This is technical report NISTIR 6264.

The FERET Evaluation Methodology for Face-Recognition Algorithms *

P. Jonathon Phillips[1], Hyeonjoon Moon[2], Syed A. Rizvi[3], and Patrick J. Rauss,[4]

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

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

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

[4]US Army Research Laboratory
2800 Powder Mill Rd., Adelphi, MD 20783-1197

January 7, 1999

Abstract

Two of the most critical requirements in support of producing reliable face-recognition systems are a large database of facial images and a testing procedure to evaluate systems. The Face Recognition Technology (FERET) program has addressed both issues through the FERET database of facial images and the establishment of the FERET tests. To date, 14,126 images from 1199 individuals are included in the FERET database, which is divided into development and sequestered portions of the database. In September 1996, the FERET program administered the third in a series of FERET face-recognition tests. The primary objectives of the third test were to (1) assess the state of the art, (2) identify future areas of research, and (3) measure algorithm performance.

1 Introduction

Over the last decade, face recognition has become an active area of research in computer vision, neuroscience, and psychology. Progress has advanced to the point that face-recognition systems are being demonstrated in real-world settings [5]. The rapid development of face recognition is due to a combination of factors: active development of

*The work reported here is 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 was done while Jonathon Phillips was at the U.S. Army Research Laboratory (ARL). Jonathon Phillips acknowledges the support of the National Institute of Justice.

algorithms, the availability of a large database of facial images, and a method for evaluating the performance of face-recognition algorithms. The FERET database and evaluation methodology address the latter two points and are de facto standards. There have been three FERET evaluations with the most recent being the Sep96 FERET test.

The Sep96 FERET test provides a comprehensive picture of the state-of-the-art in face recognition from still images. This was accomplished by evaluating algorithms' ability on different scenarios, categories of images, and versions of algorithms. Performance was computed for identification and verification scenarios. In an identification application, an algorithm is presented with a face that it must identify the face; whereas, in a verification application, an algorithm is presented with a face and a claimed identity, and the algorithm must accept or reject the claim. In this paper, we describe the FERET database, the Sep96 FERET evaluation protocol, and present identification results. Verification results are presented in Rizvi et al. [8].

To obtain a robust assessment of performance, algorithms are evaluated against different categories of images. The categories are broken out by lighting changes, people wearing glasses, and the time between the acquisition date of the database image and the image presented to the algorithm. By breaking out performance into these categories, a better understanding of the face recognition field in general as well as the strengths and weakness of individual algorithms is obtained. This detailed analysis helps to assess which applications can be successfully addressed.

All face recognition algorithms known to the authors consist of two parts: (1) face detection and normalization and (2) face identification. Algorithms that consist of both parts are referred to as *fully automatic algorithms*, and those that consist of only the second part are *partially automatic algorithms.* The Sep96 test evaluated both fully and partially automatic algorithms. Partially automatic algorithms are given a facial image and the coordinates of the center of the eyes. Fully automatic algorithms are only given facial images.

The availability of the FERET database and evaluation methodology has made a significant difference in the progress of development of face-recognition algorithms. Before the FERET database was created, a large number of papers reported outstanding recognition results (usually $> 95\%$ correct recognition) on limited-size databases (usually < 50 individuals). (In fact, this is still true.) Only a few of these algorithms reported results on images utilizing a common database, let alone met the desirable goal of being evaluated on a standard testing protocol that included separate training and testing sets. As a consequence, there was no method to make informed comparisons among various algorithms.

The FERET database has made it possible for researchers to develop algorithms on a common database and to report results in the literature using this database. Results reported in the literature do not provide a direct comparison among algorithms because each researcher reported results using different assumptions, scoring methods, and images. The independently administered FERET test allows for a direct quantitative assessment of the relative strengths and weaknesses of different approaches.

More importantly, the FERET database and tests clarify the current state of the art in face recognition and point out general directions for future research. The FERET tests allow the computer vision community to assess overall strengths and weaknesses in the field, not only on the basis of the performance of an individual algorithm, but in addition on the aggregate performance of all algorithms tested. Through this type of assessment, the community learns in an unbiased and open manner of the important technical problems to be addressed, and how the community is progressing toward solving these problems.

2 Background

The first FERET tests took place in August 1994 and March 1995 (for details of these tests and the FERET database and program, see Phillips *et al* [5, 6] and Rauss *et al* [7]); The FERET database collection began in September 1993 along with the FERET program.

The August 1994 test established, for the first time, a performance baseline for face-recognition algorithms. This test was designed to measure performance on algorithms that could automatically locate, normalize, and identify faces from a database. The test consisted of three subtests, each with a different gallery and probe set. The *gallery* contains the set of known individuals. An image of an unknown face presented to the algorithm is called a *probe*, and the collection of probes is called the *probe set*. The first subtest examined the ability of algorithms to recognize faces from a gallery of 316 individuals. The second was the false-alarm test, which measured how well an algorithm rejects faces not in the gallery. The third baselined the effects of pose changes on performance.

The second FERET test, that took place in March 1995, measured progress since August 1994 and evaluated algorithms on larger galleries. The March 1995 evaluation consisted of a single test with a gallery of 817 known individuals. One emphasis of the test was on probe sets that contained duplicate images. A *duplicate* is defined as an image of a person whose corresponding gallery image was taken on a different date.

The FERET database is designed to advance the state of the art in face recognition, with the images collected directly supporting both algorithm development and the FERET evaluation tests. The database is divided into a development set, provided to researchers, and a set of sequestered images for testing. The images in the development set are representative of the sequestered images.

The facial images were collected in 15 sessions between August 1993 and July 1996. Collection sessions lasted one or two days. In an effort to maintain a degree of consistency throughout the database, the same physical setup and location was used in each photography session. However, because the equipment had to be reassembled for each session, there was variation from session to session (figure 1).

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). The remaining images were collected at various aspects between right and left profile. To add simple variations to the database, photographers sometimes took a second set of images, for which the subjects were asked to put on their glasses and/or pull their hair back. Sometimes a second set of images of a person was taken on a later date; such a set of images is referred to as a duplicate set. Such duplicates sets result in variations in scale, pose, expression, and illumination of the face.

