This is technical report NISTIR 6281, October 1998.

The FERET Verification Testing Protocol for Face Recognition Algorithms

Syed A. Rizvi [a,1], P. Jonathon Phillips [b] and Hyeonjoon Moon [c]

[a] *Department of Applied Sciences, College of Staten Island of City University of New York, Staten Island, NY 10314*

[b] *National Institute of Standards and Technology, Gaithersburg, MD 20899, jonathon@nist.gov*

[c] *Department of Electrical and Computer Engineering, State University of New York at Buffalo, Amherst, NY 14260*

Abstract

Two critical performance characterizations of biometric algorithms, including face recognition, are identification and verification. In face recognition, FERET is the de facto standard evaluation methodology. Identification performance of face recognition algorithms on the FERET tests has been previously reported. In this paper we report on verification performance obtained from the Sep96 FERET test. Results are presented for images taken on the same day, for images taken on different days, for images taken at least one year apart, and for images taken under different lighting conditions.

Key words: Face Recognition, FERET, Algorithm Evaluation, Verification

1 Introduction

Identification and verification of a person's identity are two potential areas for applications of face recognition systems. In identification applications, a system identifies an unknown face in an image; i.e., searching an electronic

[1] This work was performed as part of the Face Recognition Technology (FERET) program, which is sponsored by the U.S. Department of Defense Counterdrug Technology Development Program. Portions of this work were support by the National Institute of Justice. Portions of this were done while Jonathon Phillips was at the U.S. Army Research Laboratory (ARL). Please direct correspondence to Jonathon Phillips.

mugbook for the identity of suspect. In verification applications, a system confirms the claimed identity of a face presented to it. Proposed applications for verification systems include, controlling access to buildings and computer terminals, confirming identities at automatic teller machines (ATMs), and verifying identities of passport holders at immigration ports of entry. These applications have a potential to influence and impact our everyday life.

For systems to be successfully fielded, it is critical that their performance is known. To date the performance of most algorithms has only been reported on identification tasks, which implies that characterization on identification tasks holds for verification. For face recognition systems to successfully meet the demands of verification applications, it is necessary to develop testing and scoring procedures that specifically address these applications.

A scoring procedure is one of two parts of an evaluation protocol. In the first part, an algorithm is executed on a test set of images and the output from executing the algorithm is written to a file(s). This produces the raw results. In the second part, a scoring procedure processes raw results and produces performance statistics. If the evaluation protocol and its associated scoring procedure are properly designed, the performance statistics can be computed for both identification and verification scenarios.

The Sep96 FERET evaluation method is such a protocol [9,10]; it used images from the FERET database of facial images [11]. The Sep96 FERET test is the latest in a series of FERET tests to measure the progress, assess the state-of-the-art, identify strengths and weakness of individual algorithms, and point out future directions of research in face recognition. Prior analysis of the FERET results has concentrated on identification scenarios. In this paper we present (1) a verification analysis method for the Sep96 FERET test, and (2) results for verification.

2 The Sep96 FERET test

The Sep96 FERET testing protocol was designed so that algorithm performance can be computed for identification and verification evaluation protocols for a variety of different galleries and probe sets [9,10]. (The *gallery* is the set of known individuals. An image of an unknown face presented to an algorithm is called a *probe*, and the collection of probes is called the *probe set*.)

In the Sep96 protocol, an algorithm is given two sets of images: the *target set* and the *query set*. We introduce this terminology to distinguish these sets from the gallery and probe sets that are used in computing performance statistics. The target set is given to the algorithm as the set of known facial images.

The images in the query set are the unknown facial images to be identified. For each image q_i in the query set Q, an algorithm reports the similarity $s_i(k)$ between q_i and each image t_k in the target set T. The key property of this protocol, which allows for greater flexibility in scoring, is that for any two images s_i and t_k, we know $s_i(k)$.

From the output files, algorithm performance can be computed for virtual galleries and probe sets. A gallery G is a virtual gallery if G is a proper subset of the target set; i.e., $G \subset T$. Similarly, P is a virtual probe set if $P \subset Q$. For a given gallery G and probe set P, the performance scores are computed by examination of the similarity measures $s_i(k)$ such that $q_i \in P$ and $t_k \in G$.

The virtual gallery and probe set technique allows us to characterize algorithm performance for identification and verification. Also, performance can be broken out by different categories of images, e.g., probes taken on the same or different days than the corresponding gallery image. We can create a gallery of 100 people and estimate an algorithm's performance at recognizing people in this gallery. Using this as a starting point, we can create virtual galleries of $200, 300, \ldots, 1000$ people and determine how performance changes as the size of the gallery increases. Another avenue of investigation is to create n different galleries of size 100, and calculate the variation in algorithm performance for these galleries.

In the September 1996 FERET test, the target set contained 3323 images and the query set 3816 images. All the images in the target set were frontal images. The query set consisted of all the images in the target set plus rotated images and digitally modified images. For each query image q_i, an algorithm outputs the similarity measure $s_i(k)$ for all images t_k in the target set. For a given query image q_i, the target images t_k are sorted by the similarity scores $s_i(\cdot)$.

Except for a set of rotated and digitally modified images, the target and query sets are the same. Thus, the test output contains every target image matched with itself. This allowed a detailed analysis of performance on multiple galleries and probe sets. (We do not present results in this paper for the rotated or digitally modified images.)

There are two versions of the September 1996 test. The target and query sets are the same for each version. The first version requires that the algorithms be fully automatic. (In the fully automatic version the test algorithms are given a list of the images in the target and query sets. Locating the faces in the images must be done automatically.) In the second version, the eye coordinates are given. Thus, algorithms do not have to locate the face in the image.

We report the results for 12 algorithms. The test was administered in September 1996 and March 1997 (see Table 1 for details of when the test was administer to which groups and which version of the test was taken). Two of

Table 1
List of groups that took the Sept96 test broken out by versions taken and dates administered. (The 2 by MIT indicates that two algorithms were tested.)

