Multimodal Biometric Authentication Methods: A COTS Approach

M. Indovina[1], U. Uludag[2], R. Snelick[1], A. Mink[1], A. Jain[2]
[1]*National Institute of Standards and Technology,* [2]*Michigan State University*
{mindovina, rsnelick, amink}@nist.gov, {uludagum, jain}@cse.msu.edu

Abstract

We examine the performance of multimodal biometric authentication systems using state-of-the-art Commercial Off-the-Shelf (COTS) fingerprint and face biometrics on a population approaching 1000 individuals. Prior studies of multimodal biometrics have been limited to relatively low accuracy non-COTS systems and populations approximately 10% of this size. Our work is the first to demonstrate that multimodal fingerprint and face biometric systems can achieve significant accuracy gains over either biometric alone, even when using already highly accurate COTS systems on a relatively large-scale population. In addition to examining well-known multimodal methods, we introduce novel methods of fusion and normalization that improve accuracy still further through population analysis.

1. Introduction

It has recently been reported [1] to the U.S. Congress that approximately two percent of the population does not have a legible fingerprint and therefore cannot be enrolled into a fingerprint biometrics system. The report recommends a system employing dual biometrics in a layered approach. Use of multiple biometric indicators for identifying individuals, so-called multimodal biometrics, has been shown to increase accuracy [2, 3, 4], and would decrease vulnerability to spoofing while increasing population coverage.

The key to multimodal biometrics is the fusion (i.e., combination) of the various biometric mode data at the feature extraction, match score, or decision level [4]. Feature level fusion combines feature vectors at the representation level to provide higher dimensional data points when producing the match score. Match score level fusion combines the individual scores from multiple matchers. Decision level fusion combines accept or reject decisions of individual systems.

Our methodology for testing multimodal biometric systems focuses on the match score level [2]. This approach has the advantage of utilizing as much information as possible from each single-mode biometric, while at the same time enabling the integration of proprietary COTS systems.

Published studies examining fusion techniques have been limited to small populations (~100 individuals), while employing low performance non-commercial biometric systems. In this paper we investigate the performance gains achievable by COTS-based multimodal biometric systems using a relatively large (~1000 individuals) population. Section two and three describe the traditional and novel normalization and fusion methods that we employed for match score combination. New methods for *adaptive normalization* and fusion using user-level weighting based on the *wolf-lamb* [5] concept are introduced and compared. In section four we provide a performance analysis of these multimodal methods and investigate performance variability attributable to population differences.

2. Normalization

A normalization step is generally necessary before the raw scores originating from different matchers can be combined in the fusion stage. For example, if one matcher yields scores in the range [100, 1000] and another matcher in the range [0, 1], fusing the scores without any normalization effectively eliminates the contribution of the second matcher. We present three well-known normalization methods, and a 4th novel method, which we call *adaptive normalization* that uses the genuine and impostor distributions.

We denote a raw matcher score as s from the set S of all scores for that matcher, and the corresponding normalized score as n. Different sets are used for different matchers. The abbreviations (such as MM) next to the normalization method names are used throughout the remainder of this paper.

Min-Max (MM). This method maps the raw scores to the [0, 1] range. max(S) and min(S) specify the end

points of the score range (vendors generally provide these values):

$$n = \frac{s - min(S)}{max(S) - min(S)}$$

Z-score (ZS). This method transforms the scores to a distribution with mean of 0 and standard deviation of 1. $mean()$ and $std()$ denote the mean and standard deviation operators:

$$n = \frac{s - mean(S)}{std(S)}$$

Tanh (TH). This method is among the so-called *robust* statistical techniques [6]. It maps the scores to the (0, 1) range:

$$n = \frac{1}{2}\left[tanh\left(0.01\frac{(s - mean(S))}{std(S)} \right) + 1 \right]$$

Adaptive (AD). The errors of individual biometric matchers stem from the overlap of the genuine and impostor distributions as shown in Fig. 1. This region is characterized with its center c and its width w. To decrease the effect of this overlap on the fusion algorithm, we propose to use an adaptive normalization procedure that aims to increase the separation of the genuine and impostor distributions, as indicated by the block arrows in Fig. 1., while still mapping scores to [0,1].