By July 1996, 1564 sets of images were in the database, consisting of 14,126 total images. The database contains 1199 individuals and 365 duplicate sets of images. For

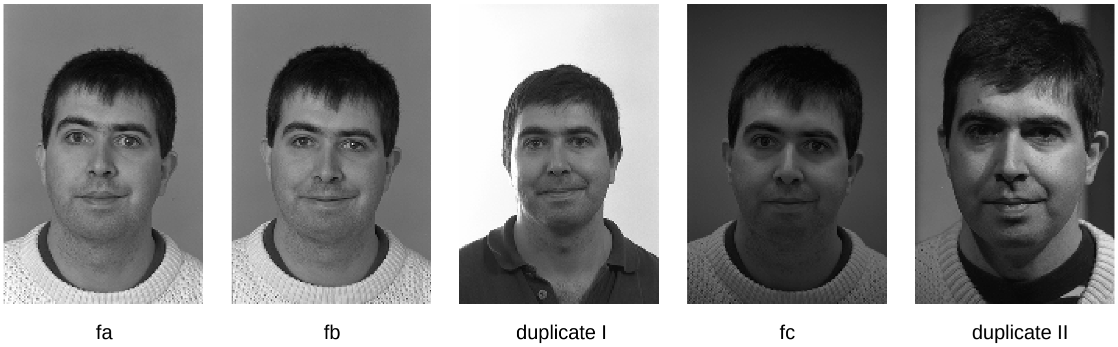


Figure 1: Examples of different categories of probes (image). The duplicate I image was taken within one year of the **fa** image and the duplicate II and **fa** images were taken at least one year apart.

some people, over two years elapsed between their first and most recent sittings, with some subjects being photographed multiple times (figure 1). The development portion of the database consisted of 503 sets of images, and was released to researchers. The remaining images were sequestered by the Government.

3 Test Design

3.1 Test Design Principles

The FERET Sep96 evaluation protocol was designed to assess the state of the art, advance the state of the art, and point to future directions of research. To succeed at this, the test design must solve the *three bears problem*. The test cannot be neither too hard nor too easy. If the test is too easy, the testing process becomes an exercise in "tuning" existing algorithms. If the test is too hard, the test is beyond the ability of existing algorithmic techniques. The results from the test are poor and do not allow for an accurate assessment of algorithmic capabilities.

The solution to the three bears problem is through the selection of images in the test set and the testing protocol. Tests are administered using a testing protocol that states the mechanics of the tests and the manner in which the test will be scored. In face recognition, the protocol states the number of images of each person in the test, how the output from the algorithm is recorded, and how the performance results are reported.

The characteristics and quality of the images are major factors in determining the difficulty of the problem being evaluated. For example, if faces are in a predetermined position in the images, the problem is different from that for images in which the faces can be located anywhere in the image. In the FERET database, variability was introduced by the inclusion of images taken at different dates and locations (see section 2). This resulted in changes in lighting, scale, and background.

The testing protocol is based on a set of design principles. Stating the design principle allows one to assess how appropriate the FERET test is for a particular face recognition algorithm. Also, design principles assist in determining if an evaluation methodology for testing algorithm(s) for a particular application is appropriate. Before discussing the design principles, we state the evaluation protocol.

In the testing 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 consists of unknown facial images to be identified. For each image q_i in the query set $\mathcal{Q}$, an algorithm reports a similarity $s_i(k)$ between q_i and each image t_k in the target set $\mathcal{T}$. The testing protocol is designed so that each algorithm can use a different similarity measure and we do not compare similarity measures from different algorithms. The key property of the new protocol, which allows for greater flexibility in scoring, is that for any two images q_i and t_k, we know $s_i(k)$.

This flexibility allows the evaluation methodology to be robust and comprehensive; it is achieved by computing scores for virtual galleries and probe sets. A gallery $\mathcal{G}$ is a virtual gallery if $\mathcal{G}$ is a subset of the target set, i.e., $\mathcal{G} \subset \mathcal{T}$. Similarly, $\mathcal{P}$ is a virtual probe set if $\mathcal{P} \subset \mathcal{Q}$. For a given gallery $\mathcal{G}$ and probe set $\mathcal{P}$, the performance scores are computed by examination of similarity measures $s_i(k)$ such that $q_i \in \mathcal{P}$ and $t_k \in \mathcal{G}$.

The virtual gallery and probe set technique allows us to characterize algorithm performance by different categories of images. The different categories include (1) rotated images, (2) duplicates taken within a week of the gallery image, (3) duplicates where the time between the images is at least one year, (4) galleries containing one image per person, and (5) galleries containing duplicate images of the same person. We can create a gallery of 100 people and estimate an algorithm's performance by recognizing people in this gallery. Using this as a starting point, we can then 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 with the different galleries.

To take full advantage of virtual galleries and probe sets, we selected multiple images of the same person and placed them into the target and query sets. If such images were marked as the same person, the algorithms being tested could use the information in the evaluation process. To prevent this from happenning, we require that each image in the target set be treated as an unique face. (In practice, this condition is enforced by giving every image in the target and query set a unique random identification.) This is the first design principle.

The second design principle is that training is completed prior to the start of the test. This forces each algorithm to have a general representation for faces, not a representation tuned to a specific gallery. Without this condition, virtual galleries would not be possible.

For algorithms to have a general representation for faces, they must be gallery (class) insensitive. Examples are algorithms based on normalized correlation or principal component analysis (PCA). An algorithm is class sensitive if the representation is tuned to a specific gallery. Examples are straight forward implementation of Fisher discriminant analysis [1, 9]. Fisher discriminant algorithms were adapted to class insensitive testing methodologies by Zhao *et al* [13, 14], with performance results of these extensions being reported in this paper.

The third design rule is that all algorithms tested compute a similarity measure between two facial images; this similarity measure was computed for all pairs of images between the target and query sets. Knowing the similarity score between all pairs of

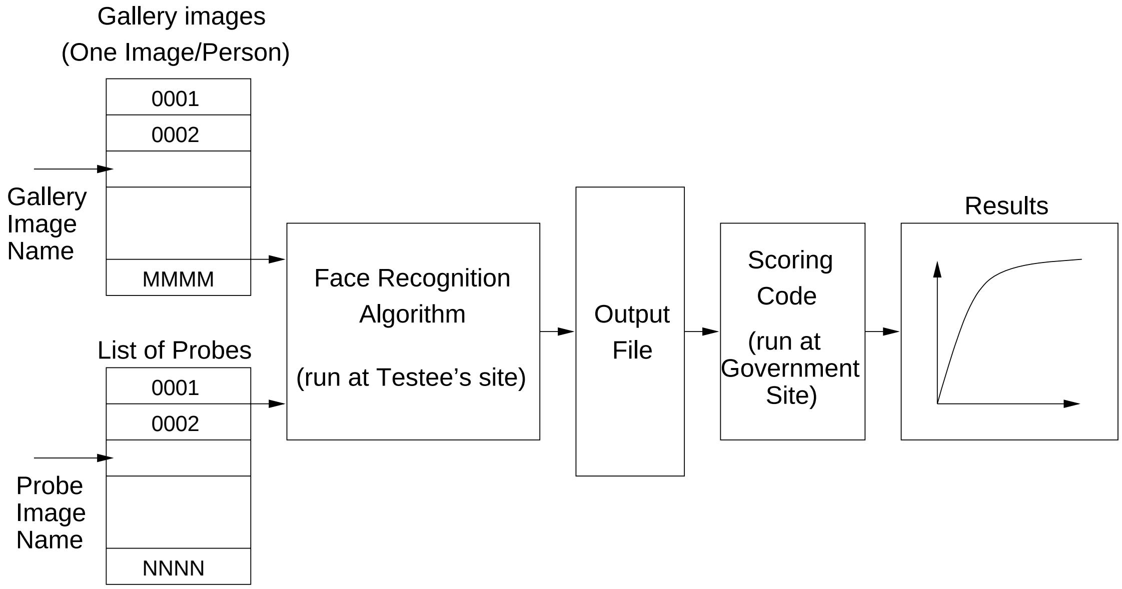


Figure 2: Schematic of the FERET testing procedure

images from the target and query sets allows for the construction of virtual galleries and probe sets.