		Test Date		
		September	March	
Version of test	Group	1996	1997	Baseline
Fully Automatic	MIT Media Lab [4,6]	•		
	U. of So. California (USC) [15]		•	
Eye Coordinates Given	Baseline PCA [7,13]			•
	Baseline Correlation			•
	Excalibur Corp.	•		
	MIT Media Lab	2		
	Michigan State U. [12]	•		
	Rutgers U. [14]	•		
	U Maryland [2]	•	•	
	USC		•	

these algorithms were developed at the MIT Media Laboratory. The first was the same MIT algorithm that was tested in March 1995 [5,8]. This algorithm was retested so that improvement since March 1995 could be measured. The second MIT algorithm was based on more recent work [4]. Algorithms were also tested from Excalibur Corp. (Carlsbad, CA), Michigan State University (MSU) [12], Rutgers University [14], University of Southern California (USC), and two from University of Maryland (UMD) [2,16]. The first algorithm from UMD was tested in September 1996 and a second version of the algorithm was tested in March 1997. The final two algorithms were our implementation of normalized correlation and a principal components analysis (PCA) based algorithm [7,13]. These algorithms provide a performance baseline. In our implementation of the PCA-based algorithm, all images were (1) translated, rotated, and scaled so that the center of the eyes were placed on specific pixels, (2) faces were masked to remove background and hair, and (3) the non-masked facial pixels were processed by a histogram equalization algorithm. The training set consisted of 500 faces. Faces were represented by their projection onto the first 200 eigenvectors and were identified by a nearest neighbor classifier using the L_1 metric. For normalized correlation, the images were (1) translated, rotated, and scaled so that the center of the eyes were placed on specific pixels and (2) faces were masked to remove background and hair.

We only report results for the semi-automatic case (eye coordinates given), because the Media Lab and U. of Southern California were the only groups to

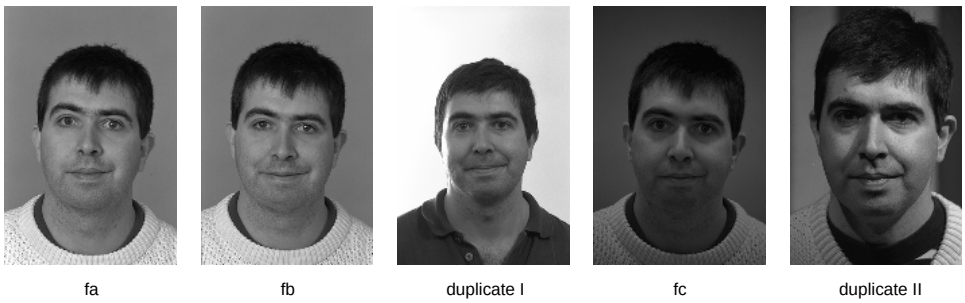


Fig. 1. Examples of different categories of probes (image). The duplicate I image was taken within one year of the **fa** image and the duplicate II and **fa** images were taken at least one year apart.

take the fully automatic test.

The images were taken from the FERET database of facial images [11]. The facial images were collected in 15 sessions between August 1993 and July 1996. Sessions lasted one or two days. To maintain a degree of consistency throughout the database, the same physical setup was used in each photography session. However, because the equipment had to reassembled for each session, there was variation from session to session.

Images of an individual were acquired in sets of 5 to 11 images, collected under relatively unconstrained conditions. Two frontal views were taken (**fa** and **fb**); a different facial expression was requested for the second frontal image. For 200 sets of images, a third frontal image was taken with a different camera and different lighting (this is referred to as the **fc** image). Figure 1 shows an example of the different categories of images.

By July 1996, 1564 sets of images were in the database, for 14,126 total images. The database contains 1199 individuals and 365 duplicate sets of images. For some people, over two years elapsed between their first and most recent sittings, with some subjects being photographed multiple times. The development portion of the database consisted of 503 sets of images, which were released to researchers. The remaining images were sequestered by the Government.

3 Verification Model

In our verification model, a person in image p claims to be the person in image g. The system either accepts or rejects the claim. (If p and g are images of the same person then we write $p \sim g$, otherwise, $p \not\sim g$.) Performance of the system is characterized by two performance statistics. The first is the probability of accepting a correct identity; formally, the probability of the

algorithm reporting $p \sim g$ when $p \sim g$ is correct. This is referred to as the verification probability, denoted by P_V (also referred to as the hit rate in the signal detection literature). The second is the probability of incorrectly verifying a claim formally, the probability of the algorithm reporting $p \sim g$ when $p \not\sim g$. This is called the false-alarm rate and is denoted by P_F.

Verifying the identity of a single person is equivalent to a detection problem where the gallery $G = \{g\}$. The detection problem consists of finding the probes in $p \in P$ such that $p \sim g$.

For a given gallery image g_i and probe p_k, the decision of whether an identity was confirmed or denied was generated from $s_i(k)$. The decisions were made by a *Neyman-Pearson* observer. A Neyman-Pearson observer confirms a claim if $s_i(k) \leq c$ and rejects it if $s_i(k) > c$. By the Neyman-Pearson theorem [3], this decision rule maximized the verification rate for a given false alarm rate α. Changing c generated a new P_V and P_F. By varying c from it's minimum to maximum value, we obtained all combinations of P_V and P_F. A plot of all combinations of P_V and P_F is a receiver operating characteristic (ROC) (also known as the relative operating characteristic) [1,3]. The input to the scoring algorithm was $s_i(k)$; thresholding similarity scores, and computing P_V, P_F, and the ROCs was performed by the scoring algorithm.

The above method computed a ROC for an individual. However, we need performance over a population of people. To calculate a ROC over a population, we performed a round robin evaluation procedure for a gallery G. The gallery contained one image per person.

The first step generated a set of partitions of the probe set. For a given $g_i \in G$, the probe set P is divided into two disjoint sets D_i and F_i. The set D_i consisted of all probes p such that $p \sim g_i$ and F_i consisted of all probes such that $p \not\sim g_i$.