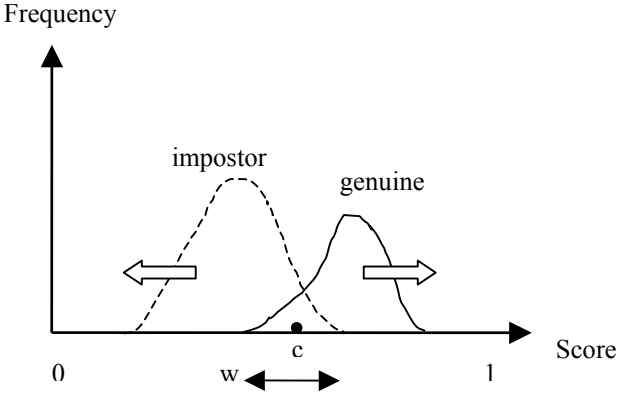

Fig. 1. Overlap of genuine and impostor distributions.

This adaptive normalization is formulated as

$$n_{AD} = f(n_{MM})$$

where $f()$ denotes the mapping function which is used on the MM normalized scores. We have considered the following three functions for $f()$. These functions use two parameters of the overlapped region, c and w, which can be provided by the vendors or estimated by the integrator from data sets appropriate for the specific application. In this work, we act as the integrator.

Two-Quadrics (QQ). This function is composed of 2 quadratic segments that change concavity at c (Fig. 2).

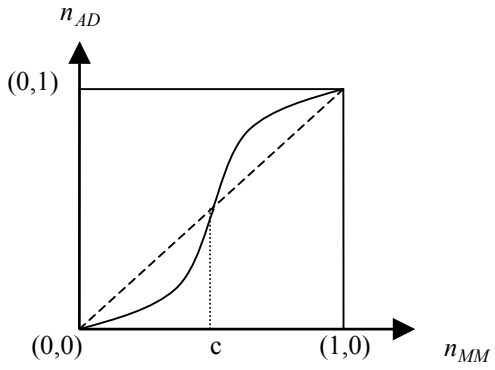

Fig. 2. Mapping function for QQ.

$$n_{AD} = \begin{cases} \frac{1}{c} n_{MM}^2, & n_{MM} \le c \\ c + \sqrt{(1-c)(n_{MM} - c)}, & \text{otherwise} \end{cases}$$

For comparison, note that the identity function, $n_{AD} = n_{MM}$, is shown by the dashed line.

Logistic (LG). Here, $f()$ takes the form of a logistic function. The general shape of the curve is similar to that shown for function QQ in Fig. 2. It is formulated as

$$n_{AD} = \frac{1}{1 + A \cdot e^{-B \cdot n_{MM}}}$$

where the constants A and B are calculated as

$$A = \frac{1}{\Delta} - 1 \quad \text{and} \quad B = \frac{\ln A}{c}$$

Here, $f(0)$ is equal to the constant Δ, which is selected to be a small value (0.01 in this study). Note the inflection point of the logistic function occurs at c, the center of the overlapped region.

Quadric-Line-Quadric (QLQ). The overlapped zone, w, is left unchanged while the other regions are mapped with two quadratic function segments (Fig. 3):

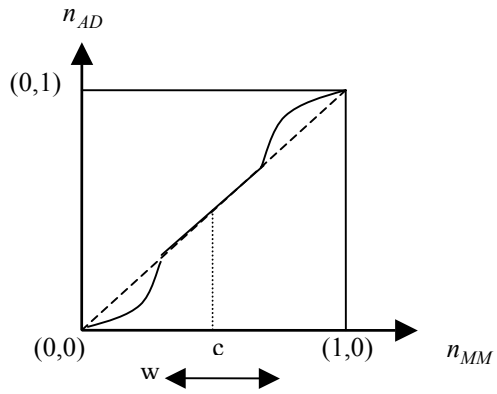

Fig. 3. Mapping function for QLQ.

$$n_{AD} = \begin{cases} \dfrac{1}{(c-\frac{w}{2})} n_{MM}^2, & n_{MM} \le (c-\frac{w}{2}) \\ n_{MM}, & (c-\frac{w}{2}) < n_{MM} \le (c+\frac{w}{2}) \\ (c+\frac{w}{2}) + \sqrt{(1-c-\frac{w}{2})(n_{MM}-c-\frac{w}{2})}, & o/w \end{cases}$$

3. Fusion

We experimented with the five different fusion methods summarized below. The first three are well-known fusion methods; the last two are novel and they utilize the performance of individual matchers in weighting their contributions. As we progress from the first three methods to the fifth, the amount of data necessary to apply the fusion method increases.