3.2 Test Details

In the Sep96 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. We designed the digitally modified images to test the effects of illumination and scale. (Results from the rotated and digitally modified images are not reported here.) 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 $s_i(\cdot)$. Since the target set is a subset of the query set, the test output contains the similarity score between all images in the target set.

There were two versions of the Sep96 test. The target and query sets were the same for each version. The first version tested partially automatic algorithms by providing them with a list of images in the target and query sets, and the coordinates of the center of the eyes for images in the target and query sets. In the second version of the test, the coordinates of the eyes were not provided. By comparing the performance between the two versions, we estimate performance of the face-locating portion of a fully automatic algorithm at the system level.

The test was administered at each group's site under the supervision of one of the authors. Each group had three days to complete the test on less than 10 UNIX workstations (this limit was not reached). We did not record the time or number of workstations because execution times can vary according to the type of machines used, machine and network configuration, and the amount of time that the developers spent optimizing their code (we wanted to encourage algorithm development, not code optimization). (We imposed the time limit to encourage the development of algorithms that could be incorporated into operational, fieldable systems.)

The images contained in the gallery and probe sets consisted of images from both the developmental and sequestered portions of the FERET database. Only images from the FERET database were included in the test; however, algorithm developers were not prohibited from using images outside the FERET database to develop or tune parameters in their algorithms.

The FERET test is designed to measure laboratory performance. The test is not concerned with speed of the implementation, real-time implementation issues, and speed and accuracy trade-offs. These issues and others, need to be addressed in an operational, fielded system, were beyond the scope of the Sep96 FERET test.

Figure 2 presents a schematic of the testing procedure. To ensure that matching was not done by file name, we gave the images random names. The nominal pose of each face was provided to the testee.

4 Decision Theory and Performance Evaluation

The basic models for evaluating the performance of an algorithm are the closed and open universes. In the closed universe, every probe is in the gallery. In an open universe, some probes are not in the gallery. Both models reflect different and important aspects of face-recognition algorithms and report different performance statistics. The open universe models verification applications. The FERET scoring procedures for verification is given in Rizvi et al [8].

The closed-universe model allows one to ask how good an algorithm is at identifying a probe image; the question is not always "is the top match correct?" but "is the correct answer in the top n matches?" This lets one know how many images have to be examined to get a desired level of performance. The performance statistics are reported as cumulative match scores. The rank is plotted along the horizontal axis, and the vertical axis is the percentage of correct matches. The cumulative match score can be calculated for any subset of the probe set. We calculated this score to evaluate an algorithm's performance on different categories of probes, i.e., rotated or scaled probes.

The computation of an identification score is quite simple. Let $\mathcal{P}$ be a probe set and $|\mathcal{P}|$ the size of $\mathcal{P}$. We score probe set $\mathcal{P}$ against gallery $\mathcal{G}$, where $\mathcal{G} = \{g_1, ..., g_M\}$ and $\mathcal{P}= \{p_1, ..., p_N\}$ by comparing the similarity scores $s_i(\cdot)$ such that $p_i \in \mathcal{P}$ and $g_k \in \mathcal{G}$. For each probe image $p_i \in \mathcal{P}$, we sort $s_i(\cdot)$ for all gallery images $g_k \in \mathcal{G}$. We assume that a smaller similarity score implies a closer match. If g_k and p_i are the same image, then $s_i(k) = 0$. The function $id(i)$ gives the index of the gallery image of the person in probe p_i, i.e., p_i is an image of the person in $g_{id(i)}$. A probe p_i is correctly identified if $s_i(id(i))$ is the smallest scores for $g_k \in \mathcal{G}$. A probe p_i is in the top k if $s_i(id(i))$ is one of the k-th smallest score $s_i(\cdot)$ for gallery $\mathcal{G}$. Let R_k denote the number of probes in the top k. We reported $R_k/|\mathcal{P}|$, the fraction of probes in the top k. As an example, let $k = 5$, $R_5 = 80$ and $|\mathcal{P}| = 100$. Based on the formula, the performance score for R_5 is $80/100 = 0.8$.

In reporting identification performance results, we state the size of the gallery and the number of probes scored. The size of the gallery is the number of different faces (people) contained in the images that are in the gallery. For all results that we report, there is one image per person in the gallery, thus, the size of the gallery is also the number of images in the gallery. The number of probes scored (also, size of the probe set) is $|\mathcal{P}|$. The probe set may contain more than one image of a person and the probe set may not contain an image of everyone in the gallery. Every image in the probe set has a corresponding image in the gallery.

5 Latest Test Results

The Sep96 FERET test was designed to measure algorithm performance for identification and verification tasks. Both tasks are evaluated on the same sets of images. We report the results for 12 algorithms that includes 10 partially automatic algorithms and 2 fully automatic algorithms. The test was administered in September 1996 and March 1997 (see table 1 for details of when the test was administered to which groups and which version of the test was taken). Two of these algorithms were developed at the MIT Media Laboratory. The first was the same algorithm that was tested in March 1995. This algorithm was retested so that improvement since March 1995 could be measured. The second algorithm was based on more recent work [2, 3]. Algorithms were also tested from Excalibur Corp. (Carlsbad, CA), Michigan State University (MSU) [9, 14], Rutgers University [11], University of Southern California (USC) [12], and two from University of Maryland (UMD) [1, 13, 14]. The first algorithm from UMD was tested in September 1996 and a second version of the algorithm was tested in March 1997. For the fully automatic version of test, algorithms from MIT and USC were evaluated.

The final two algorithms were our implementation of normalized correlation and a principal components analysis (PCA) based algorithm [4, 10]. 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.

5.1 Partially automatic algorithms

We report identification scores for four categories of probes. The first probe category was the **FB** probes (fig 3). For each set of images, there were two frontal images. One of the images was randomly placed in the gallery, and the other image 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 contained all duplicate frontal images in the FERET database for 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). 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 1995 and the probe images were acquired after January 1996.