The second step computed the verification and false alarm rates for each gallery image g_i for a given cut-off value c, denoted by $P_V^{c,i}$ and $P_F^{c,i}$, respectively. The verification rate was computed by

$$P_V^{c,i} = \begin{cases} 0 & \text{if } |D_i| = 0 \\ \frac{|s_i(k) \leq c \text{ given } p_k \in D_i|}{|D_i|} & \text{otherwise,} \end{cases}$$

where $|s_i(k) \leq c \text{ given } p \in D_i|$ was the number of probes in D_i such that $s_i(k) \leq c$. The false alarm rate is computed by

$$P_F^{c,i} = \begin{cases} 0 & \text{if } |F_i| = 0 \\ \frac{|s_i(k) \leq c \text{ given } p_k \in F_i|}{|F_i|} & \text{otherwise.} \end{cases}$$

The third step computed the overall verification and false alarm rates, which was a weighted average of $P_V^{c,i}$ and $P_F^{c,i}$. The overall verification and false-alarm rates are denoted by P_V^c and P_F^c, and was computed by

$$P_V^c = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{|D_i|}{\frac{1}{|G|}\sum_i |D_i|} P_V^{c,i} = \frac{1}{\sum_i |D_i|} \sum_{i=1}^{|G|} |s_i(k) \leq c \text{ given } p_k \in D_i| \cdot P_V^{c,i}$$

and

$$P_F^c = \frac{1}{|G|} \sum_{i=1}^{|G|} \frac{|F_i|}{\frac{1}{|G|}\sum_i |F_i|} P_F^{c,i} = \frac{1}{\sum_i |F_i|} \sum_{i=1}^{|G|} |s_i(k) \leq c \text{ given } p_k \in F_i| \cdot P_F^{c,i}.$$

The verification ROC was computed by varying c from $-\infty$ to $+\infty$.

In reporting verification scores, we state the size of the gallery G which was the number of images in the gallery set G and the number of images in the probe set P. All galleries contained one image per person, and probe sets could contain more than one image per person. Probe sets did not necessarily contain an image of everyone in the associated gallery. For each probe p, there existed a gallery image g such that $p \sim g$.

For a given algorithm, the choice of a suitable hit and false alarm rate pair depends on a particular application. However, for performance evaluation and comparison among algorithms, the equal error rate is often quoted. The equal error rate occurs at the threshold c where the incorrect rejection and false alarm rates are equal; that is $1 - P_V^c = P_F^c$ (incorrect rejection rate is one minus the verification rate.)

4 Verification Results

To provide a detailed analysis of algorithm performance, we report verification scores for four categories of probes. The first probe category was the **FB** probes. For each set of images, there were two frontal images. One of the images was randomly placed in the gallery, and the other images was placed in the **FB** probe set. (This category is denoted by **FB** to differentiate it from the **fb** images in the FERET database.) The second probe category was all duplicates of the gallery images. We refer to this category as the duplicate I probes. The third category was the **fc** (images taken the same day, but with a different camera and lighting), and the fourth consisted of duplicates where there is at least one year between the acquisition of the probe image and corresponding gallery image. We refer to this category as the duplicate II probes. For this category, the gallery images were acquired before January

Table 2
Figures reporting results broken out by probe category.

Figure no.	Probe category	Gallery size	Probe set size
2	**FB**	1196	1195
3	Duplicate I	1196	722
4	**fc**	1196	194
5	Duplicate II	864	234

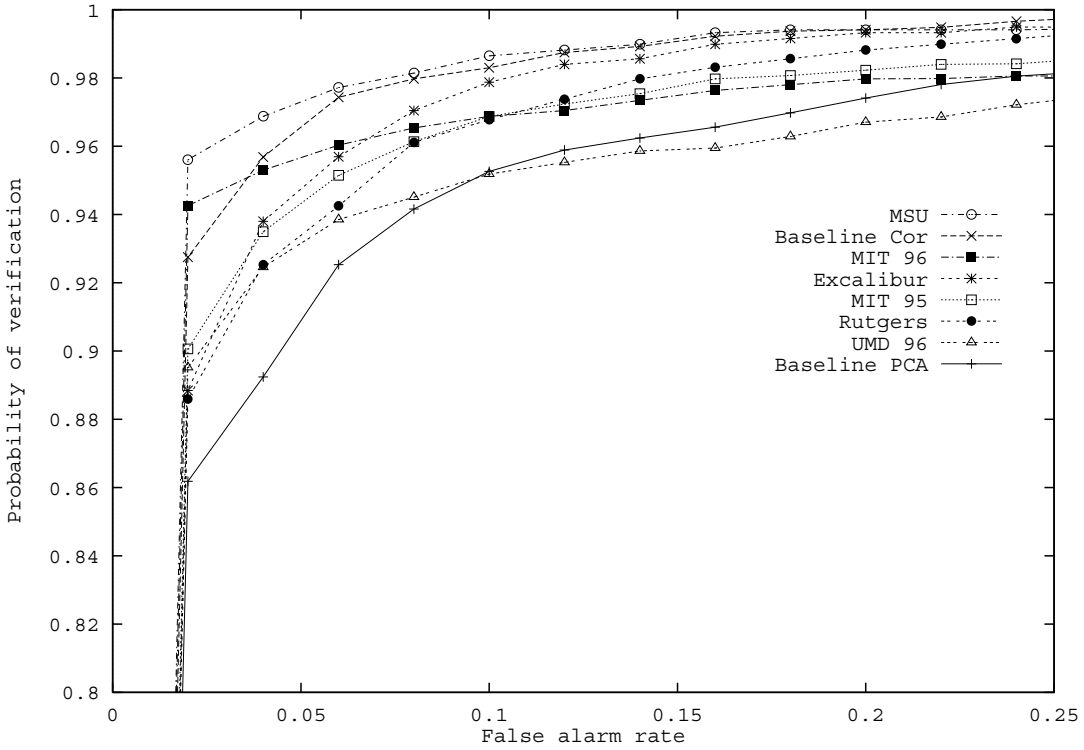
1995 and the probe images were acquired after January 1996. The gallery for the **FB**, duplicate I, and **fc** probes was the same and consisted of 1196 frontal images with one image person in the gallery (thus the gallery contained 1196 individuals). Also, none of the faces in the gallery images wore glasses. The gallery for the duplicate II probes was a subset of 864 images from the gallery for the other categories. The identification results presented in Phillips et al. [9,10] use the same gallery and probe sets for **FB**, **fc**, duplicate I, and duplicate II probe sets.