Our notation is as follows: n_i^m represents the normalized score for the matcher m ($m = 1, 2, ..., M$, where M is the number of different matchers) and for the user i ($i = 1, 2, ..., I$, where I is the number of individuals in the database). The fused score is denoted as f_i.

Simple Sum (SS). Scores for an individual are summed:

$$f_i = \sum_{m=1}^{M} n_i^m, \forall i$$

Min Score (MIS). Choose the minimum of an individual's scores:

$$f_i = min(n_i^1, n_i^2, ..., n_i^M), \forall i$$

Max Score (MAS). Choose the maximum of an individual's scores:

$$f_i = max(n_i^1, n_i^2, ..., n_i^M), \forall i$$

Matcher Weighting (MW). Matcher weighting-based fusion makes use of the Equal Error Rate (EER). Denote the EER of matcher m as e^m, $m = 1, 2, ..., M$ and the weight w^m associated with a matcher m is calculated as

$$w^m = \frac{\frac{1}{e^m}}{\sum_{m=1}^{M} \frac{1}{e^m}} \qquad (1)$$

Note that $0 \le w^m \le 1, \forall m$, $\sum_{m=1}^{M} w^m = 1$ and the weights are inversely proportional to the corresponding errors; the weights for *more accurate* matchers are higher than those of *less accurate matchers* (Although the EER value alone may not be a good estimator for the accuracy of a matcher, we chose to use it for spanning the amount of data available to the integrator mentioned above). The MW fused score is calculated as

$$f_i = \sum_{m=1}^{M} w^m n_i^m, \forall i$$

User Weighting (UW). The User Weighting fusion method applies weights to individual matchers differently for every user (individual). Previously, Ross and Jain [7] proposed a similar scheme, but they *exhaustively* search a coarse sampling of the weight space, where weights are multiples of 0.1. Their method can be prohibitively expensive if the number of fused matchers, M, is high, since the weight space is $\Re^M$; further, coarse sampling may hinder the calculation of an optimal weight set. In our method, the UW fused score is calculated as

$$f_i = \sum_{m=1}^{M} w_i^m n_i^m, \forall i$$

where w_i^m represents the weight of matcher m for user i.

The calculation of these user-dependent weights make use of the *wolf-lamb* concept introduced by Doddington, et al. [5] for unimodal speech biometrics. They label the users who can be imitated easily as *lambs*; *wolves* on the other hand are those who can successfully imitate some

others. Lambs and wolves decrease the performance of biometric systems since they lead to false accepts.

We extend these notions to multimodal biometrics by developing a metric of *lambness* for every user and matcher, (i,m), pair. This lambness metric is then used to calculate weights for fusion. Thus, if user i is a *lamb* (can be imitated easily by some *wolves)* in the space of matcher m, the weight associated with this matcher is decreased. The main aim is to decrease the lambness of user i in the space of combined matchers.

We assume that for every (i, m) pair, the mean and standard deviation of the associated genuine and impostor distributions are known (or can be calculated, as is done in this study). Denote the means of these distributions as ${}^{gen}\mu_i^m$ and ${}^{imp}\mu_i^m$, respectively, and denote the standard deviations as ${}^{gen}\sigma_i^m$ and ${}^{imp}\sigma_i^m$, respectively.

We use the d-prime metric [8] as a measure of the separation of these two distributions in formulating the lambness metric as:

$$d_i^m = \frac{{}^{gen}\mu_i^m - {}^{imp}\mu_i^m}{\sqrt{({}^{gen}\sigma_i^m)^2 + ({}^{imp}\sigma_i^m)^2}}$$

If d_i^m is small, user i is a lamb for some wolves; if d_i^m is large, i is not a lamb. We structure the user weights to be proportional to this lambness metric as follows

$$w_i^m = \frac{1}{\sum_{m=1}^{M} d_i^m} \cdot d_i^m \qquad (2)$$

Note that $0 \le w_i^m \le 1, \forall i, \forall m$, and $\sum_{m=1}^{M} w_i^m = 1, \forall i$.

Fig. 4 shows the location of potential wolves for a specific (i,m) pair with a block arrow, along with the associated genuine and impostor distributions. This user dependent weighting scheme addresses the issue of matcher-user relationship: namely, a user can be lamb for a specific matcher, but also can be a wolf for some other matcher. We find the user weights by measuring the respective threat of wolves *living* in different matcher spaces for every user.