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 [2, 3]	●		
	U. of So. California (USC) [12]		●	
Eye Coordinates Given	Baseline PCA [4, 10]			●
	Baseline Correlation			●
	Excalibur Corp.	●		
	MIT Media Lab	2		
	Michigan State U. [9, 14]	●		
	Rutgers U. [11]	●		
	U Maryland [1, 13, 14]	●	●	
	USC		●	

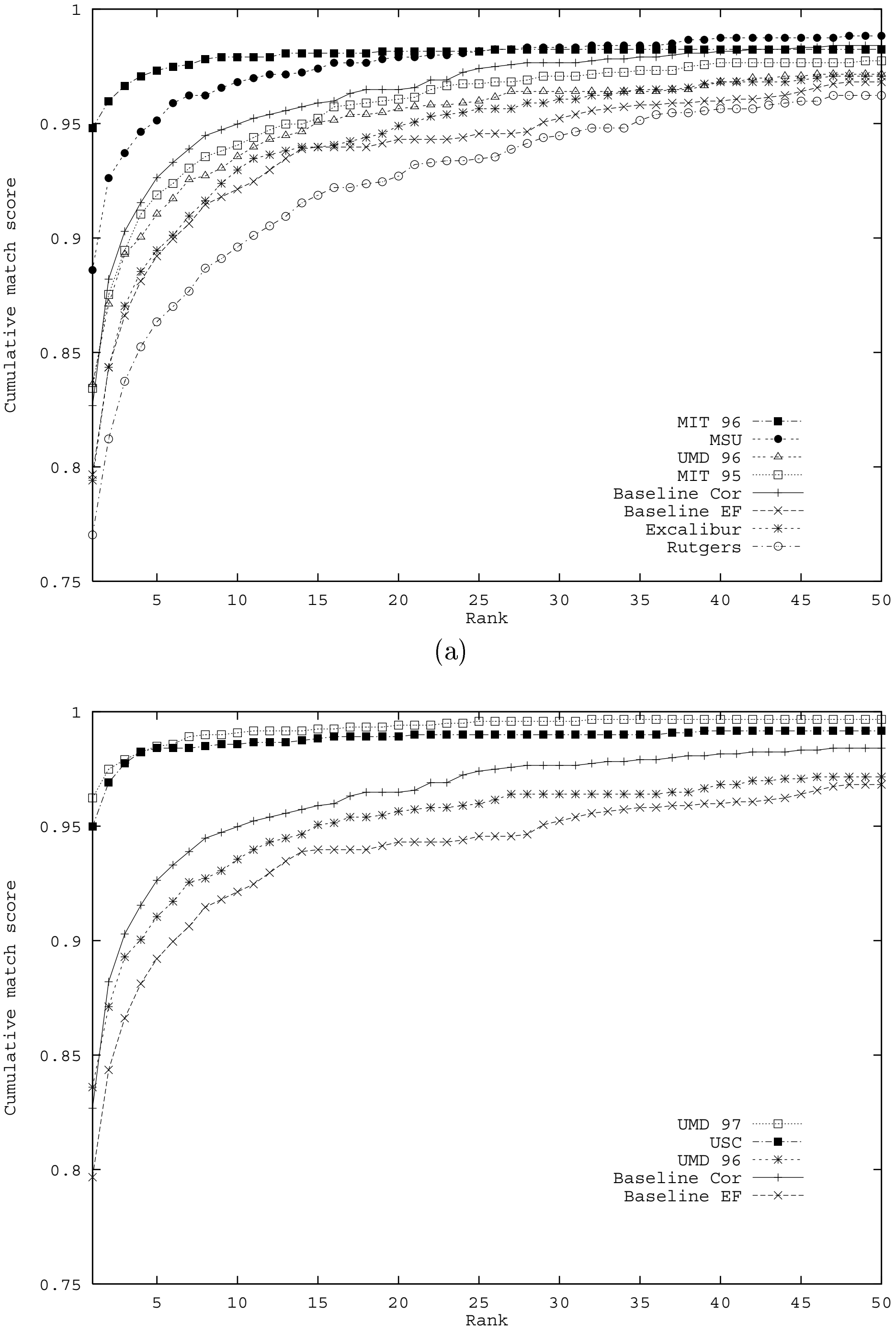
The gallery for the **FB**, duplicate I, and **fc** probes was the same and consisted of 1196 frontal images with one image person in the gallery (thus the gallery contained 1196 individuals). Also, none of the faces in the gallery images wore glasses. The gallery for duplicate II probes was a subset of 864 images from the gallery for the other categories.

The results for identification are reported as cumulative match scores. Table 2 shows the categories corresponding to the figures presenting the results, type of results, and size of the gallery and probe sets (figs 3 to 6).

In figures 7 and 8, we compare the difficulty of different probe sets. Whereas, figure 4 reports identification performance for each algorithm, figure 7 shows a single curve that is an average of the identification performance of all algorithms for each probe category. For example, the first ranked score for duplicate I probe sets is computed from an average of the first ranked score for all algorithms in figure 4. In figure 8, we presented current upper bound for performance on partially automatic algorithms for each probe category. For each category of probe, figure 8 plots the algorithm with the highest top rank score (R_1). Figures 7 and 8 reports performance of four categories of probes, **FB**, duplicate I, **fc**, duplicate II.

Table 2: Figures reporting results for partially automatic algorithms. Performance is broken out by probe category.

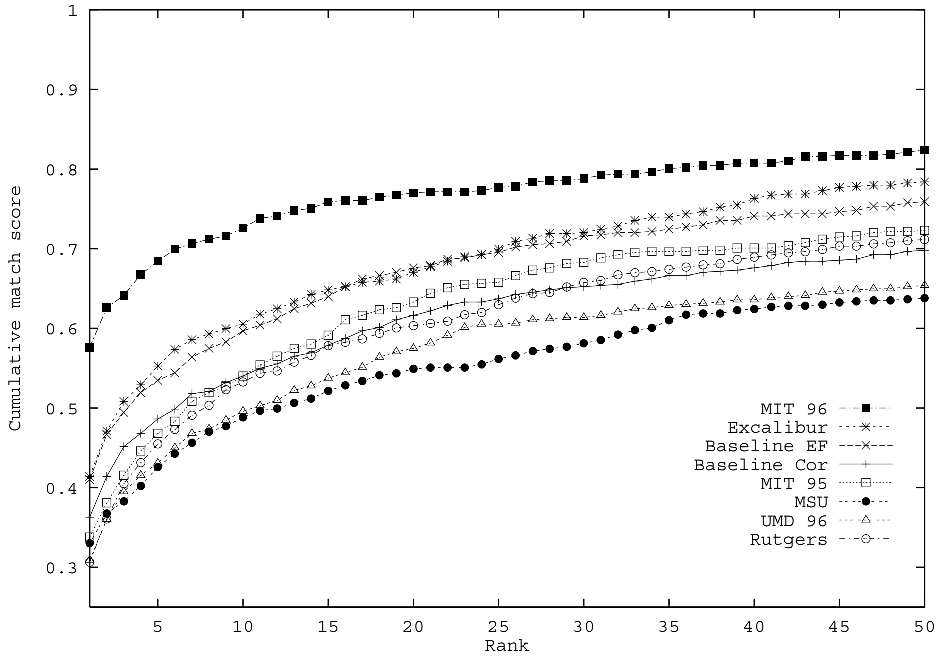
Figure no.	Probe Category	Gallery size	Probe set size
3	**FB**	1196	1195
4	duplicate I	1196	722
5	**fc**	1196	194
6	duplicate II	864	234