The verification results are reported on ROCs. The results are broken out by probe category and are presented in figures 2 to 5. Table 2 shows categories corresponding to the figures presenting these results, and size of the gallery and probe sets. For each probe category, there are two ROCs. First ROC reports results for the two baseline algorithms and the algorithms tested in September 1996. The second ROC reports for the two baseline algorithms, the algorithms tested in March 1997, and the UMD algorithm algorithm test in September 1996. Table 3 reports the equal error rates. We also report the average and best equal error rate for each probe category.

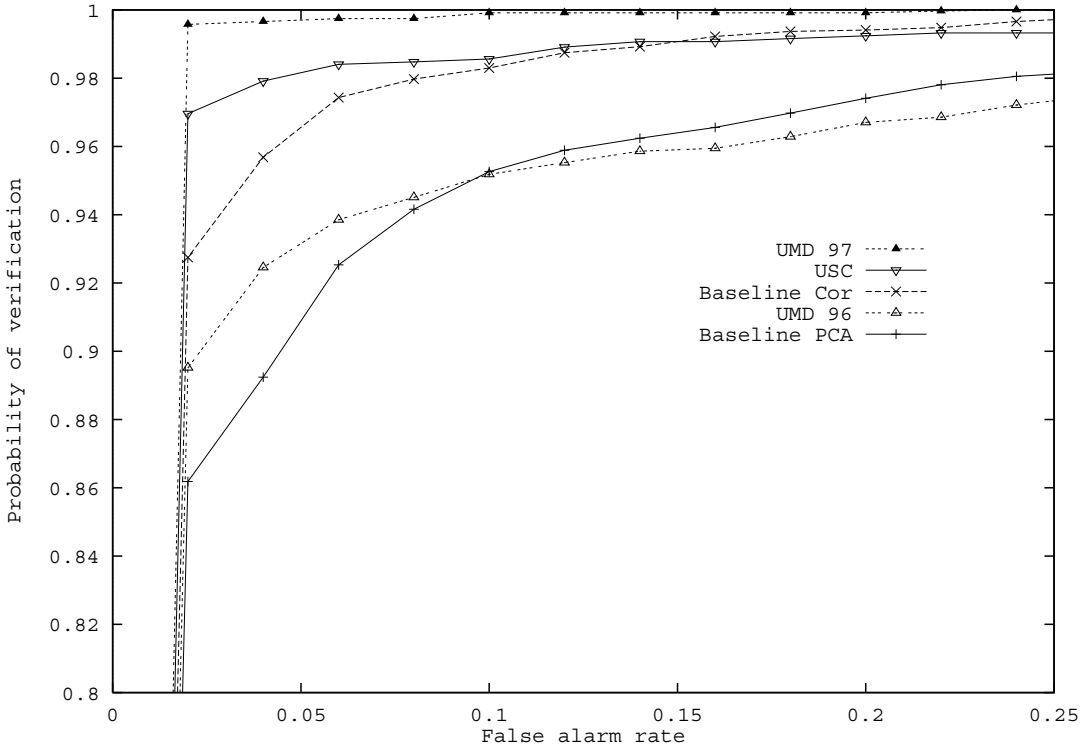
Performance of algorithms from a particular group will improve, and also, the performance level of face recognition algorithms in general improves over time. Thus, one should not comparing test results from different test dates. This illustrated by the improvement in performance of the UMD algorithm between September 1996 and March 1997. In consideration of this fact, we present results for September 1996 and March 1997 on different ROCs.

In figure 6, we compare the difficulty of different probe sets. Whereas, figure 3 reports verification performance for each algorithm, figure 6 shows a single curve that is an average of verification performance of all the algorithms. The average ROC is computed by averaging the P_V values for each P_F. Figure 6 reports performance for four categories of probes, **FB**, duplicate I, **fc**, and duplicate II.

Average performance provides an overall measure of the state-of-the-art. For

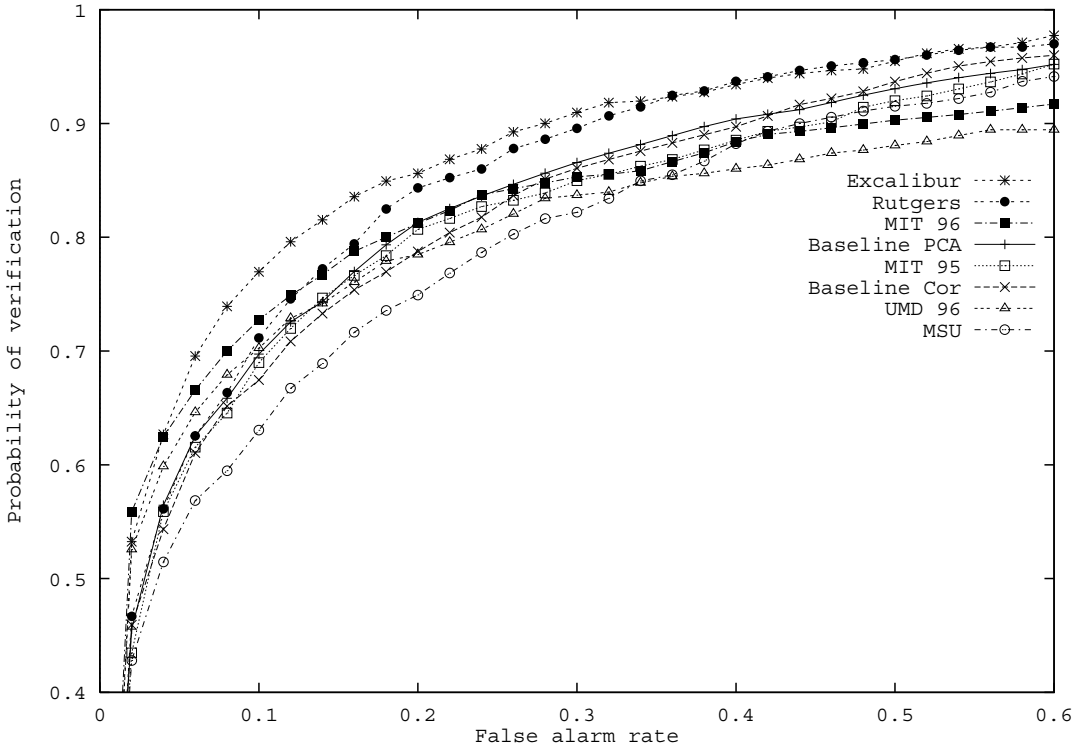


(a)