4. Experimental Results

4.1. Databases

Our experiments were conducted on a population of consistently paired fingerprint and facial images from two groups of 972 individuals, using our previously developed test methodology and framework [2]. Since the paired fingerprint and facial images come from different individuals, we are assuming that they are statistically independent – a widely accepted practice. The images were taken from two separate groups of 972 individuals, with the first group contributing a pair of facial images and the second a pair of fingerprint images. This creates a database of 972 *virtual* individuals. Each pair consists of a primary and a secondary image, with all primary images assigned to the *target* set, and all secondary images assigned to the *query* set.

Match scores were generated from four COTS biometric systems – three fingerprint and one face. For each biometric system, all query set images were matched against all target set images, yielding 972 genuine scores (correct matches) and 943,812 imposter scores.

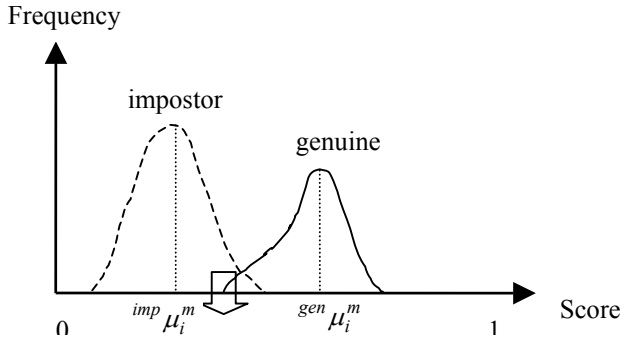

Fig. 4. Distributions for a (user, matcher) pair: the arrow indicates location of wolves for lamb i

4.2. Approach

Among the three adaptive normalization methods (QQ, QLQ and LG), the QLQ method gave the best results in our experiments, so it is selected as the representative method.

We carried out all possible permutations of (normalization, fusion) techniques on our database of 972 users. Table 1 shows the EER values for these permutations. Note that EER values for the 3 individual fingerprint matchers and the face matcher are found to be 3.96%, 3.72%, 2.16% and 3.76%, respectively. The best EER values in individual columns are indicated with **bold** typeface; the best EER values in individual rows are indicated with a star (*) symbol.

Table 1. EER values for permutations (%).

Normalization Technique	*Fusion Technique*				
	SS	MIS	MAS	MW	UW
MM	0.99	5.43	0.86	**1.16**	***0.63**
ZS	*1.71	5.28	1.79	1.72	1.86
TH	1.73	**4.65**	1.82	*1.50	1.62
QLQ	**0.94**	5.43	***0.63**	**1.16**	***0.63**

4.3. Normalization

Figures 5-9 show the effect of each normalization method on system performance for different (but fixed) fusion methods. The ROCs (Receiver Operating Characteristics) for the individual fingerprint matchers and the face matcher are also shown for better comparison.

For UW fusion (Fig. 9), the scatter plot of user weights (Fig. 10) form a distinctive band-like behavior for each fingerprint matcher V1, V2, V3, and the face matcher. The mean user weights for the individual biometric matchers, calculated from (2), are 0.14, 0.64, 0.17 and 0.05, respectively, which implies that on average, fingerprint matcher V2 is the safest for the lambs; whereas the space of the face matcher is filled with wolves (i.e., those waiting to be falsely accepted as one of the lambs). Note that individual matcher performance, shown in the previous ROC curves, is not reflected directly in the set of user weights and their means. Namely, V2 has a higher mean user weight than V3, despite V3's generally better ROC).

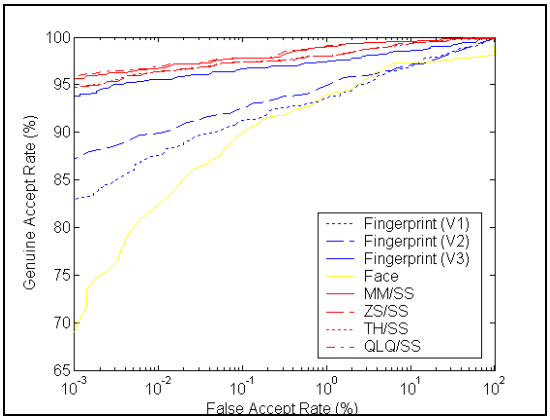

Fig. 5. ROC curves for SS, normalization varied.

