Multimodal Biometrics: Issues in Design and Testing

LONG PAPER

PAPER ID = 9

ABSTRACT

Experimental studies show that multimodal biometric systems for small-scale populations perform better performance than single-mode biometric systems. We examine if such techniques scale to larger populations, introduce a methodology to test the performance of such systems, and assess the feasibility of using commercial off-the-shelf (COTS) products to construct deployable multimodal biometric systems. A key aspect of our approach is to leverage confidence level scores from preexisting single-mode data. An example presents a multimodal biometrics system analysis that explores various normalization and fusion techniques for face and fingerprint classifiers. This multimodal analysis uses a population of about 1000 subjects, a number ten-times larger than seen in any previously reported study. Experimental results combining face and fingerprint biometric classifiers reveal significant performance improvement over single-mode biometric systems.

General Terms

Algorithms, Measurement, Performance, Design, Experimentation, Security, Human Factors, Standardization, Verification.

Keywords

Evaluation, Fusion, Multimodal Biometrics; Normalization; System Design; Testing Methodology;

1. INTRODUCTION

Single-mode biometric solutions have limitations in terms of accuracy, enrollment rates, and susceptibility to spoofing. A recent report [4] 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. The report recommends a system employing dual biometrics in a layered approach. Combining multiple sources of evidence improves performances, as demonstrated in several small-scale experimental studies performed in academia [1,2,3].

The key to multimodal biometrics is the fusion (i.e., combination) of the various biometric mode data and, if necessary, the normalization of that data to achieve values in a common range. Fusion can occur at the feature extraction, match-score, or decision level [2]. Feature level fusion combines feature vectors at the representation level that essentially provides higher dimensional data points when comparing the matching score. Match-score level fusion combines the disjoint confidence scores. Decision level fusion combines accept or reject decisions of individual systems. A majority vote scheme can then be employed, for example, to make the final judgment [10]. Our approach addresses fusion at the match-score level.

Limitations upon deployment of multimodal systems include lack of a common testing framework and the absence of tools to evaluate and build such systems. Core components of this work present (i) a verification testing methodology for multimodal biometric systems, (ii) an evaluation of normalization and fusion algorithms for a subject population ten-times larger than previously reported, and (iii) recommendations for designing multimodal biometric systems that can accommodate COTS products.

2. TESTING FRAMEWORK

We begin by introducing a methodology for testing multimodal biometric systems; the methodology provides a general framework for conducting normalization and fusion technique evaluations. The basis of this methodology is that fusion is applied after the individual biometric match-scores are determined. An advantage of fusion at this stage is that existing and proprietary biometric systems are not affected, allowing for a common middleware layer to handle the multimodal application but with a modicum of common information. Another advantage of using match (or confidence) scores is that data from prior evaluations of single-mode biometric systems can be reused. This avoids live testing or re-running individual biometric algorithms. One source of such data is the 2002 face recognition vendor test (FRVT 2002) [5].

The following is an overview of our adoption and extension of a single-mode biometric testing methodology proposed by Phillips et al. [5, 6]. A biometric signature is any form of biometric identifying data (e.g., a still fingerprint image or template of that information).

1. Assemble two sets of biometric signatures: a target and query set. The target set contains the set of signatures that are known to the system (i.e., the Biometric database). The query set contains signatures of subjects that are to be compared against the target set. The intersection of these two sets contains the subjects that should be found in the database. For practical tests the intersection should not be null. Although the same subjects are in both sets, separate instances of their biometric signatures should be used.

2. For each pair of query and target signatures obtain a match-score and store in a matrix, called a similarity matrix, whose size is query set size by target set size. The match-score is a measure of how similar two biometric signatures are. The match-score could represent, for example, a similarity or distance score.

3. Gallery and probe subsets can be extracted from the target/query similarity matrix, respectively, to perform "virtual" experiments on a subset of the population. A gallery is any arbitrary subset of the target set. A probe is any arbitrary subset of the query set.

4. Repeat steps 1-3 for each biometric mode.

5. Assemble and align the similarity matrices from step 2; this includes converting data to a common format, forming subsets to obtain matrices of the same size, and data mating to create real or virtual subjects. If the scores were produced by different sets of subjects, we rely upon the assumption that the individual modalities concerned are statistically independent of one another and could thus be assigned arbitrarily (though consistently) to form a set of mated virtual subjects for the purpose of testing. The result is a set of similarity matrices of equal size representing match-score data for mated subjects in a common format convenient for processing

6. Normalize the assembled similarity matrices to a common number range. Since this is an optional step, the transformation could be null and the output is equal to the input. A decision tree based fusion algorithm is a case where normalization may not be necessary. Normalization can be any post-processing transformation of the score data, but care should be taken not to reduce the dimensionality of the data [9].