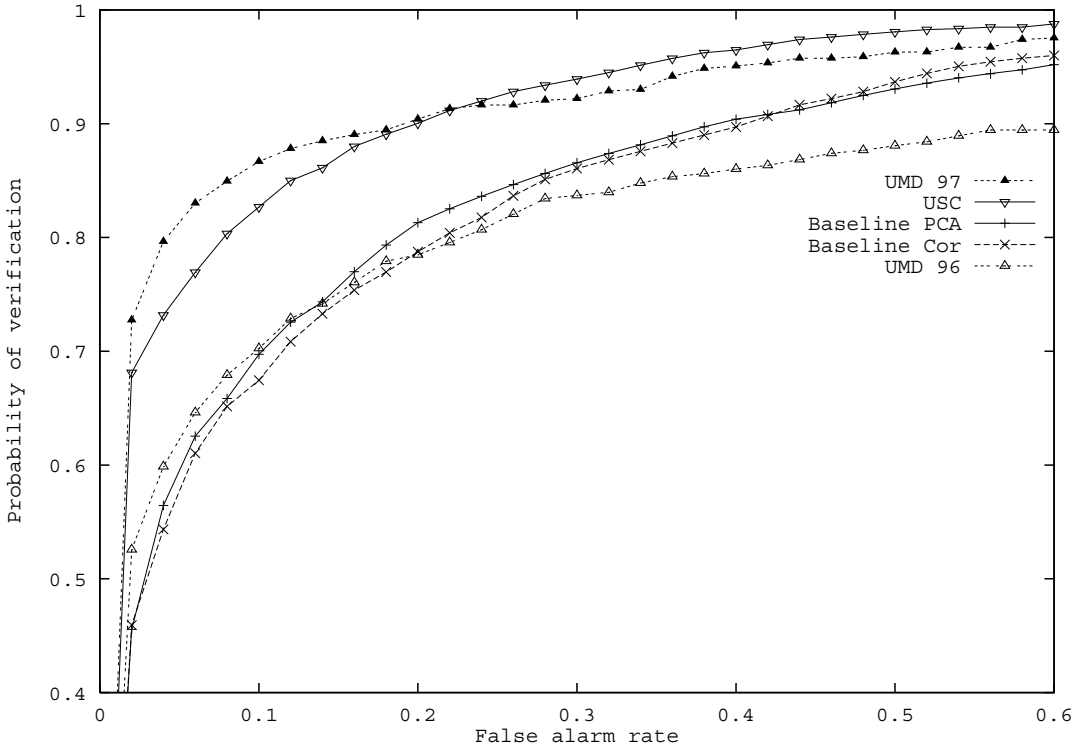


(b)

Fig. 2. Performance for **FB** probes. (a) Algorithms tested in September 1996. (b) Algorithms tested in March 1997.

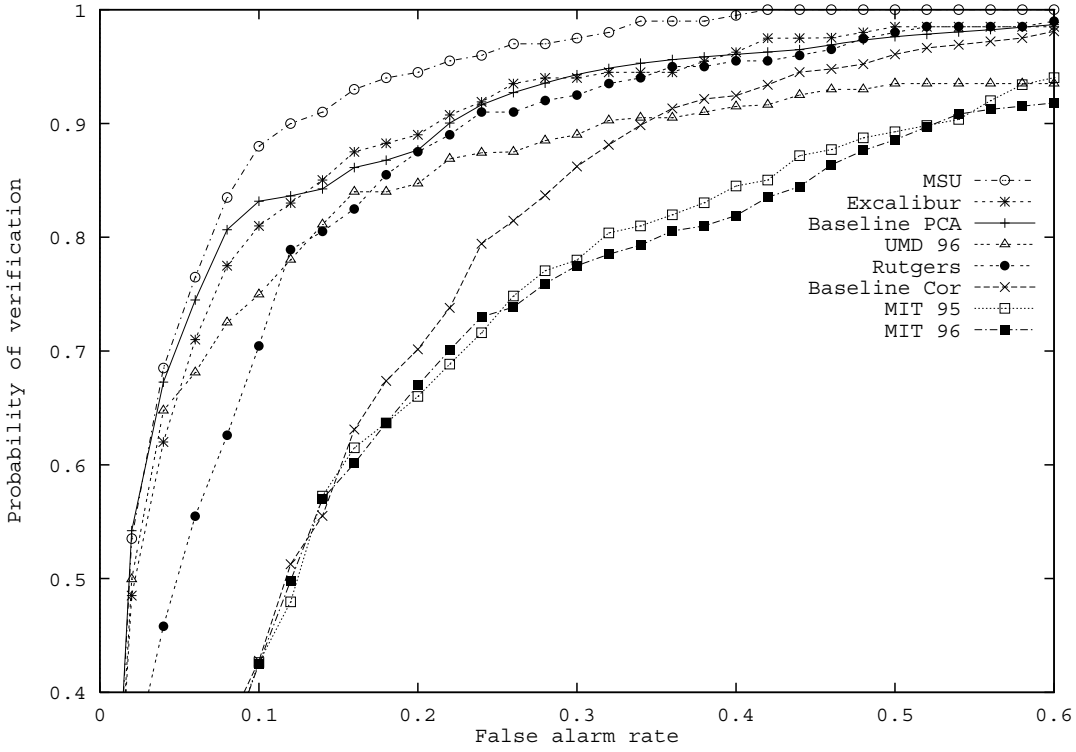


(a)

