

# Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems

Robert Snelick<sup>1</sup>, Umut Uludag<sup>2\*</sup>, Alan Mink<sup>1</sup>, Michael Indovina<sup>1</sup> and Anil Jain<sup>2</sup>

<sup>1</sup>*National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD, 20899*

<sup>2</sup>*Michigan State University, Computer Science and Engineering, East Lansing, MI, 48824*

*{rsnelick, amink, mindovina}@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 biometric systems on a population approaching 1,000 individuals. Majority of prior studies of multimodal biometrics have been limited to relatively low accuracy non-COTS systems and populations of a few hundred users. 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 highly accurate COTS systems on a relatively large-scale population. In addition to examining well-known multimodal methods, we introduce new methods of normalization and fusion that further improve the accuracy.

**Index Terms:** Multimodal biometrics, authentication, matching score, normalization, fusion, fingerprint, face.

## 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

---

\* Corresponding author

biometrics system. The report recommends a system employing dual biometrics in a layered approach for large-scale applications such as border crossing. Use of multiple biometric indicators for identifying individuals, known as multimodal biometrics, has been shown to increase accuracy [2] and population coverage, while decreasing vulnerability to spoofing.

The key to multimodal biometrics is the fusion of various biometric modality data at the feature extraction, matching score, or decision levels [3]. Our methodology focuses on fusion at the matching score level. This approach has the advantage of utilizing as much information as possible from each biometric modality, while at the same time enabling the integration of proprietary Commercial Off-the-Shelf (COTS) biometric systems. Most vendors of biometric systems do not like to release the feature values computed by their systems. Note that a normalization step is generally necessary before combining scores originating from different matchers.

Majority of published studies examining fusion techniques have been limited to small populations (a few hundred individuals at most), while employing low performance non-commercial (e.g., locally developed) biometric systems. In this paper, we investigate the performance gains achievable by COTS multimodal biometric systems using a relatively large (nearly 1,000 individuals) population. Further, we propose new normalization and fusion methods that improve the multimodal system performance. A preliminary version of this research appeared in [4]. A version of this paper including color figures can be found at <http://biometrics.cse.msu.edu/publications.html>

## **2. Related Work**

A number of studies showing the advantages of multimodal biometrics have appeared in the literature. Brunelli and Falavigna [5] used hyperbolic tangent ( $\tanh$ ) for normalization and

weighted geometric average for fusion of voice and face biometrics. They also proposed a hierarchical combination scheme for a multimodal identification system. Kittler et al. [6] have experimented with several fusion techniques for face and voice biometrics, including sum, product, minimum, median, and maximum rules and they have found that the sum rule outperformed others. Kittler et al. [6] note that the sum rule is not significantly affected by the probability estimation errors and this explains its superiority.

Hong and Jain [7] proposed an identification system based on face and fingerprint, where fingerprint matching is applied after pruning the database via face matching. Ben-Yacoub et al. [8] considered several fusion strategies, such as support vector machines, tree classifiers and multi-layer perceptrons, for face and voice biometrics. The Bayes classifier is found to be the best method. Ross and Jain [9] combined face, fingerprint and hand geometry biometrics with sum, decision tree and linear discriminant-based methods. The authors report that sum rule outperforms others.

It should be noted that the number of samples per subject in the databases used by researchers affects the complexity of the appropriate fusion systems. More samples may allow utilizing complex knowledge-based (e.g., perceptron) techniques.

### 3. Score Normalization

In this section, we present three well-known normalization methods, and a new method, which we call *adaptive normalization*. We denote a raw matching score as  $s$  from the set  $S$  of all scores for that matcher, and the corresponding normalized score as  $n$ .

**Min-Max (MM):** This method maps the raw scores to the  $[0, 1]$  range. The quantities  $\max(S)$  and  $\min(S)$  specify the end points of the score range:

$$n = \frac{s - \min(S)}{\max(S) - \min(S)} \quad (1)$$

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

$$n = \frac{s - mean(S)}{std(S)} \quad (2)$$

**Tanh (TH):** This method is among the so-called *robust* statistical techniques [10]. It maps the raw scores to the (0, 1) range:

$$n = \frac{1}{2} \left[ \tanh \left( 0.01 \frac{(s - mean(S))}{std(S)} \right) + 1 \right] \quad (3)$$

**Adaptive (AD):** The errors of individual biometric matchers stem from the overlap of the genuine and impostor score distributions. We characterize this overlap region by 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, while still mapping the scores to [0,1] range.

Previously, test normalization (T-norm) [11] that can be thought of as adaptive normalization considering impostor scores is proposed.

Our adaptive normalization is formulated as  $n_{AD} = f(n_{MM})$ , where  $f()$  denotes the mapping function that is applied to the MM normalized scores,  $n_{MM}$ . We have considered the following three choices for the function  $f()$ . These functions use two parameters of the overlapping region,  $c$  and  $w$ , which can be either provided by the vendors or estimated by the

system integrator. In this work, we estimate these parameters.