7. Fuse the set of normalized similarity matrices into a single fusion similarity matrix. A fusion function, *f(x1, ...xn),* defines a mapping from *n-space*, where each biometric represents one of the *n* dimensions, into a single fused dimension. A threshold divides this range into an *accept* and *reject* part. Alternatively, decision level fusion defines a boundary that partitions the n-space into two parts representing accept and reject space. Operationally, the threshold or boundary is derived from an estimate of the Receiver Operating Characteristic (ROC) curve developed in step 8.

8. Performance statistics for verification are computed from the genuine and imposter scores. Genuine scores are those that result from comparing elements in the target and query sets of the same subject. Imposter scores are those resulting from comparisons of different subjects. Use each fusion score as a threshold and compute the false-accept rate (FAR) and false-reject rate (FRR) by selecting those imposter scores and genuine scores, respectively, on the wrong side of this threshold and dividing by the total number of scores used in the test. A mapping table of the threshold values and the corresponding error rates (FAR and FRR) are stored. The complement of the FRR (1 – FRR) is the genuine accept-rate (GAR). The GAR and the FAR are plotted against each other to yield a ROC curve, a common system performance measure. In practice, one chooses a desired operational point on the ROC curve and uses the FAR of that point to determine the corresponding threshold from the mapping table.

This framework allows a system designer to model hypothetical multimodal biometric systems that can vary the biometric indicator, matching algorithm, normalization and fusion techniques, and sample databases (e.g., the subject population or environmental conditions can be varied). Given this framework, systems can be built to optimally suit a particular application.

3. EVALUATION

We apply the principles laid out in the framework by examining two similarity matrices representing scores from a fingerprint and a face recognition system. Steps 1 through 4 of our testing methodology were previously completed. We now proceed to apply steps 5 through 8.

3.1 Databases

The fingerprint scores were obtained from a subset of a 60,000 x 60,000 similarity matrix previously generated by NIST using public domain fingerprint matching algorithms and 120,000 fingerprint images. The images were taken from 30,000 individuals who each contributed a primary and a secondary image for both of their index fingers.

The primary images were assigned to the target set and the secondary images were assigned to the query set. Because these sets are disjoint, all scores generated were for unique pairs of images, thus eliminating any concerns about "asymmetry" of the matching algorithm (note, the matcher used was in fact symmetric).

From this original matrix, we extracted a 1005 x 1005 sub-matrix into our common format containing only scores from comparing images of left index fingers for 1005 individuals.

The face scores were obtained from a subset of a 3,323 x 3,816 similarity matrix produced during prior evaluations [6] of an MIT developed face recognition algorithm ("MIT Standard, March 1995"). The scores result from comparisons of various facial images contributed by 1201 individuals to the FERET Database [11]. From this original matrix we extracted a 1005 x 1005 sub-matrix into our common format containing only scores obtained by comparing unique pairs of images from 1005 individuals.

We then arbitrarily, although consistently, assigned each of the 1005 "virtual subjects" to a set of face and finger scores (under the assumption that face and finger scores are independent of one another). This completes step 5 of our testing methodology.

3.2 Normalization

Normalization, step 6 of our testing methodology, is recommended for certain data fusion methods. Normalization addresses the problem of incomparable classifier output scores in different combination classification systems. Table 1 provides a summary of some well-known normalization techniques that we use in this study.

Table 1. Summary of Normalization Techniques.

Note: We denote the classifier output score by *s* and normalized score by *s'*	
Min-Max	*s' = (s - min) / (max-min)*
Z-score	*s' = (s - mean)/(standard deviation)*
MAD	*s' = (s - median)/constant(median | s - median|)*
Tanh	*s' = .5[tanh (.01(s - mean)/(standard deviation)) +1]*

3.3 Fusion

We apply a number of well-known fusion techniques [7], shown in Table 2, which is step 7 of our testing methodology. The simple sum rule adds the scores of each classifier to calculate the fused score. The Minimum Score fusion method selects the score having the least value of the classifiers. Likewise, the Maximum Score fusion method selects the score having the greatest value of the classifiers. The genuine posterior probability, *P(genuine* | s_i), represents the probability of a subject being genuine, given a score for a particular classifier (s_i). The Sum of Probabilities, and Product of Probabilities fusion techniques compute the fused scores by adding or multiplying, respectively, these probabilities for all classifiers.