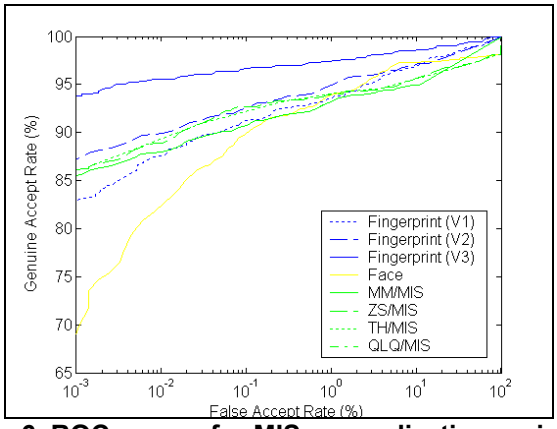

Fig. 6. ROC curves for MIS, normalization varied.

For MW fusion (Fig. 8), the matcher weights, calculated according to (1), are: 0.2, 0.22, 0.37 and 0.21, for the fingerprint matchers and the face matcher, respectively. From Figures 5-9 and Table 1, we see that QLQ and MM normalization methods lead to best performance, except for MIS fusion. Between these two normalization methods, QLQ is better than MM for fusion methods MAS and UW; and about the same as MM for the others.

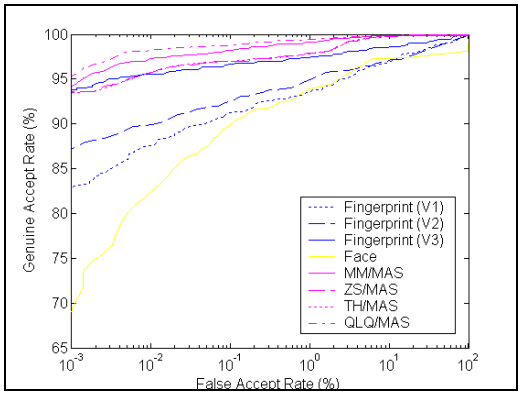

Fig. 7. ROC curves for MAS, normalization varied.

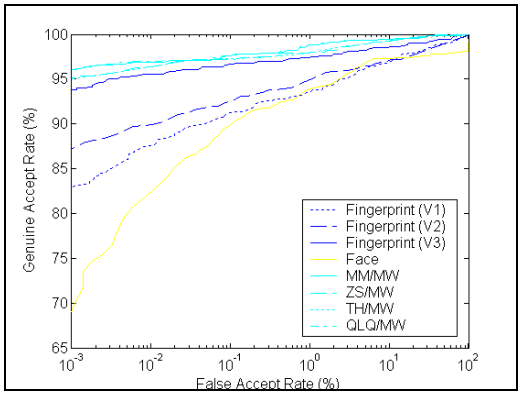

Fig. 8. ROC curves for MW, normalization varied.

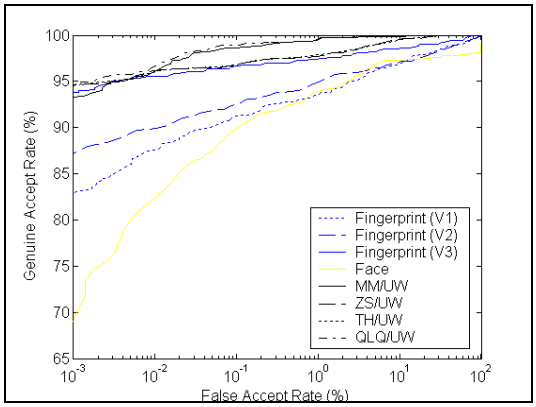

Fig. 9. ROC curves for UW, normalization varied.

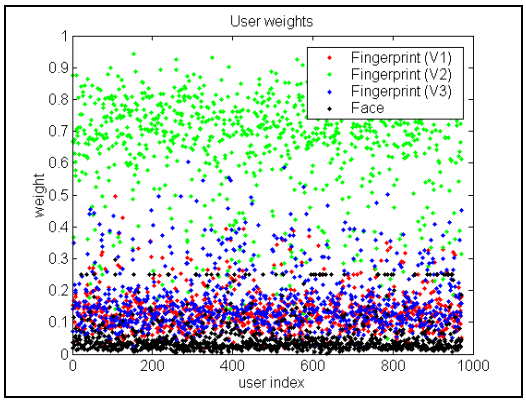

Fig. 10. Pictorial representation of user weights, for QLQ normalization.