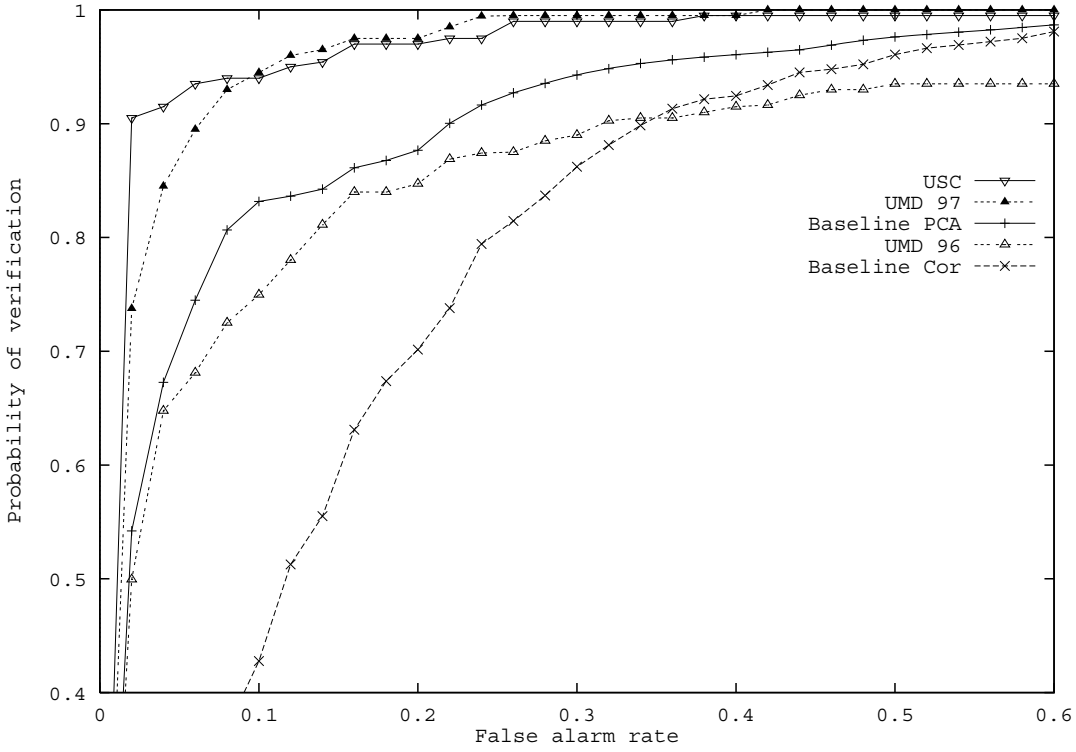


(b)

Fig. 3. Performance for duplicate I probes. (a) Algorithms tested in September 1996. (b) Algorithms tested in March 1997.

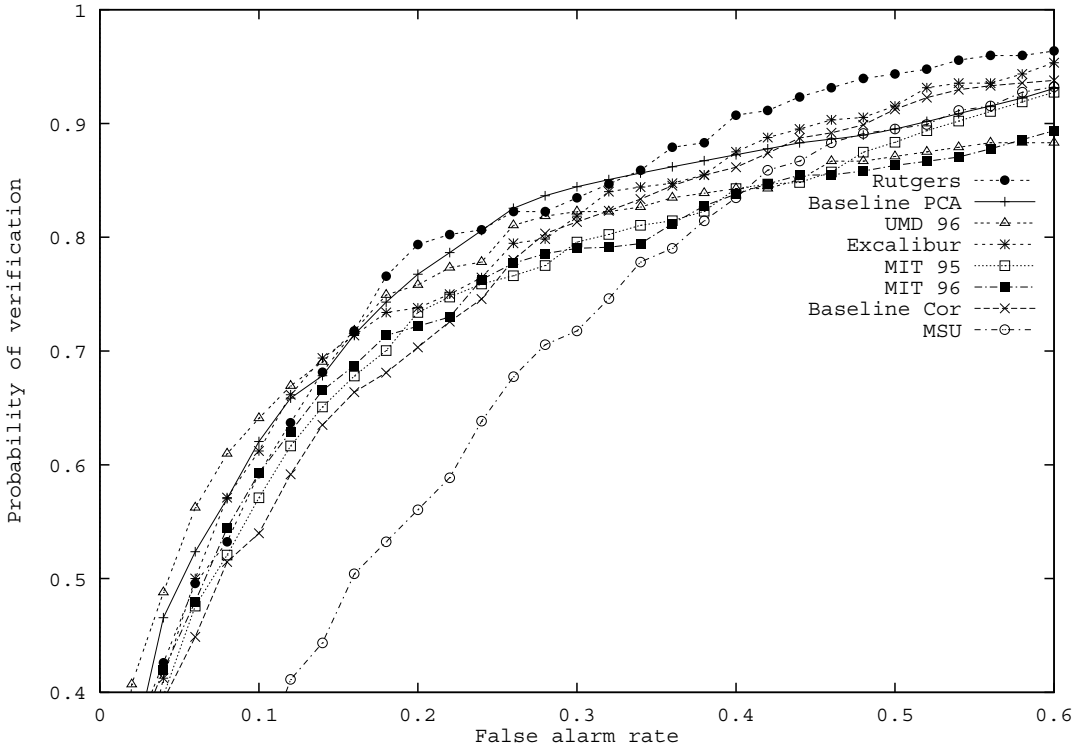


(a)