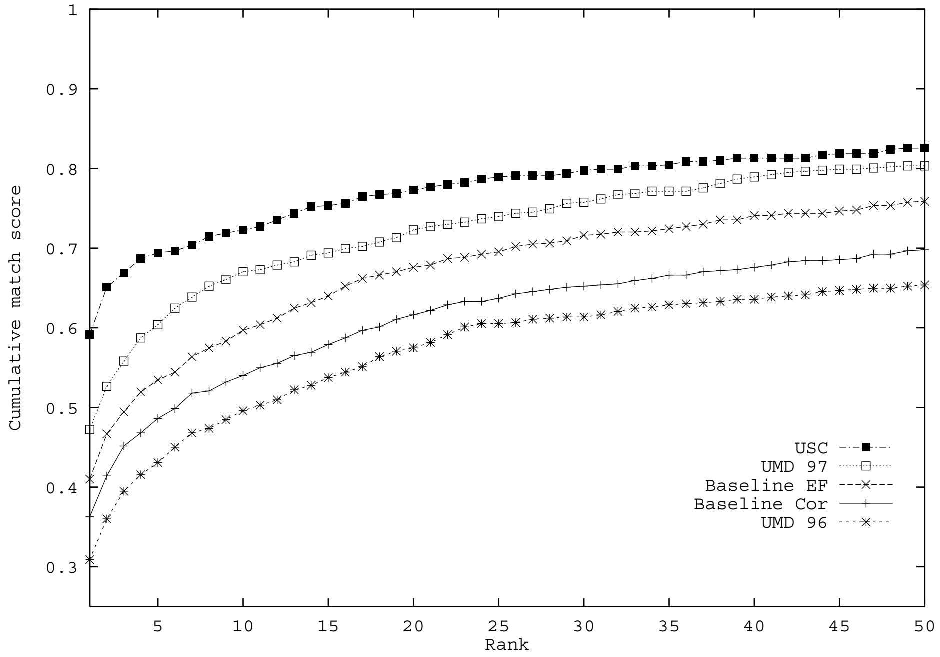
(a)

(b)

Figure 3: Identification performance against **FB** probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.

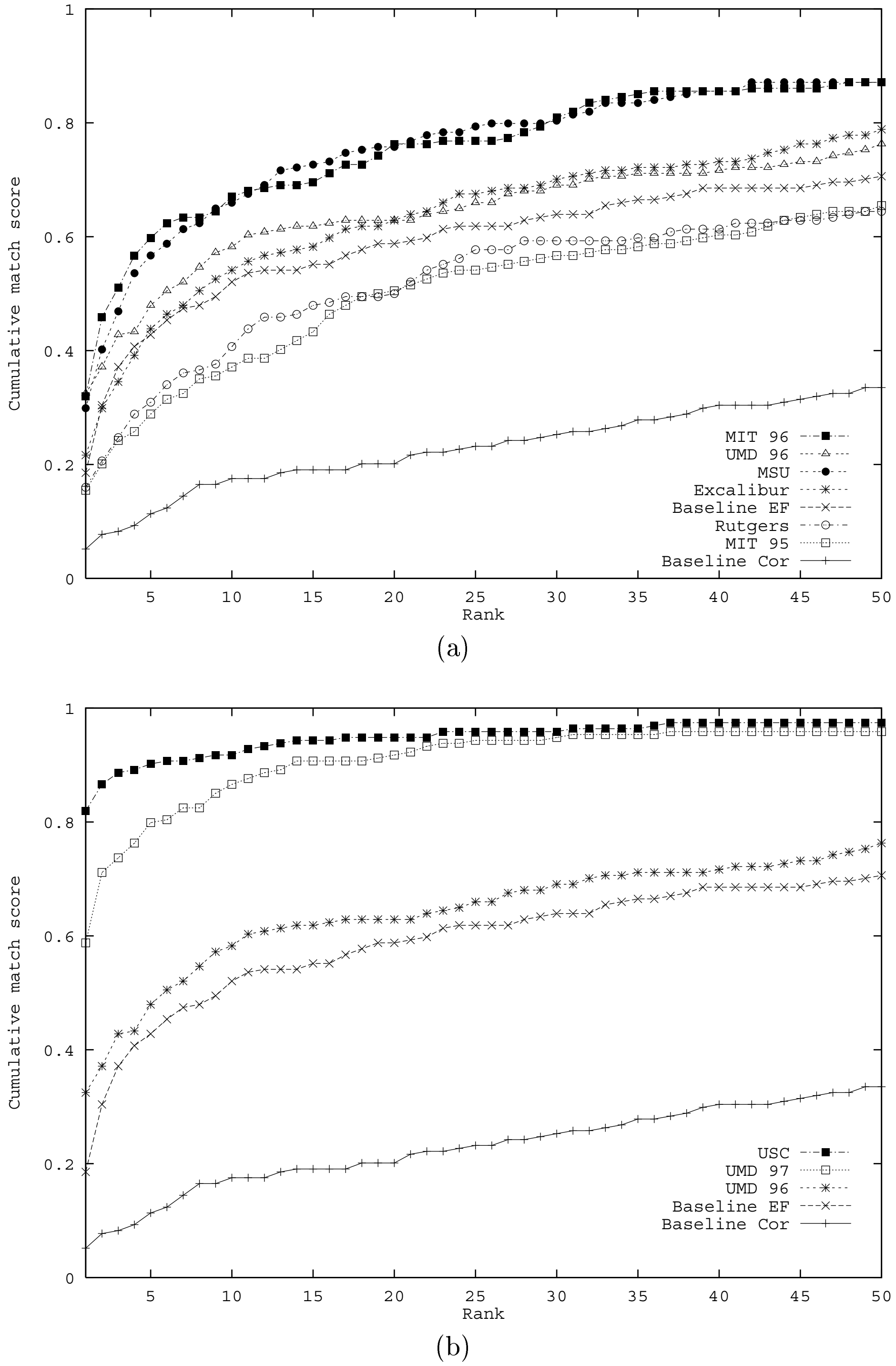


(a)



(b)