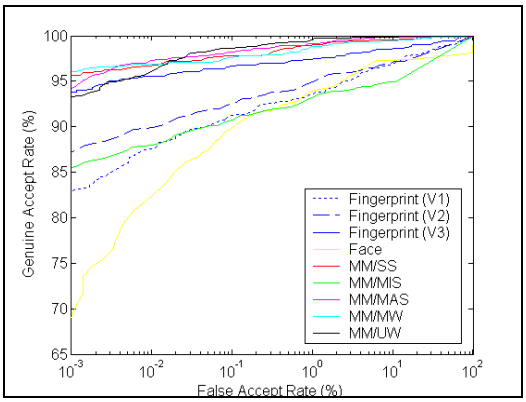

Fig. 11. ROC curves for MM, fusion varied.

4.4. Fusion

Figures 11-14 show the effect of each fusion method on system performance for different (but fixed) normalization methods. The ROCs for the individual fingerprint matchers and the face matcher are also shown for better comparison.

From Figures 11-14 and Table 1, we see that fusion methods SS, MAS and MW generally perform better than the other two (MIS and UW). But for the FAR range of [0.01%, 10%], UW fusion is better than the others. One reason that below 0.01% FAR the performance of UW fusion drops may be the estimation errors become dominant, since we have only one sample available for replacing the individual genuine distributions.

Note that parameter update (for normalization and/or fusion methods) can be employed for addressing the time varying characteristics of the target population. For example, the matcher weights can be updated every time a new set of EER figures are estimated; the user weights can be updated if the fusion system detects changes in the vulnerability of specific users, due to fluctuations in their *lambness*, etc.

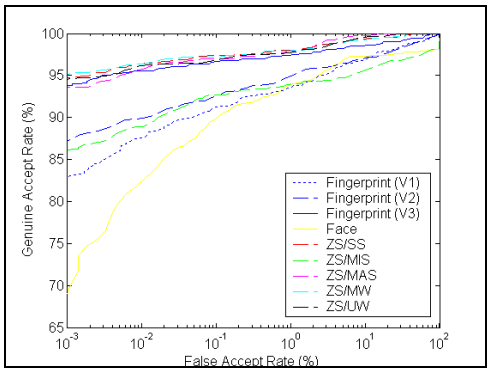

Fig. 12. ROC curves for ZS, fusion varied.

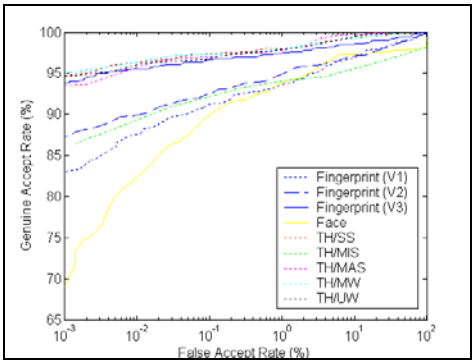

Fig. 13. ROC curves for TH, fusion varied.

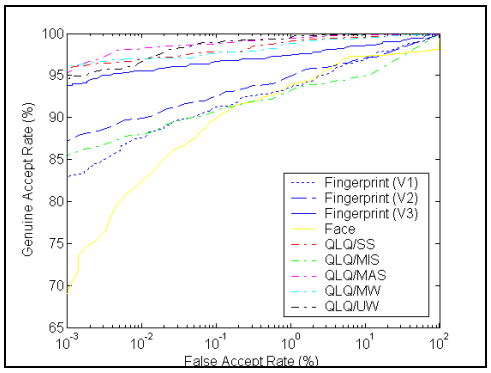

Fig. 14. ROC curves for QLQ, fusion varied.

4.5. Fusing Subsets of Matchers

ROC curves were generated for fusing subsets of the total matcher set. Here, we fixed the normalization method to QLQ and the fusion method to SS.

In Fig. 15 we see that fusing just the three fingerprint matchers (V1V2V3, with EER of 1.94%) is not as good as fusing all the available four matchers (V1, V2, V3 and Face) using QLQ/SS (see Figs. 5 and 14). This implies that even though the face matcher is not as good as any of the individual fingerprint matchers, it still provides complementary information for fusion.

Fusing individual fingerprint matchers separately with the face matcher (V1-Face, V2-Face, V3-Face; with EERs of 1.68%, 1.46% and 2.02%, respectively) we see that V2-Face performs better than V3-Face fusion. Since V3 is the better fingerprint matcher for our dataset, this result may seem counterintuitive. In fact this shows that the face matcher is best complemented with the V2 matcher, i.e., they make independent mistakes; whereas face matcher and V3 matcher make relatively more correlated mistakes.