- **Two-Quadratics (QQ):** This function is composed of two quadratic segments that change the concavity at  $c$  (Fig. 1a):

$$n_{AD} = \begin{cases} \frac{1}{c} n_{MM}^2, & n_{MM} \leq c \\ c + \sqrt{(1-c)(n_{MM} - c)}, & \text{otherwise} \end{cases} \quad (4)$$

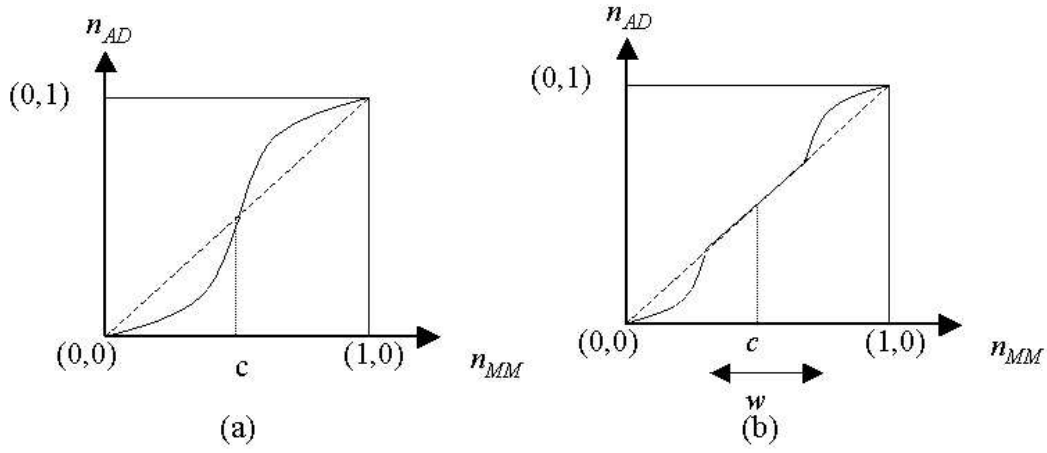


Fig. 1. Mapping functions for (a) QQ and (b) QLQ adaptive normalizations.

For comparison, the identity function,  $n_{AD} = n_{MM}$ , is also shown by the dashed lines in Fig. 1.

- **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. 1a. It is formulated as

$$n_{AD} = \frac{1}{1 + A \cdot e^{-B \cdot n_{MM}}}, \quad (5)$$

where the constants  $A$  and  $B$  are calculated as  $A = \frac{1}{\Delta} - 1$  and  $B = \frac{\ln A}{c}$ . Here,  $f(0)$  is equal to the constant  $\Delta$ , which is selected to be a small value (0.01 in this study). Note that, due to this

specification, the inflection point of the logistic function occurs at  $c$ , the center of the overlap region.

- **Quadric-Line-Quadric (QLQ):** The overlap zone, with center  $c$  and width  $w$ , is left unchanged while the other regions are mapped with two quadratic function segments (Fig. 1b):

$$n_{AD} = \begin{cases} \frac{1}{(c - \frac{w}{2})} n_{MM}^2, & n_{MM} \leq (c - \frac{w}{2}) \\ n_{MM}, & (c - \frac{w}{2}) < n_{MM} \leq (c + \frac{w}{2}) \\ (c + \frac{w}{2}) + \sqrt{(1 - c - \frac{w}{2})(n_{MM} - c - \frac{w}{2})}, & \text{otherwise.} \end{cases} \quad (6)$$

#### 4. Biometric Fusion

We experimented with five different fusion methods, namely simple-sum, min-score, max-score, matcher weighting and user weighting. The first three are well-known fusion methods; the last two are new and they take into account the performance of individual matchers in weighting their contributions. The quantity  $n_i^m$  represents the normalized score for matcher  $m$  ( $m = 1, 2, \dots, M$ , where  $M$  is the number of matchers) applied to user  $i$  ( $i = 1, 2, \dots, I$ , where  $I$  is the number of individuals in the database). The fused score for user  $i$  is denoted as  $f_i$ .

**Simple-Sum (SS):**  $f_i = \sum_{m=1}^M n_i^m, \forall i$

**Min-Score (MIS):**  $f_i = \min(n_i^1, n_i^2, \dots, n_i^M), \forall i$

**Max-Score (MAS):**  $f_i = \max(n_i^1, n_i^2, \dots, n_i^M), \forall i$

**Matcher Weighting (MW):** Weights are assigned to the individual matchers based on their Equal Error Rates (EER's). Denote the EER of matcher  $m$  as  $e^m$ ,  $m=1, 2, \dots, M$ . Then, the weight  $w^m$  associated with matcher  $m$  is calculated as

$$w^m = \frac{\left(1 / \sum_{m=1}^M \frac{1}{e^m}\right)}{e^m}. \quad (7)$$

Note that  $0 \leq w^m \leq 1, \forall m$ ,  $\sum_{m=1}^M w^m = 1$  and the weights are inversely proportional to the corresponding errors; the weights for *more accurate* matchers are higher than those of *less accurate* matchers. The MW fused score for user  $i$  is calculated as

$$f_i = \sum_{m=1}^M w^m n_i^m, \forall i. \quad (8)$$

**User Weighting (UW):** The User Weighting fusion method assigns weights to individual matchers that may be different for different users. Jain and Ross [12] proposed a similar scheme, but they exhaustively searched a coarse sampling of the weight space, where weights are multiples of 0.1 in the range [0, 1]. Their method can be prohibitively expensive if the number of fused matchers,  $M$ , is high, since the weight space is  $\mathfrak{R}^M$ ; further, coarse sampling as used in [12] may not find the optimal weight set. In our method, the UW fused score for user  $i$  is calculated as