For the probability fusion techniques, we follow the theoretical framework of Kittler et al. [7] that uses a training set of the first *n* (*n* = 100 in this study) subjects to estimate the population posterior probabilities of genuineness *P(genuine* | s_i *)* to combine these probabilities for a fused similarity score. We used the mean and variance of the genuine and imposter scores from this training set and assumed a normal distribution for their probability density function, *p(s | genuine)* and *p(s | imposter)*, to evaluate *P(genuine | S)= p(s | genuine) / [p(s | genuine) + p(s | imposter)]*. Using the actual density function, rather than assuming a normal distribution, may yield better results. Note for the sum of probabilities and product of probabilities fusion techniques the normalization step is not needed—normalization is implied in the algorithm.

Table 2. Summary of Fusion Techniques.

s_i is the score from the ith-classifier, assuming *N* classifiers; Let *P (genuine* | s_i *) and P (imposter* | s_i *)* be the posteriori probability of s_i being *genuine* or *imposter*	
Simple Sum	$\sum_{i=1}^{N} s_i$
Minimum Score	$\min(s_1, s_2, \cdots s_n)$
Maximum Score	$\max(s_1, s_2, \cdots s_n)$
Sum of Probabilities	$\sum_{i=1}^{N} P(genuine \mid s_i)$
Product of Probabilities	$\prod_{i=1}^{N} P(genuine \mid s_i)$

3.4 Experiments

Performance statistics, step 8 of the testing methodology, computes the ROC curves for our study. Figure 1 shows a ROC curve for the simple sum fusion rule with various normalization techniques. Clearly the use of these fusion and normalization techniques enhances the performance significantly over the single-modal face or fingerprint classifiers. For example, at a FAR of 0.1% the simple sum fusion with the min-max normalization has a GAR of 94.9%, which is considerably better than that of face, 75.3%, and fingerprint, 83.0%. Also, using any of the normalization techniques in lieu of not normalizing the data proves beneficial. The simplest normalization technique, the min-max, yields the best performance in this example.

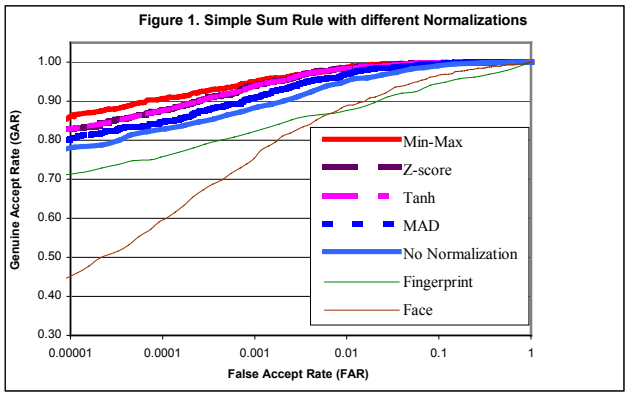
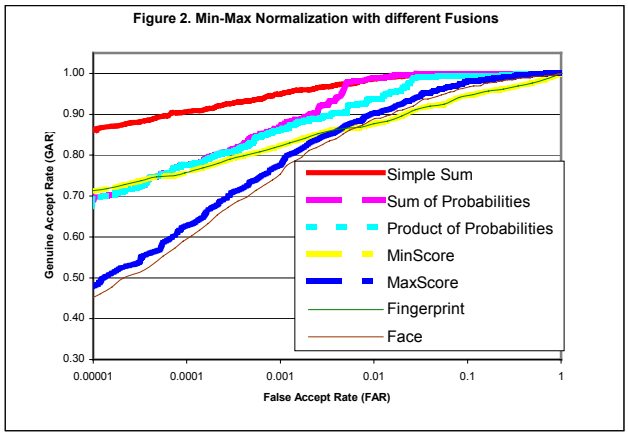


Figure 2 illustrates the results of Min-Max normalization for a spectrum of fusion methods. The simple sum fusion method yields the best performance over the range of FARs. Interestingly, the genuine-accept rate for sum and product probability rules falls off dramatically at a lower FAR.