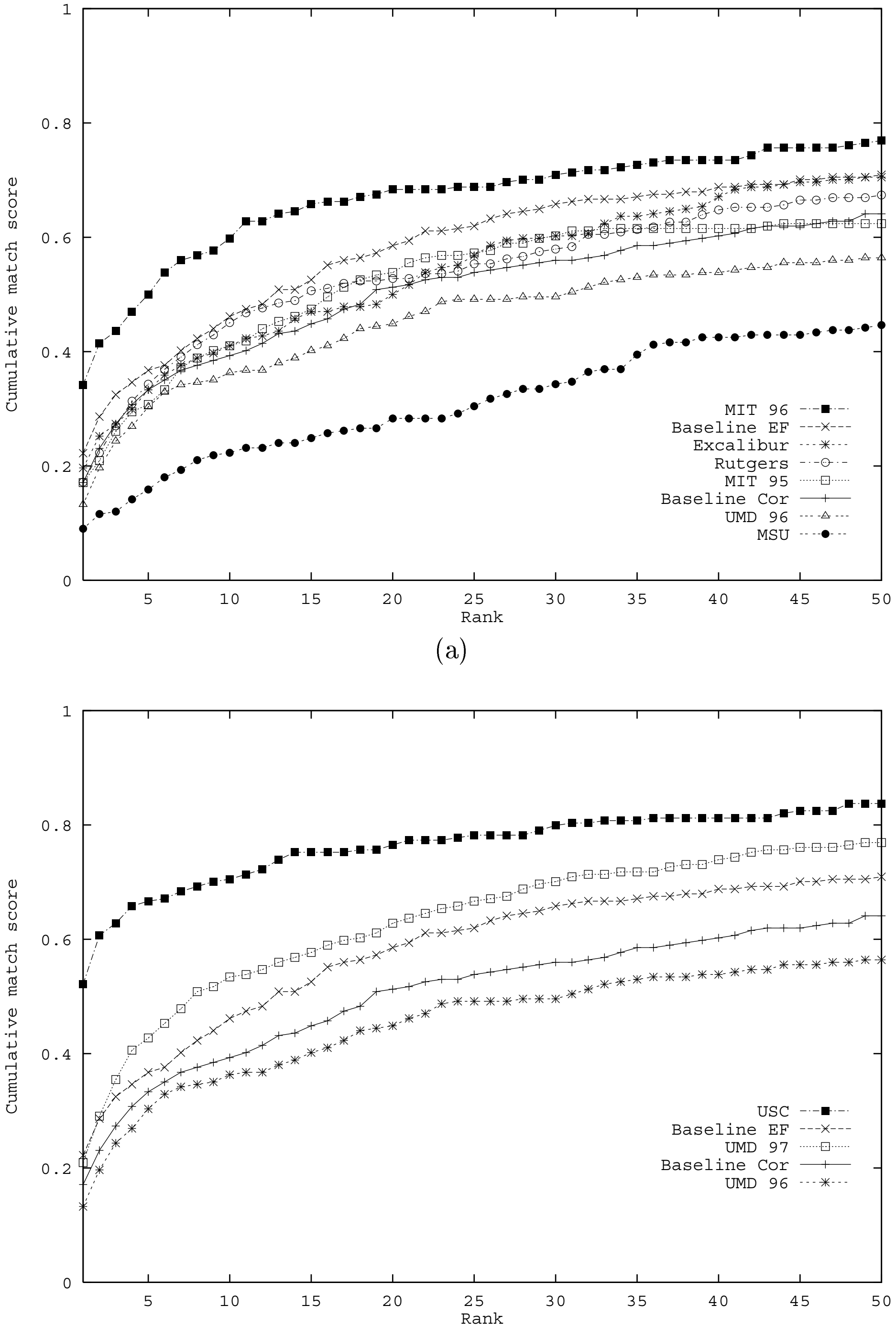
Figure 4: Identification performance against all duplicate I probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.



(a)

(b)

Figure 5: Identification performance against **fc** probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.



(a)

(b)

Figure 6: Identification performance against duplicate II probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.

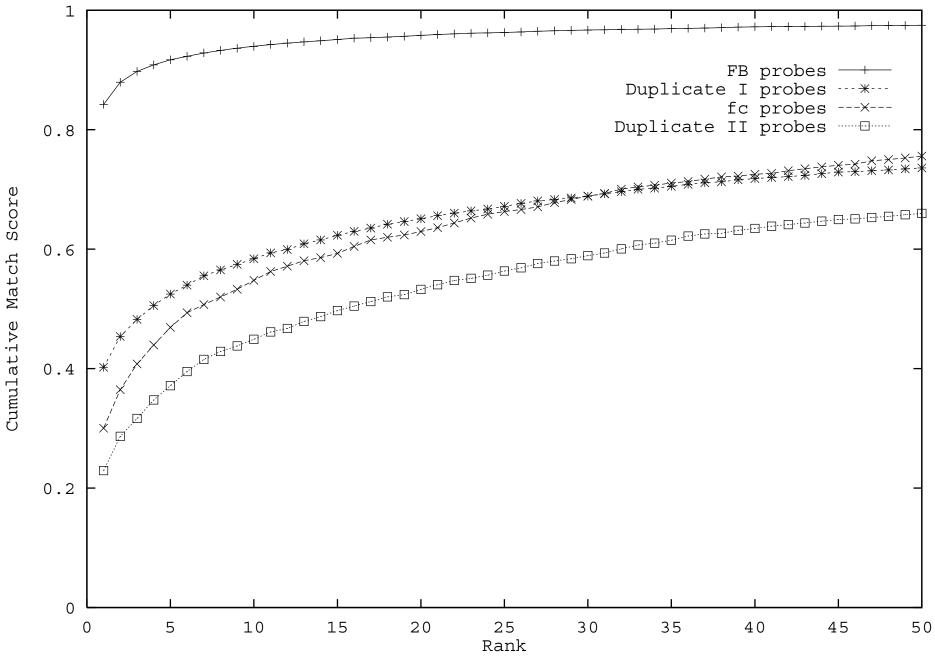


Figure 7: Average identification performance of partially automatic algorithms on each probe category.

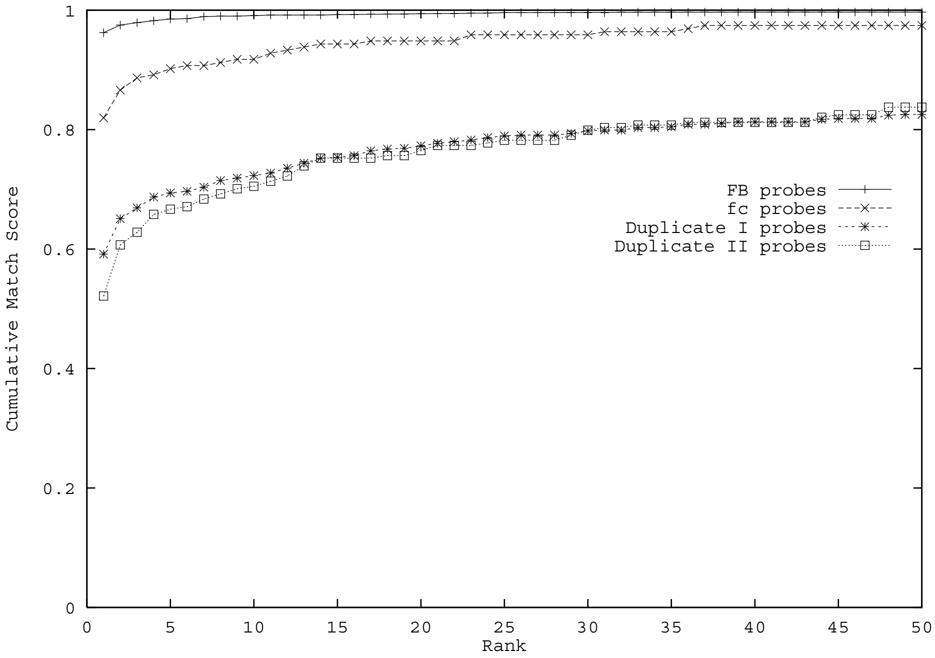


Figure 8: Current upper bound identification performance of partially automatic algorithm for each probe category.