$$f_i = \sum_{m=1}^M w_i^m n_i^m, \forall i, \quad (9)$$

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

The calculation of these user-dependent weights is based on the *wolf-lamb* concept introduced by Doddington et al. [13] for unimodal speech biometrics. They label the users who can be imitated easily as *lambs* (namely, impostors can provide biometric data similar to that of lambs); *wolves* on the other hand are those who can successfully imitate some other users. 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 pair of user and matcher,  $(i, m)$ . This lambness metric is then used to calculate the weights for biometric 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 for user  $i$ . 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 estimated, as is done in this study). Denote the means of these distributions as  $\mu_i^m(gen)$  and  $\mu_i^m(imp)$ , respectively, and denote the standard deviations as  $\sigma_i^m(gen)$  and  $\sigma_i^m(imp)$ , respectively. We use the d-prime metric [14] as a measure of the separation of these two distributions in formulating the lambness metric for user  $i$  and matcher  $m$  as:

$$d_i^m = \frac{\mu_i^m(gen) - \mu_i^m(imp)}{\sqrt{(\sigma_i^m(gen))^2 + (\sigma_i^m(imp))^2}} \quad (10)$$

If  $d_i^m$  is small, user  $i$  is a lamb for some wolves and 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

$$w_i^m = \frac{1}{\sum_{m=1}^M d_i^m} \cdot d_i^m \quad (11)$$

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

Fig. 2 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 she 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. Different biometric modalities or matchers can affect the lambness of each user differently.

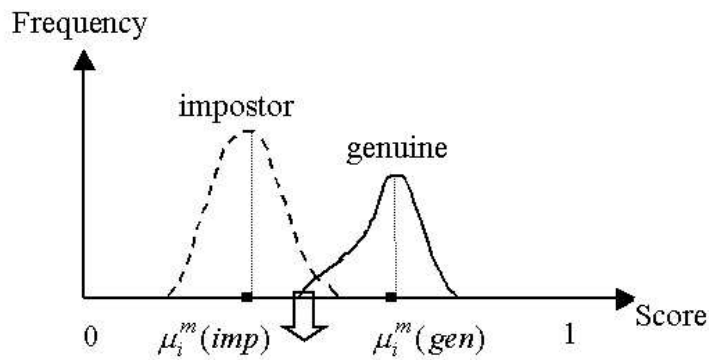


Fig. 2. Score distributions for a (user, matcher) pair: the arrow indicates the location of wolves for lamb  $i$ .

## 5. Experimental Results

We used the FERET image database [15] for face matching. The fingerprint image database that we used is proprietary and we cannot reveal many of its details; the fingerprint images were obtained with a live-scan, 500 dpi sensor, and their characteristics (e.g., size) are similar to those of public fingerprint databases. We had two fingerprint images for each of the 972 individuals, and we used two frontal face images of 972 individuals from the FERET database. Assuming

that face and fingerprint biometrics are statistically independent for an individual, a widely accepted and reasonable practice in multimodal biometrics research, we associated an individual from the face database with an individual from the fingerprint database, to create a *virtual* subject. Continuing in this fashion consistently, we arrived at our database consisting of 972 subjects, each having two face and two fingerprint images. One face and one fingerprint image for each subject are labeled as target, the remaining face and fingerprint image are labeled as query. For determining the normalization and fusion parameters, we used the entire database. The need for *virtual* subjects arises since there is no real multimodal database (where multiple biometrics attributes are measured on the same individual) of comparable size available in the public domain.

Matching scores were generated from four COTS biometric systems – three fingerprint systems and one face system. For each of these four systems, all query set images were matched against all target set images, yielding 972 genuine scores (where images are from the same subject) and 943,812 ( $972 \times 971$ ) imposter scores. The normalization and fusion operations are carried out using the generated similarity matrices to arrive at the final fused matching scores. The performance of individual matchers and different (normalization, fusion) permutations are presented via EER values, number of false rejections for subjects, and Receiver Operating Characteristics (ROC) curves. Among the three adaptive normalization methods (QQ, LG and QLQ) proposed before, the QLQ method gave the best results in our experiments, so it is selected as the representative adaptive normalization method. We carried out all possible permutations of (normalization, fusion) methods on our database of 972 subjects. Table 1 shows the EER values for these permutations. Note that EER values for the three individual fingerprint matchers (ordered Vendor 1, Vendor 2 and Vendor 3) and the face matcher are found to be 3.96%, 3.72%, 2.16% and 3.76%, respectively. The best, namely the lowest, 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 (normalization, fusion) permutations (%).

| <i>Normalization Method</i> | <i>Fusion Method</i> |             |              |             |              |
|-----------------------------|----------------------|-------------|--------------|-------------|--------------|
|                             | SS                   | MIS         | MAS          | MW          | UW           |
| MM                          | 0.99                 | 5.43        | 0.86         | <b>1.16</b> | <b>*0.63</b> |
| ZS                          | *1.71                | 5.28        | 1.79         | 1.72        | 1.86         |
| TH                          | 1.73                 | <b>4.65</b> | 1.82         | *1.50       | 1.62         |
| QLQ                         | <b>0.94</b>          | 5.43        | <b>*0.63</b> | <b>1.16</b> | <b>*0.63</b> |

As seen in Table 1, all of the fusion methods, except MIS fusion, lead to better performance than any of the individual matchers. Generally, MM and QLQ normalization methods outperform other normalization methods; SS, MW and UW fusion methods outperform other fusion methods.