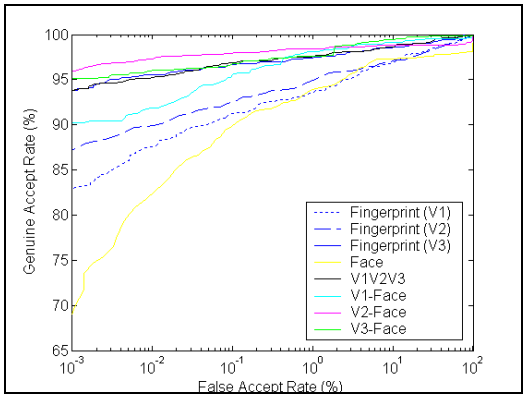

Fig. 15. Fusing subsets of matcher set.

4.6. Peformance Variability

We are interested in determining how the performance of the fused system changes when using (i) an increasingly larger database, (ii) different equal-size databases, and (iii) many randomly assigned virtual subject databases.

Scalability. We created five new user databases from subsets of our original 972 user database: (i) the first 20% of all the users (194 users), (ii) the first 40% of all the users (389 users), (iii) the first 60% of all the users (583 users), (iv) the first 80% of all the users (778 users) and (v) 100% of all the users (972 users). Fig. 16 shows the associated ROC curves for an MM/SS based multimodal system using these datasets. The EERs corresponding to these five sets are 0.42%, 0.75%, 0.67%, 0.8%, and 0.99%, respectively.

We observe that the performance initially drops, but then quickly converges. For this relatively large, but limited, dataset we are unable to draw any general conclusions. It is widely believed that performance decreases as the database size increases. A possible explanation for this belief is that as the state space becomes more populated, differentiation within it, or some clustered areas, becomes more difficult. Another viewpoint is that performance trends cannot be extrapolated to larger populations. Thus a representative database of the intended size may be necessary to predict performance.

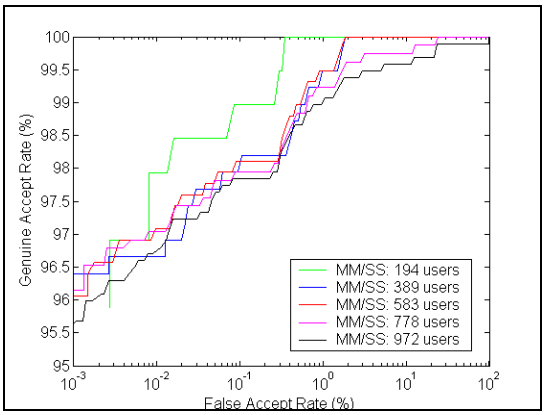

Fig. 16. Scalability: ROC curves for overlapping portions of the whole database.

Generalizability. We created two new user databases of 486 users each from *disjoint* subsets of the original database of 972 users. Fig. 17 shows the associated ROC curves for an MM/SS based multimodal system using these disjoint datasets. The EERs corresponding to these datasets are 0.68% and 1.45%, respectively. We see that the portion of the ROC curves above 0.4% FAR, show a considerable performance difference. Although we can draw no general trends, this implies that its necessary to use a representative database when determining expected performance, but there are presently no clear measurements/methods to determine if a database is representative. Similar results have been reported for performance variation of unimodal systems in [9].

Virtual Subjects. As explained previously, it is common practice to create virtual subjects in multimodal experiments. In our previous experiments, we consistently assigned a "physical finger" to a "physical face" to create a virtual subject. In this section, we randomly created 1000 virtual user sets (i.e., we randomly assigned the 972 face samples to the 972 fingerprint samples, 1000 times). In Fig. 18, we plot the ROC's for all of these 1000 cases, with the one used previously in this paper highlighted in red.

The minimum, mean, maximum and standard deviation of the EER set (with 1000 members) is found to be 0.82%, 1.1%, 1.5% and 0.11, respectively. The EER for the one case used previously in this paper is 0.99%. The close clustering of these curves, and the low standard deviation, supports the independence assumption between face and fingerprint biometrics and would seem to validate the use of virtual subjects. Furthermore the "thickness" of this cluster of curves supports other observations that performance estimates vary by +/- 1%.