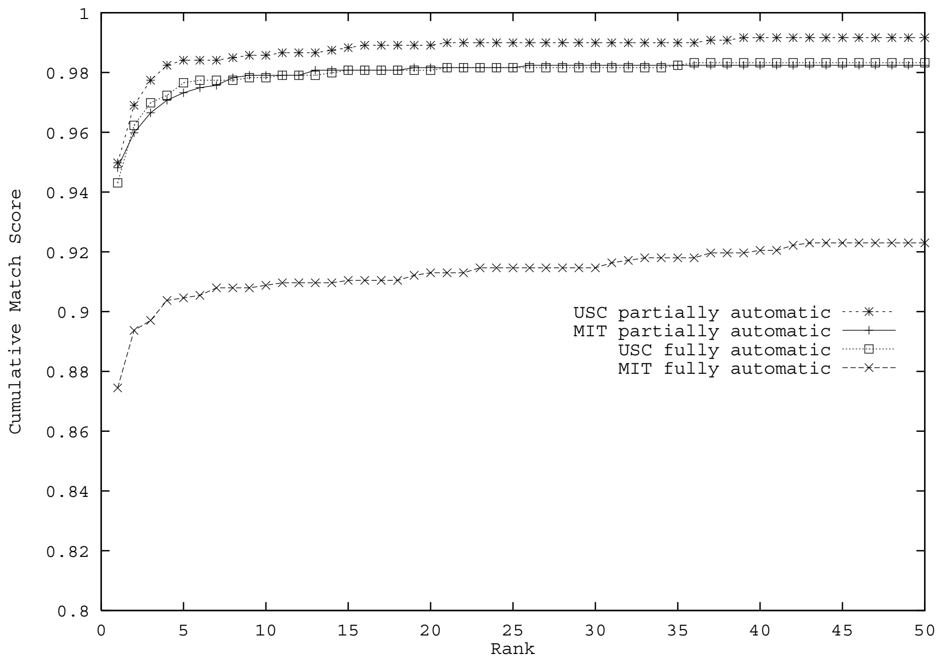


Figure 9: Identification performance of fully automatic algorithms against partially automatic algorithms for **FB** probes.

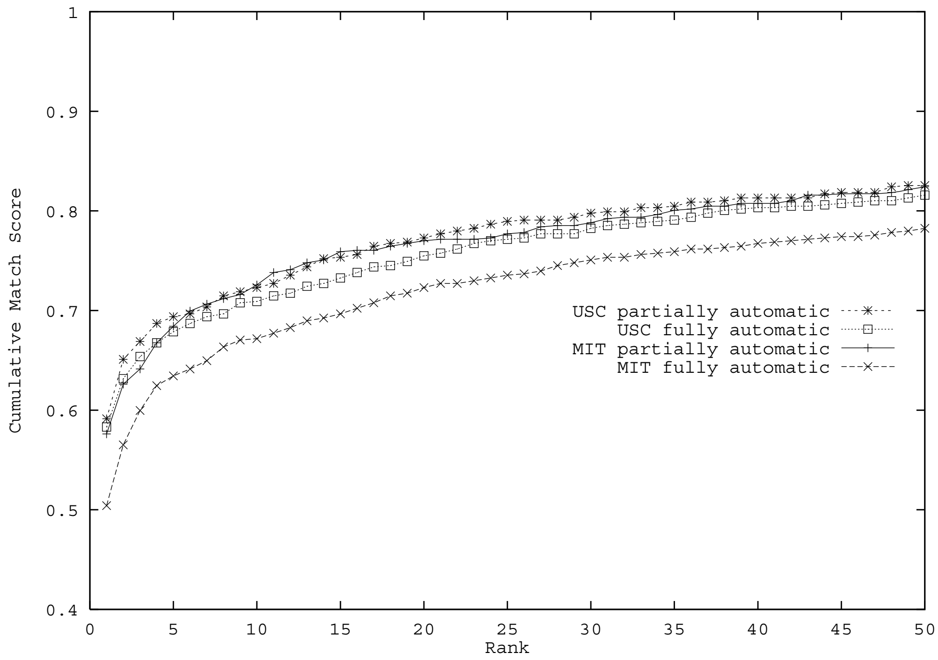


Figure 10: Identification performance of fully automatic algorithms against partially automatic algorithms for duplicate I probes.

5.2 Fully Automatic Performance

In this subsection, we report performance for the fully automatic algorithms of the MIT Media Lab and USC. To allow for a comparison between the partially and fully automatic algorithms, we plot the results for the partially and fully automatic algorithms. Figure 9 shows performance for **FB** probes and figure 10 shows performance for duplicate I probes. (The gallery and probe sets are the same as in subsection 5.1.)

5.3 Variation in Performance

From a statistical point of view, a face-recognition algorithm estimates the identity of a face. Consistent with this view, we can ask about the variance in performance of an algorithm: “For a given category of images, how does performance change if the algorithm is given a different gallery and probe set?” In tables 3 and 4, we show how algorithm performance varies if the people in the galleries change. For this experiment, we constructed six galleries of approximately 200 individuals, in which an individual was in only one gallery (the number of people contained within each gallery versus the number of probes scored is given in tables 3 and 4). Results are reported for the partially automatic algorithms. For the results in this section, we order algorithms by their top rank score on each gallery; for example, in table 3, the UMD Mar97 algorithm scored highest on gallery 1 and the baseline PCA and correlation tied for 9th place. Also included in this table is average performance for all algorithms. Table 3 reports results for **FB** probes. Table 4 is organized in the same manner as table 3, except that duplicate I probes are scored. Tables 3 and 4 report results for the same gallery. The galleries were constructed by placing images within the galleries by chronological order in which the images were collected (the first gallery contains the first images collected and the 6th gallery contains the most recent images collected). In table 4, mean age refers to the average time between collection of images contained in the gallery and the corresponding duplicate probes. No scores are reported in table 4 for gallery 6 because there are no duplicates for this gallery.