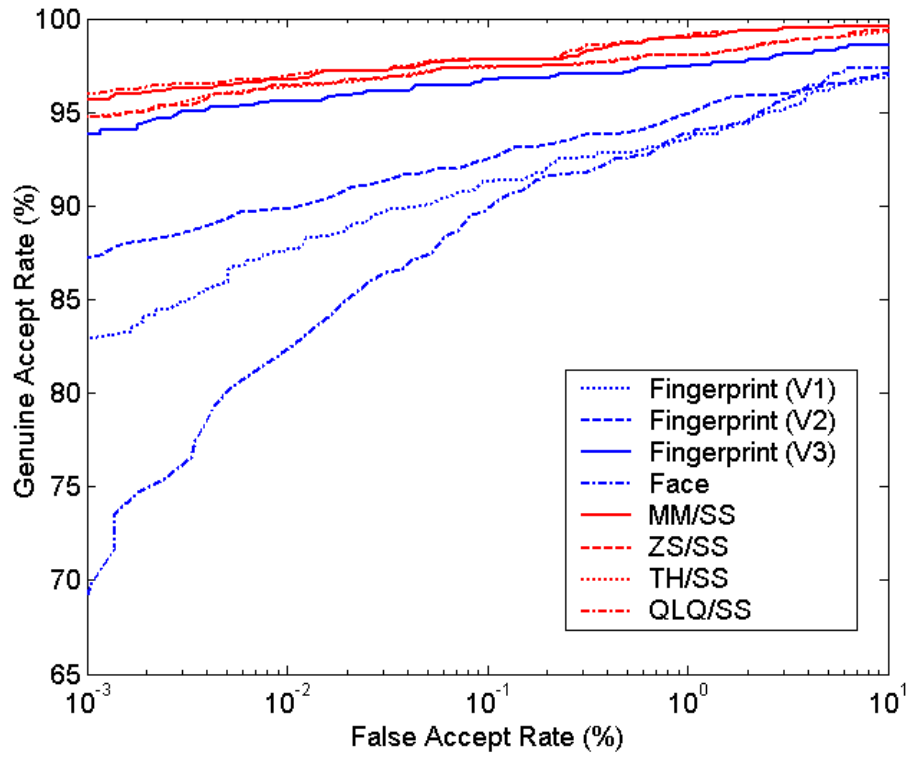
Further, we analyzed the system performance in terms of the number of falsely rejected subjects: At 1% and 0.1% FAR (False Accept Rate) values, we counted the number of false rejects for the individual matchers and QLQ/SS (namely, scores are normalized with QLQ method, and they are combined using SS fusion method) multimodal system. As shown in Table 2, the number of false rejects is considerably lower for the multimodal system compared to all of the unimodal matchers.

Table 2. Number of false rejects with matchers operating at 1% and 0.1% FAR.

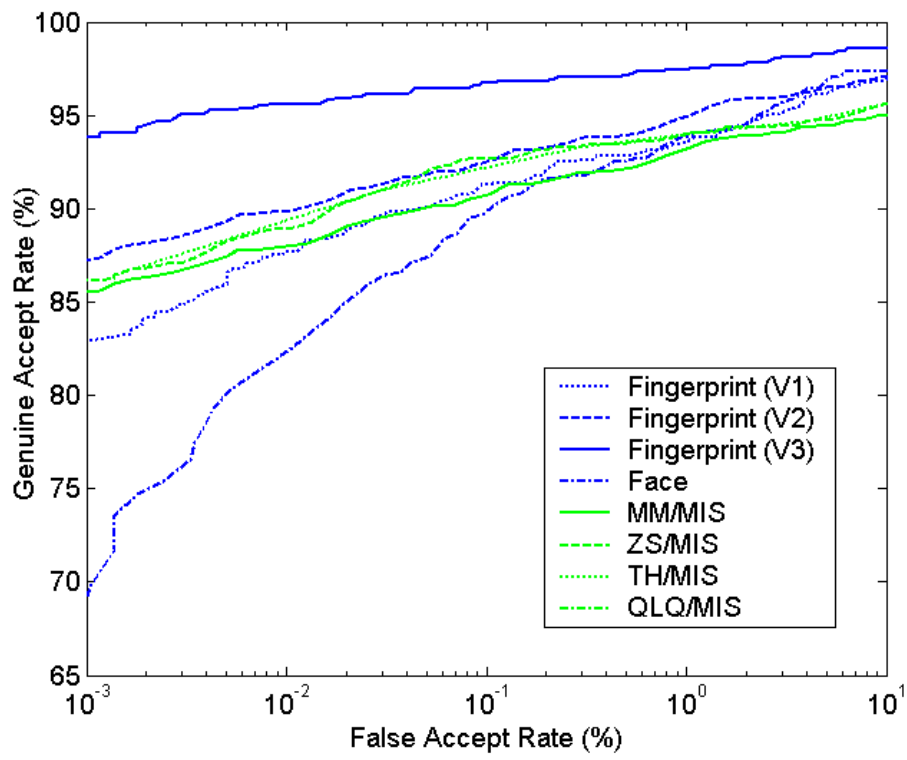
| <i>Matcher</i>           | <i>FAR</i> |      |
|--------------------------|------------|------|
|                          | 1%         | 0.1% |
| Fingerprint (Vendor 1)   | 62         | 85   |
| Fingerprint (Vendor 2)   | 48         | 72   |
| Fingerprint (Vendor 3)   | 25         | 32   |
| Face                     | 59         | 100  |
| QLQ/SS Multimodal System | 9          | 21   |