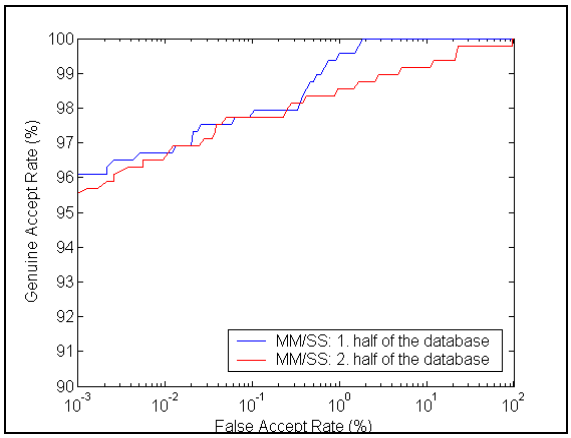

Fig. 17. Generalizability: ROC curves for disjoint portions of the whole database.

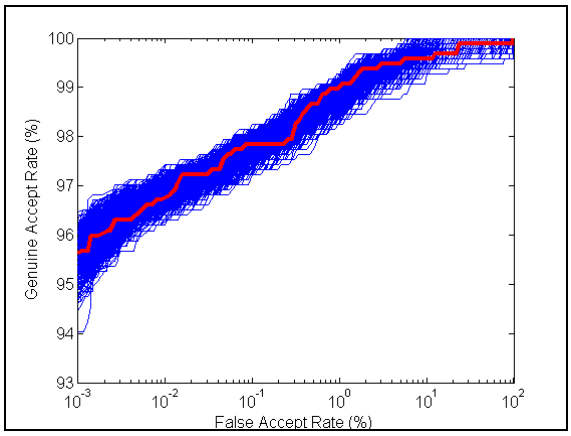

Fig. 18. Effects of virtual subject creation.

5. Conclusions

We examined the performance of multimodal biometric authentication systems using state-of-the-art Commercial Off-the-Shelf (COTS) fingerprint and face biometrics on a population approaching 1000 individuals, 10 times larger than previous studies. We introduced novel normalization and fusion methods along with well-known methods to accomplish match score level multimodal biometrics. Our work shows that COTS-based multimodal fingerprint and face biometric systems can achieve better performance than unimodal COTS systems. However, the performance gains are smaller than those reported by prior studies of non-COTS based multimodal systems (a ~2.3% gain here as compared to a ~12.9% gain reported in [2], at 0.1% FAR). This was expected, given that higher-accuracy COTS systems leave less room for improvement. Our analysis of fusion and normalization methods suggests that for authentication applications, which normally deal with open populations (e.g., airports) whose specific information is not known in advance, Min-Max normalization and Simple-Sum fusion generally out perform unimodal biometrics. For applications which deal with closed populations (e.g., a laboratory), where repeated samples and their statistics can be accumulated, our novel QLQ *adaptive normalization* and UW fusion methods tend to out perform Min-Max normalization and Simple-Sum fusion.

Our analysis of multimodal face-fingerprint pair systems shows that better performance is obtained by combining complementary systems rather than the best individual systems. And our investigations of performance variability across different datasets have provided evidence that the use of virtual subjects is valid, and offer initial estimates of variability for COTS-based multimodal systems .

6. References

[1] NIST report to the United States Congress, "Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability", November 13, 2000.

[2] R. Snelick, M. Indovina, J. Yen, A. Mink, "Multimodal Biometrics: Issues in Design and Testing", Proc. of The 5th International Conference on Multimodal Interfaces (IMCI 2003), November 2003, Vancouver, British Columbia, Canada.

[3] A.K. Jain, R. Bolle, and S. Pankanti, Eds. Biometrics: Personal Identification in Networked Society, Kluwer Academic Publishers, 1999.

[4] A. Ross and A.K. Jain, "Information Fusion in Biometrics", Proc. of AVBPA, Halmstad, Sweden, June 2001, pp. 354-359.

[5] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheeps, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation", Proc. of ICSLD 98, Sydney, Australia, November 1998.

[6] P.J. Huber, Robust Statistics, Wiley, 1981.

[7] A.K. Jain and A. Ross, "Learning User-Specific Parameters in Multibiometric System", Proc. of International Conference on Image Processing (ICIP), Rochester, NY, September 2002, pp. 57-60.

[8] R.M. Bolle, S. Pankanti, and N.K. Ratha, "Evaluation techniques for biometrics-based authentication systems (FRR)", Proc. of ICPR 2000, 15th International Conference on Pattern Recognition, Sept 2000, vol. 2, pp. 831 -837.

[9] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone, "Face Recognition Vendor Test 2002, Evaluation Report", March 2003, ftp://sequoyah.nist.gov/pub/nist_internal_reports/ir_6965/FRVT_2002_Evaluation_Report.pdf