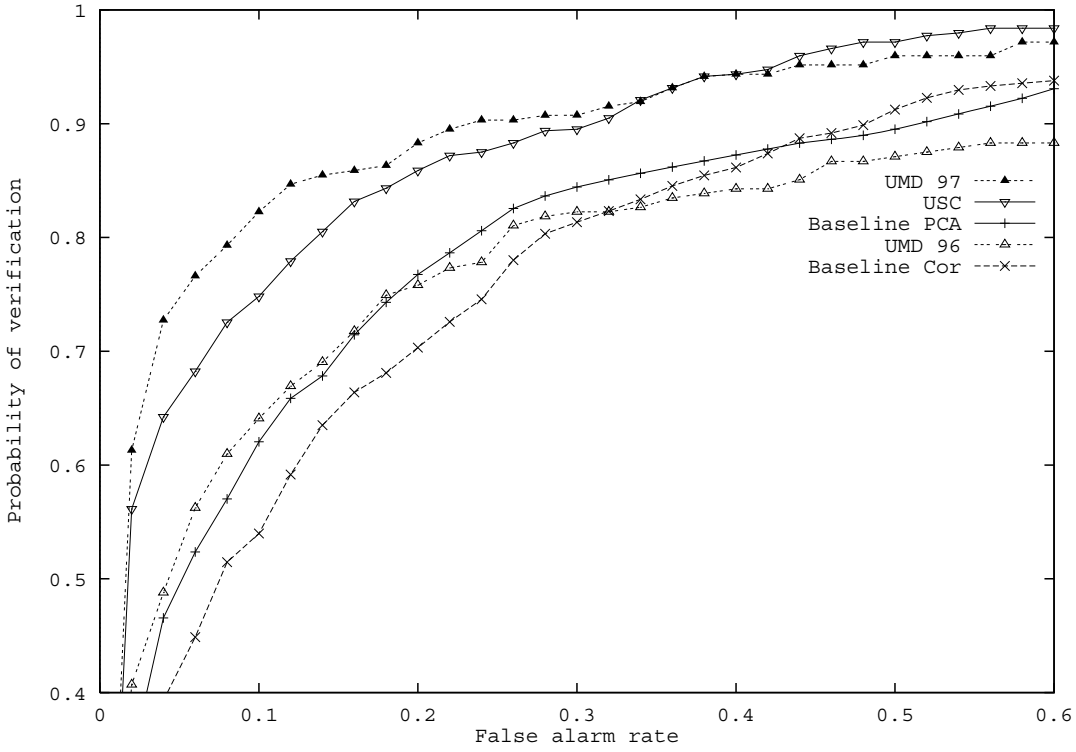


(b)

Fig. 4. Performance for **fc** probes. (a) Algorithms tested in September 1996. (b) Algorithms tested in March 1997.



(a)



(b)

Fig. 5. Performance for duplicate II probes. (a) Algorithms tested in September 1996. (b) Algorithms tested in March 1997.

Table 3
Equal error rates by probe category.

	Equal error rate by probe category(%)			
Algorithm	**FB**	Duplicate I	**fc**	Duplicate II
Baseline PCA	7	19	15	22
Baseline correlation	4	21	23	27
Excalibur	5	16	14	24
MIT Mar95	5	20	25	26
MIT Sep96	4	20	26	26
MSU	3	23	11	31
Rutgers	6	18	17	21
UMD Sep96	7	22	16	23
UMD Mar97	1	12	8	14
USC	2	14	6	17
Average	4	19	16	23
Minimum	1	12	6	14

applications, one is interested in the currently achievable upper performance bounds. In figure 7, we present the current upper bound on performance for each probe category in figure 6. For the upper bounds, we plotted the algorithm with minimum equal error rate in table 3.

5 Conclusion

We have devised a verification scoring procedure for the Sep96 FERET test, and reported results for this procedure. This allows for an independent assessment of face recognition algorithms in a key potential application.

This FERET test shows improvement in performance for both face recognition as a field and for individual algorithms. The improvement in the field is exhibited by the overall increase in performance of the algorithms tested between September 1996 and March 1997. Individual increase is demonstrated by the performance improvement of the U. of Maryland algorithm. This increase shows that algorithm performance should only be directly compared if they are tested at the same time.

Phillips et al. [9,10] presented identification results for the same algorithms on the same galleries and probes sets. The Sep96 MIT algorithm was the top

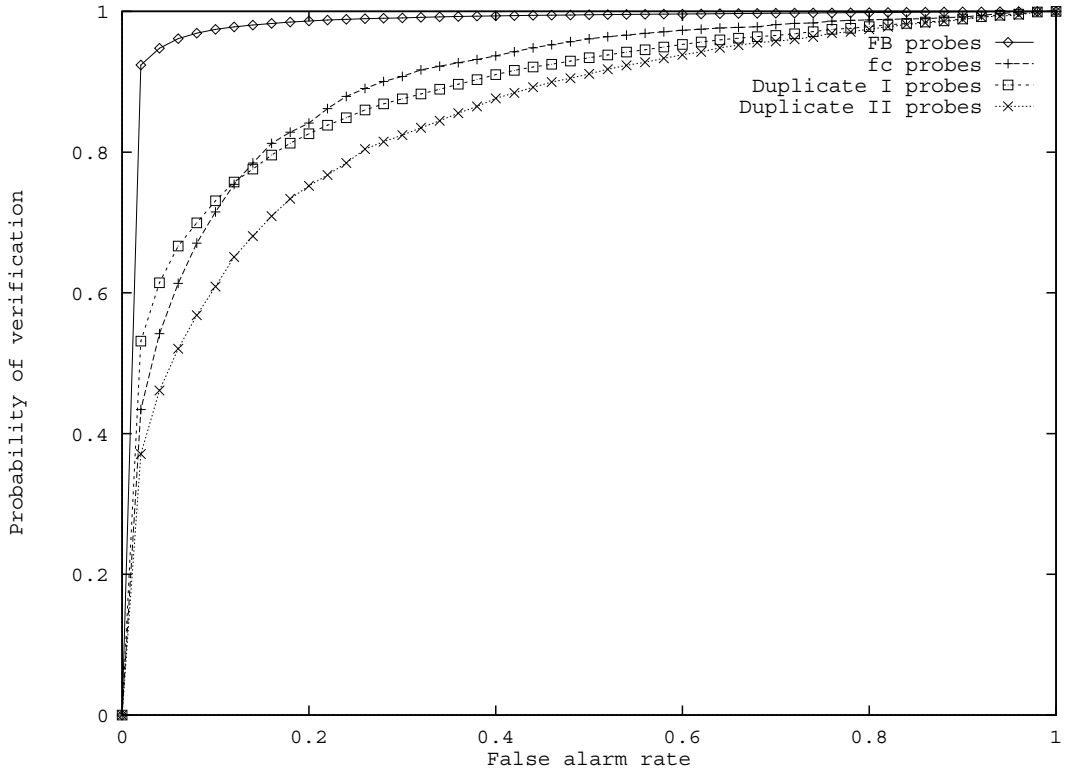


Fig. 6. Average performance of the algorithms on each probe category.

performer for the algorithms tested in September 1996. Among the algorithms tested in September 1996, no algorithm was among the top performers for all probe categories. This shows that relative performance on one task may not be predictive of relative performance on another task.