## 5.1. Normalization

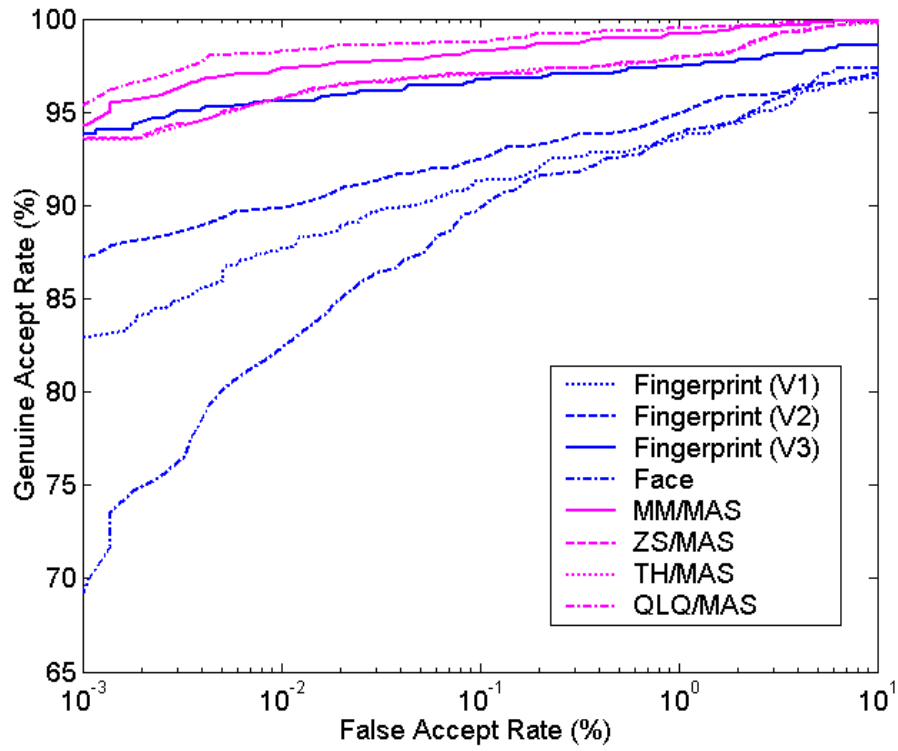
Fig. 3 shows the effect of each normalization method on system performance for different (but fixed) fusion methods. The ROC curves for the three fingerprint matchers and the face matcher are also shown for comparison. For MW fusion (Fig. 3d), the matcher weights, calculated according to Eq. (7), are: 0.2, 0.22, 0.37 and 0.21, for the three fingerprint matchers and the face matcher, respectively. For UW fusion (Fig. 3e), the mean user weights for these four individual biometric matchers, calculated from Eq. (11), are 0.14, 0.64, 0.17 and 0.05, respectively. This implies that, on average, fingerprint matcher V2 (corresponding to a mean user weight of 0.64) is the safest matcher for the lambs; whereas the space of the face matcher (corresponding to a mean user weight of 0.05) is filled with wolves (i.e., those waiting to be falsely accepted as some of the lambs). From Fig. 3 and Table 1, we see that QLQ and MM normalization methods lead to the 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.



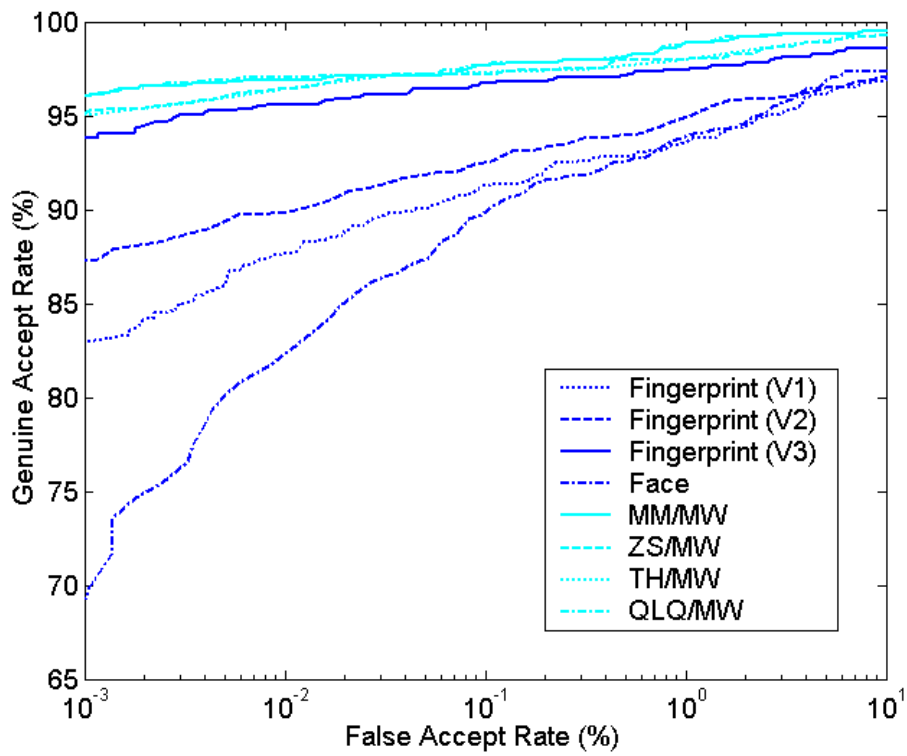
(a)