Tables 3 and 4 show the GAR for the spectrum of normalization and fusion techniques at FARs of 1% and 0.1% respectively. At 1% FAR, the sum of probabilities fusion works the best. However, these results do not hold true at a FAR of 0.1%. The simple sum rule generally performs well over the range of normalization techniques. These results demonstrate the utility of using multimodal biometric systems for achieving better matching performance. They also indicate that the method chosen for fusion has a significant impact on the resulting performance.



In operational biometric systems, application requirements drive the selection of tolerable error rates, and in both single-modal and multimodal biometric systems, implementers are forced to make a trade-off between usability and security. Implementers produce ROC curves for their systems from their own test data based on

these guidelines. Operators use these ROC curves to determine the FAR of the security level needed for their application. The mapping table, from step 8 of our testing methodology, is used to determine the threshold value corresponding to that FAR. This mapping is usually done via an implementer provided utility, which may need to use extrapolation to determine certain values.

Table 3. Summary of Fusion Techniques, GAR at 1% FAR.

	Fusion Techniques				
Normalization Techniques	**Simple Sum**	**Max Score**	**Min Score**	**Sum of Prob.**	**Prod. of Prob.**
Min-Max	98.7 %	90.2 %	87.7 %	N/A	N/A
Z-Score	98.5 %	98.3 %	91.1 %	N/A	N/A
Tanh	98.5 %	98.1 %	91.1 %	N/A	N/A
MAD	96.9 %	93.4 %	91.1 %	N/A	N/A
None (implied)	94.6 %	93.4 %	87.7 %	99.0 %	93.7 %

Looking at the data from a slightly different perspective, we count the number of subjects who were rejected by either face or fingerprint, or by both classifiers, but accepted by fusion. Table 5 summarizes the false-rejections for the various classifiers at a given FAR. Of the 1005 genuine subjects at a FAR of 1%, there were 4 cases where a subject was rejected for both the face and fingerprint indicator, but was accepted with the min-max normalization/simple sum fusion system. Likewise, at a FAR of 0.1% there were 11 such cases. As expected, the acceptance rates are more dramatic when compared to those for the individual modalities. These results suggest that multimodal biometric systems can be deployed that will increase security while reducing the number of false rejections.

Table 4. Summary of Fusion Techniques, GAR at .1% FAR.

	Fusion Techniques				
Normalization Techniques	**Simple Sum**	**Max Score**	**Min Score**	**Sum of Prob.**	**Product of Prob.**
Min-Max	94.9 %	77.9 %	83.0 %	N/A	N/A
Z-Score	94.2 %	87.9 %	85.1 %	N/A	N/A
Tanh	94.4 %	87.5 %	85.1 %	N/A	N/A
MAD	90.7 %	83.2 %	84.3 %	N/A	N/A
None (implied)	88.5 %	83.0 %	82.6 %	87.3 %	86.2 %

Conversely, we also examine those subjects who were accepted by either face or fingerprint classifier but rejected by fusion. At a FAR of 1% 4 subjects passed the fingerprint system but failed fusion. There were no such cases for face. At 0.1% 20 subjects passed the fingerprint system but failed fusion. Likewise, 3 subjects passed the face system but failed fusion.

Table 5. False Rejections for 1005 subjects in the Unimodal and Multimodal Biometric Systems

	False Rejections	
Classifier	**0.1% FAR**	**1% FAR**
Face	248	124
Fingerprint	183	112
Simple Sum	51	13
Both Face and Finger	39	8
All Three	28	4

It is important to note that although our findings support the results from earlier small-scale studies, the results presented here are applicable only for the data in this study. No inferences can be drawn to predict performance of a system as we scale the subject population [8]. This emphasizes the need to conduct experiments on representative data sets for even larger populations.

4. SYSTEM DESIGN

The advantage of fusion at the match-score level is that existing and proprietary single-mode biometric systems can easily be integrated into a multimodal biometric environment if some basic information is provided by these existing systems. The needed information does not expose any of the internal operations of these systems. The following is a list of preliminary recommendations for the information needed from existing systems that could hasten interoperability and plug-n-play in such an environment:

- The match-score (confidence level), its range and distribution should be exposed in a common format.
- A set of training data or distributions for sample test populations.

Our long-term goal is to develop a middleware environment that would support multimodal biometric applications. Plug-n-play architectures can be built from individual single-mode biometric systems supporting the requirements stated above. As a first step towards achieving this goal we are constructing a prototype multimodal biometric system that combines face and fingerprint classifiers from two independent COTS products of different vendors, as shown in figure 3. This system is built at the application level and fuses match-score data provided by each of the vendor's software development kits.