We broke out performance for four categories of probes. Each category represents a different degree of difficulty. To estimate the degree of difficulty for each category, we compared the average and current upper bounds of performance for each category. For average performance, our results rank **FB** probes as easiest, duplicate II probes as most difficult, and **fc** and duplicate I probes as tied in the middle. For current upper bounds, duplicate I probes are more difficult than **fc** probes. Our results show that we can expect that the best performance will be significantly better than the average performance. Upper bound performance for all probe categories is superior to all average performance categories except for **FB** probes.

The results in this paper show that algorithm development is a dynamic process and evaluations such as FERET make an important contribution to face recognition and computer vision. These evaluations let researchers know the strengths of their algorithms and where improvements could be made. By knowing their weaknesses, researchers know where to concentrate their efforts to improve performance.

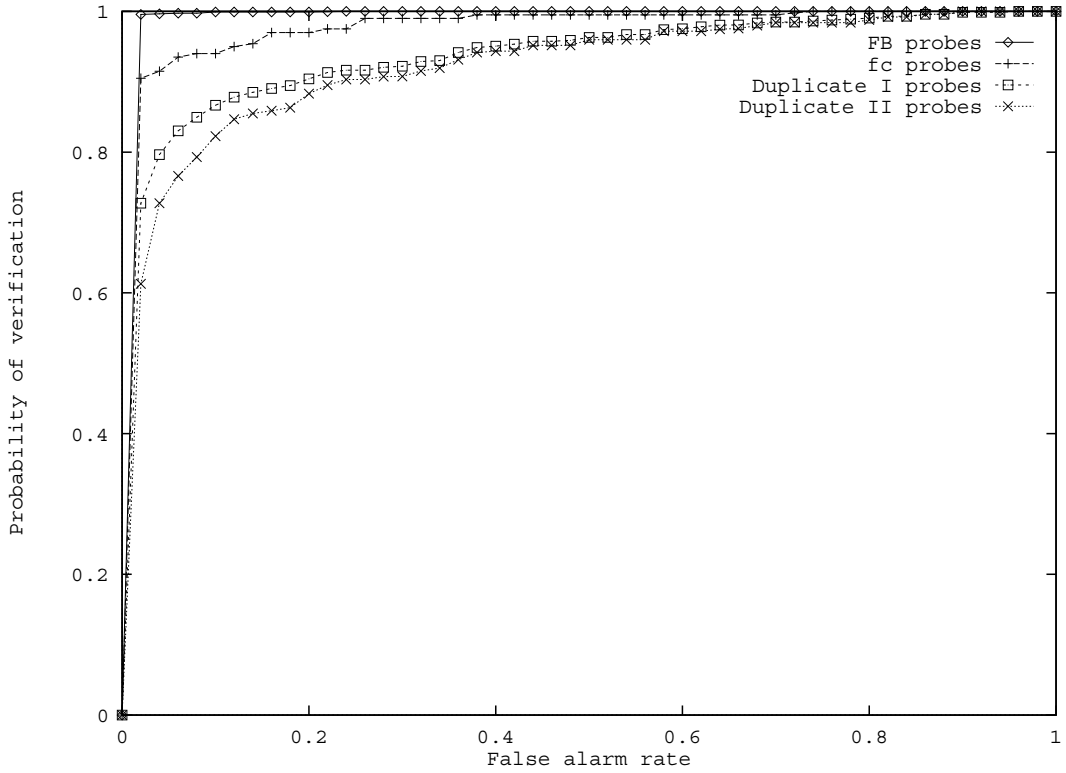


Fig. 7. Current upper bound on algorithm performance for each probe category.

References

[1] J. P. Egan. *Signal Detection Theory and ROC Analysis.* Academic Press, 1975.

[2] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *J. Opt. Soc. Am. A*, 14:1724–1733, August 1997.

[3] D. Green and J. Swets. *Signal Detection Theory and Psychophysics.* John Wiley & Sons Ltd., 1966.

[4] B. Moghaddam, C. Nastar, and A. Pentland. Bayesian face recognition using deformable intensity surfaces. In *Proceedings Computer Vision and Pattern Recognition 96*, pages 638–645, 1996.

[5] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *Proceedings of the Inter. Conf. on Computer Vision*, pages 786–793, 1995.

[6] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. *IEEE Trans. PAMI*, 17(7):696–710, 1997.

[7] H. Moon and P. J. Phillips. Analysis of PCA-based face recognition algorithms. In K. W. Bowyer and P. J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision.* IEEE Computer Society Press, Los Alamitos, CA, 1998.

[8] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eignspaces for face recognition. In *Proceedings Computer Vision and Pattern Recognition 94*, pages 84–91, 1994.

[9] P. J. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings Computer Vision and Pattern Recognition 97*, pages 137–143, 1997.

[10] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The feret evaluation. In P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.

[11] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, 16(5):295–306, 1998.

[12] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. PAMI*, 18(8):831–836, 1996.

[13] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[14] J. Wilder. Face recognition using transform coding of gray scale projection projections and the neural tree network. In R. J. Mammone, editor, *Artifical Neural Networks with Applications in Speech and Vision*, pages 520–536. Chapman Hall, 1994.

[15] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elasric bunch graph matching. *IEEE Trans. PAMI*, 17(7):775–779, 1997.

[16] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *3rd International Conference on Automatic Face and Gesture Recognition*, pages 336–341, 1998.