(b)



(c)



(d)

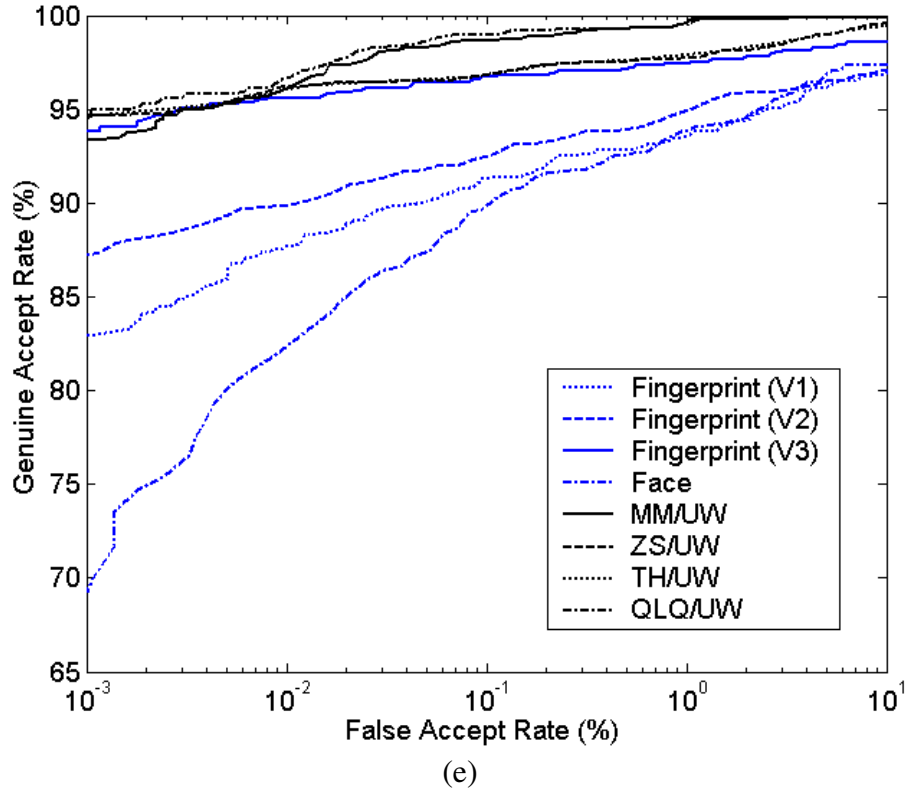


Fig. 3. Effects of normalization methods on system performance for different fusion methods:

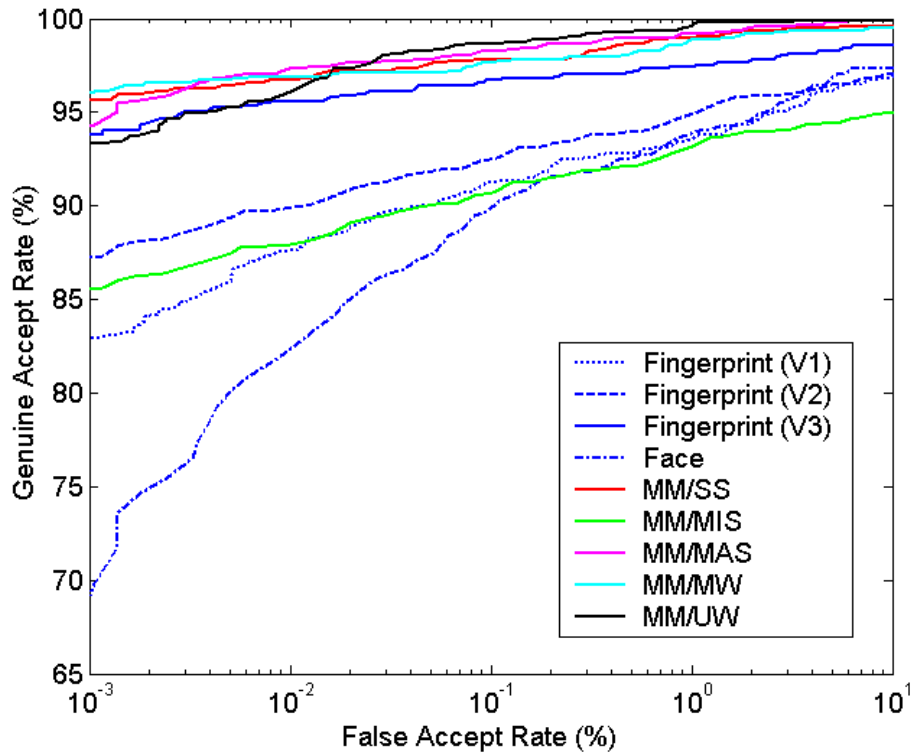
(a) SS fusion, (b) MIS fusion, (c) MAS fusion, (d) MW fusion, (e) UW fusion.

