System.* In Proceedings of International Symposium on Pattern Recognition "In Memoriam Pierre Devijver", Brussels, Belgium, February 1999.

(Wang, 2006)
P. Wang and Q. Ji. *Performance Modeling and Prediction of Face Recognition Systems.* Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2, pages: 1566-1573. Washington, DC, USA. 2006

(Wikipedia, 2007)
Wikipedia website, "US-VISIT (United States Visitor and Immigrant Status Indicator Technology)", 2007. Available at http://en.wikipedia.org/wiki/US-VISIT_(United_States_Visitor_and_Immigrant_Status_Indicator_Technology)

(Wilson, 2004a)
E. Wilson, C. Tabassi, and Watson. *Fingerprint Image Quality. Technical Report 7151,* National Institute of Standards and Technology (NIST). 2004

(Wilson, 2004b)
C.L. Wilson, M.D.Garris, and C.I. Watson. *Matching Performance for the US-VISIT IDENT System Using Flat Fingerprints.* Technical Report 7110. National Institute of Standards and Technology (NIST). 2004

VITA AUCTORIS

NAME:	Wei Gan
PLACE OF BIRTH:	Changsha, Hunan, China
YEAR OF BIRTH:	1976
EDUCATION:	Hunan University, Hunan, China
	1994 – 1998 B.Eng.
	University of Windsor, Windsor, Ontario, Canada
	2005 – 2007 M.Sc.