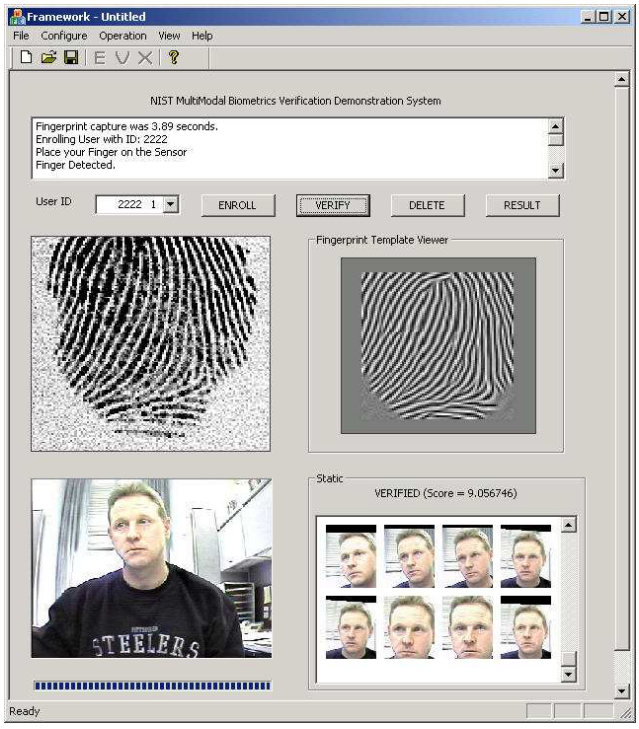


Figure 3. Prototype Multimodal Biometric System.

5. SUMMARY AND FUTURE WORK

We have established a framework capable of assessing the performance of multimodal biometric systems. We have demonstrated the utility of this methodology by examining relatively large face and fingerprint data sets over a spectrum of normalization and fusion techniques. The results of this study, which uses a population ten-times larger than previously reported, supports the results of smaller studies that show multimodal biometric systems out perform single-mode biometric systems. An additional advantage of fusion at this level is that existing and proprietary biometric systems do not need to be modified, allowing for a common middleware layer to handle the multimodal applications with a modicum of common information. Future work will investigate alternative normalization and fusion methods, while honing our proposed testing methodology.

NIST, in its extensive single-mode biometrics testing, has concluded [4,8] that to accurately evaluate the performance of biometric systems, tests must be performed with data sets on the order of tens-of-thousands subjects and that no inferences be drawn from tests conducted on small subject populations to assess system scalability. Thus, future plans include expanding the test databases to attain these larger sizes. In addition, to assess the feasibility of such systems for large-scale deployments, we will perform these tests using COTS products.

6. ACKNOWLEDGMENTS

Professor Anil Jain provided insight that helped shape the research efforts and reviewed earlier drafts of this paper. Ross Micheals, Mike Garris, Patrick Grother, and Stan Janet provided access to the single-mode biometric data and assistance in interpreting the data.

7. REFERENCES

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

[2] Ross, A. and Jain, A., Information Fusion in Biometrics. In Proceedings AVBPA, Halmstad, Sweden, June 2001, pp. 354-359.

[3] Jain, A. and Ross, A., "Learning User-Specific Parameters in Multibiometric System" In proceedings of IEEE ICIP, Rochester, NY. September 2002.

[4] NIST report to the United State Congress. "Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability", November 13, 2000. http://www.itl.nist.gov/iad/894.03/NISTAPP_Nov02.pdf

[5] Phillips, J., et al. "Face Recognition Vendor Test 2002: Evaluation Report". NISTIR 6965, March 2003. http://www.frvt.org.

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

[7] J. Kittler, M. Hatef, R.P. Duin, J.G. Matas, "On Combining Classifiers", IEEE Transactions on PAMI 20 (3) (1998) 226-239.

[8] Wilson, C. National Institute of Standards and Technology (NIST). Personal Communication.

[9] Altincay, H. and Demireler, M. "Undesirable effects of output normalization in multiple classifier systems", Pattern Recognition Letter 24 (2003) 1163-1170.

[10] Y. Zuev, S. Ivanon, "The Voting as a way to increase the decision reliability, in: Foundations of Information/Decision Fusion with Applications to Engineering Problems", Washington D.C. USA, 1996, pp. 206-210.

[11] The Facial Recognition Technology (FERET) Database, NIST, http://www.itl.nist.gov/iad/humanid/feret/feret_master.html