## 5.2. Fusion

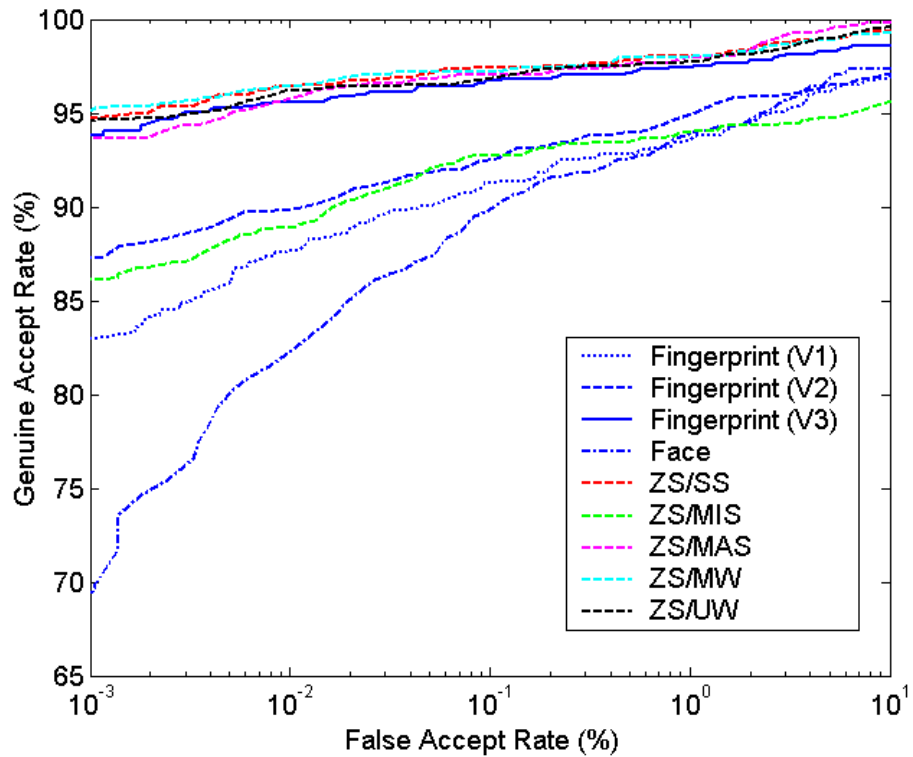
Fig. 4 shows the effect of each fusion method on system performance for different (but fixed) normalization methods. From Fig. 4 and Table 1, we see that fusion methods SS, MAS and MW generally perform better than the other two (MIS and UW). But for FAR in the range of [0.01%, 10%], UW fusion is better than the others. One reason that the performance of UW fusion drops below 0.01% FAR may be that the estimation errors become dominant.

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 is estimated; the user weight can be

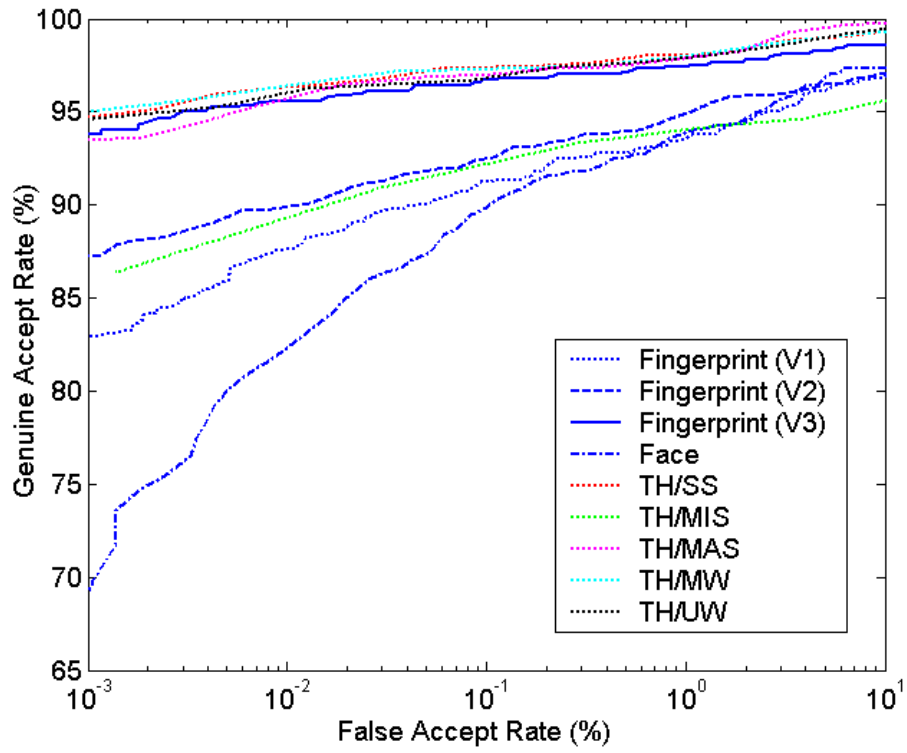
updated if the fusion system detects changes in the vulnerability of that user, due to fluctuations in their *lambness*, etc.