6 Discussion and Conclusion

In this paper we presented the Sep96 FERET evaluation protocol for face recognition algorithms. The protocol makes it possible to independently evaluate algorithms. The protocol was designed to evaluate algorithms on different galleries and probe sets for different scenarios. Using this protocol, we computed performance on identification and verification tasks. The verification results are presented in Rizvi et al. [8], and all verification results mentioned in this section are from that paper. In this paper we presented detailed identification results. Because of the Sep96 FERET evaluation protocol’s ability to test algorithms performance on different tasks for multiple galleries and probe sets, it is the de facto standard for measuring performance of face recognition algorithms. These results show that factors effecting performance include scenario, date tested, and probe category.

The Sep96 test was the latest FERET test (the others were the Aug94 and Mar95 tests [6]). One of the main goals of the FERET tests has been to improve the performance of face recognition algorithms, and is seen in the Sep96 FERET test. The first case is the improvement in performance of the MIT Media Lab September 1996 algorithm over

Table 3: Variations in identification performance on six different galleries on FB probes. Images in each gallery do not overlap. Ranks range from 1–10.

	Algorithm Ranking by Top Match					
	Gallery Size / Scored Probes					
	200/200	200/200	200/200	200/200	200/199	196/196
Algorithm	gallery 1	gallery 2	gallery 3	gallery 4	gallery 5	gallery 6
Baseline PCA	9	10	8	8	10	8
Baseline correlation	9	9	9	6	9	10
Excalibur Corp.	6	7	7	5	7	6
MIT Sep96	4	2	1	1	3	3
MIT Mar95	7	5	4	4	5	7
Michigan State Univ.	3	4	5	8	4	4
Rutgers Univ.	7	8	9	6	7	9
UMD Sep96	4	6	6	10	5	5
UMD Mar97	1	1	3	2	2	1
USC	2	3	2	2	1	1
Average Score	0.935	0.857	0.904	0.918	0.843	0.804

Table 4: Variations in identification performance on five different galleries on duplicate probes. Images in each of the gallery does not overlap. Ranks range from 1–10.

	Algorithm Ranking by Top Match				
	Gallery Size / Scored Probes				
	200/143	200/64	200/194	200/277	200/44
Mean Age of Probes (months)	9.87	3.56	5.40	10.70	3.45
Algorithm	gallery 1	gallery 2	gallery 3	gallery 4	gallery 5
Baseline PCA	6	10	5	5	9
Baseline correlation	10	7	6	6	8
Excalibur Corp.	3	5	4	4	3
MIT Sep96	2	1	2	2	3
MIT Mar95	7	4	7	8	10
Michigan State Univ.	9	6	8	10	6
Rutgers Univ.	5	7	10	7	6
UMD Sep96	7	9	9	9	3
UMD Mar97	4	2	3	3	1
USC	1	3	1	1	1
Average Score	0.238	0.620	0.645	0.523	0.687

the March 1995 algorithm; the second is the improvement of the UMD algorithm between September 1996 and March 1997.

By looking at progress over the series of FERET tests, one sees that substantial progress has been made in face recognition. The most direct method is to compare the performance of fully automatic algorithms on **fb** probes (the two earlier FERET tests only evaluated fully automatic algorithms. The best top rank score for **fb** probes on the Aug94 test was 78% on a gallery of 317 individuals, and for Mar95, the top score was 93% on a gallery of 831 individuals [6]. This compares to 87% in September 1996 and 95% in March 1997 (gallery of 1196 individuals). This method shows that over the course of the FERET tests, the absolute scores increased as the size of the database increased. The March 1995 score was from one of the MIT Media Lab algorithms, and represents an increase from 76% in March 1995.

On duplicate I probes, MIT Media Lab improved from 39% (March 1995) to 51% (September 1996); USC's performance remained approximately the same at 57-58% between March 1995 and March 1997. This improvement in performance was achieved while the gallery size increased and the number of duplicate I probes increased from 463 to 722. While increasing the number of probes does not necessarily increase the difficulty of identification tasks, we argue that the Sep96 duplicate I probe set was more difficult to process then the Mar95 set. The Sep96 duplicate I probe set contained the duplicate II probes and the Mar95 duplicate I probe set did not contain a similar class of probes. Overall, the duplicate II probe set was the most difficult probe set.

Another goal of the FERET tests is to identify areas of strengths and weaknesses in the field of face recognition. We addressed this issue by computing algorithm performance for multiple galleries and probe sets. From this evaluation, we concluded that algorithm performance is dependent on the gallery and probe sets. We observed variation in performance due to changing the gallery and probe set within a probe category, and by changing probe categories. The effect of changing the gallery while keeping the probe category constant is shown in tables 3 and 4. For **fb** probes, the range for performance is 80% to 94%; for duplicate I probes, the range is 24% to 69%. Equally important, tables 3 and 4 shows the variability in relative performance levels. For example, in table 4, UMD Sep96 duplicate performance varies between number three and nine. Similar results were found in Moon and Phillips [4] in their study of principal component analysis-based face recognition algorithms. This shows that an area of future research could measure the effect of changing galleries and probe sets, and statistical measures that characterize these variations.

Figures 7 and 8 shows probe categories characterized by difficulty. These figures show that **fb** probes are the easiest and duplicate II probes are the most difficult. On average, duplicate I probes are easier to identify than **fc** probes. However, the best performance on **fc** probes is significantly better than the best performance on duplicate I and II probes. This comparative analysis shows that future areas of research could address processing of duplicate II probes and developing methods to compensate for changes in illumination.

The scenario being tested contributes to algorithm performance. For identification, the MIT Media Lab algorithm was clearly the best algorithm tested in September 1996. However, for verification, there was not an algorithm that was a top performer for all probe categories. Also, for the algorithms tested in March 1997, the USC algorithm performed overall better than the UMD algorithm for identification; however, for verification, UMD overall performed better. This shows that performance on one task is not predictive of performance on another task.

The September 1996 FERET test shows that definite progress is being made in face recognition, and that the upper bound in performance has not been reached. The improvement in performance documented in this paper shows directly that the FERET series of tests have made a significant contribution to face recognition. This conclusion is indirectly supported by (1) the improvement in performance between the algorithms tested in September 1996 and March 1997, (2) the number of papers that use FERET images and report experimental results using FERET images, and (3) the number of groups that participated in the Sep96 test.

References

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

[2] 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.

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

[4] 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.

[5] P. J. Phillips and P. Rauss. The face recognition technology (FERET) program. In *Proceedings of Office of National Drug Control Policy, CTAC International Technology Symposium*, pages 8–11 — 8–20, August 1997.

[6] 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.

[7] P. Rauss, P. J. Phillips, A. T. DePersia, and M. Hamilton. The FERET (Face Recognition Technology) program. In *Surveillance and Assessment Technology for Law Enforcement*, volume SPIE Vol. 2935, pages 2–11, 1996.

[8] S. Rizvi, P. J. Phillips, and H. Moon. The feret verification testing protocol for face recognition algorithms. *Image and Vision Computing Journal*, (to appear) 1999.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.