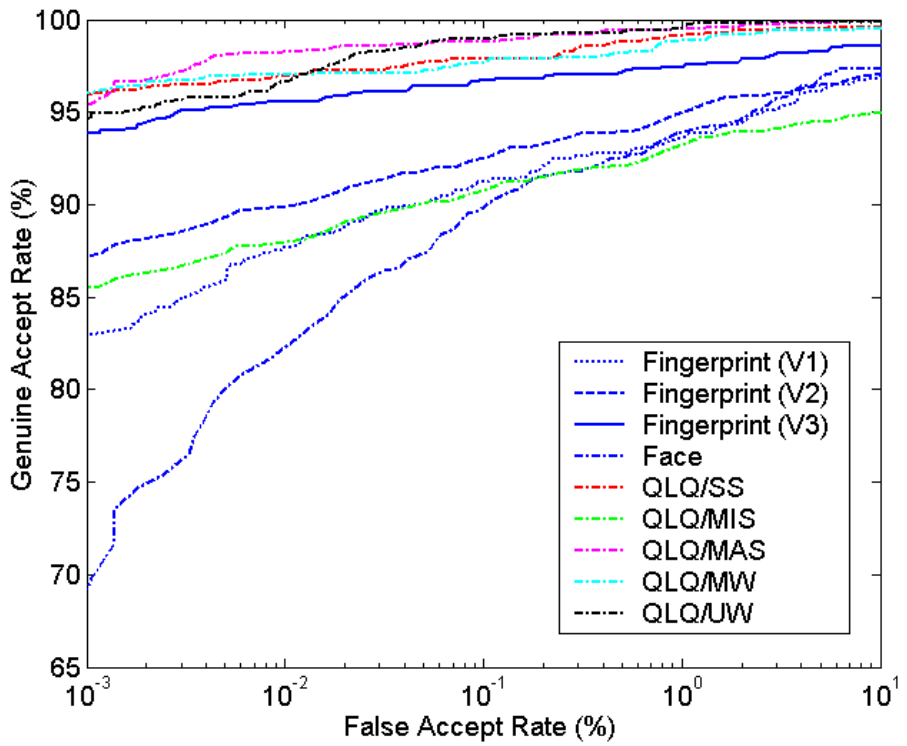
(a)



(b)



(c)



(d)

Fig. 4. Effects of fusion methods on system performance for different normalization methods:

(a) MM normalization, (b) ZS normalization, (c) TH normalization, (d) QLQ normalization.

## 6. Conclusions

We have examined the performance of multimodal biometric authentication systems using state-of-the-art Commercial Off-the-Shelf (COTS) fingerprint and face biometric matchers on a population approaching 1,000 individuals, which is considerably larger than previous studies. We have introduced new normalization and fusion methods to accomplish matching score level fusion of 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. This was expected, given that higher-accuracy COTS systems leave less room for improvement via fusion. Further, if we consider relative performance gains, an EER improvement of 1% will mean halving of false accept and false reject numbers when we have a highly accurate system (e.g., originally having 2% EER). But this 1% EER decrease may not translate to a large improvement if the underlying system was less accurate (e.g., originally having 5% EER), as it will lead to just 20% decrease in false accept and false reject numbers.

Our analysis of normalization and fusion methods suggests that for authentication applications that normally deal with open populations (e.g., airports), whose specific characteristics are not known in advance, Min-Max normalization and Simple-Sum fusion methods can be employed. For applications that deal with closed populations (e.g., an office environment), where repeated user samples and their statistics can be accumulated, the proposed QLQ *adaptive normalization* and UW *user weighting* fusion methods can be used.

## References

- [1] NIST Report to the United States Congress, “Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability”, Nov. 13, 2002.

- [2] A.K. Jain, R. Bolle, and S. Pankanti, (Eds.), *Biometrics: Personal Identification in Networked Society*, Kluwer Academic Publishers, 1999.
- [3] D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*, Springer, 2003.
- [4] M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach", *Proc. MMUA 2003, Workshop on Multimodal User Authentication*, pp. 99-106, Santa Barbara, CA, Dec. 11-12, 2003.
- [5] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues", *IEEE Trans. PAMI*, vol. 17, no. 10, pp. 955-966, 1995.
- [6] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers", *IEEE Trans. PAMI*, vol. 20, no. 3, pp. 226-239, 1998.
- [7] L. Hong and A.K. Jain, "Integrating Faces and Fingerprints for Personal Identification", *IEEE Trans. PAMI*, vol. 20, no. 12, pp. 1295-1307, 1998.
- [8] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of Face and Speech Data for Person Identity Verification", *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 1065-1075, 1999.
- [9] A. Ross and A.K. Jain, "Information Fusion in Biometrics", *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115-2125, 2003.
- [10] P.J. Huber, *Robust Statistics*, Wiley, 1981.
- [11] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems", *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.
- [12] A.K. Jain and A. Ross, "Learning User-Specific Parameters in a Multibiometric System", *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 57-60, Rochester, NY, Sept. 2002.
- [13] 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. ICSLD 98*, Sydney, Australia, Nov. 1998.

- [14] R.M. Bolle, S. Pankanti, and N.K. Ratha, "Evaluation Techniques for Biometrics-based Authentication Systems (FRR)", *Proc. 15th International Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 831-837, Sept. 2000.
- [15] The Facial Recognition Technology (FERET) Database, [http://www.itl.nist.gov/iad/humanid/feret/feret\\_master.html](http://www.itl.nist.gov/iad/humanid/feret/feret_master.html)