The FERET Evaluation Methodology for Face-Recognition Algorithms

P. Jonathon Phillips, *Member*, *IEEE*, Hyeonjoon Moon, *Member*, *IEEE*,
Syed A. Rizvi, *Member*, *IEEE*, and Patrick J. Rauss

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 1,199 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.

Index Terms—Face recognition, algorithm evaluation, FERET database.

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 [6]. The rapid development of face recognition is due to a combination of factors: active development of 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 September 1996 FERET test.

The September 1996 FERET test provides a comprehensive picture of the state-of-the-art in face recognition from still images. This was accomplished by evaluating the 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 an unknown face that is to be identified, whereas, in a verification application, an algorithm is presented with a face and a claimed identity, and the algorithm either accepts or rejects the claim. In this paper, we describe the FERET database and the September 1996 FERET evaluation protocol and present identification results. Verification results are presented in Rizvi et al. [10].

- *P.J. Phillips is with the National Institute of Standards and Technology, 100 Bureau Dr. STOP 8940, Gaithersburg, MD 20899-8940. E-mail: jonathon@nist.gov.*
- *H. Moon is with Lau Technologies, 30 Porter Road, Littleton, MA 01460. E-mail: hm@lautechnologies.com.*
- *S.A. Rizvi is with the Department of Engineering Science and Physics, College of Staten Island/CUNY, Staten Island, NY 10314. E-mail: rizvi@wagner.csi.cuny.edu.*
- *P.J. Rauss is with the Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783-1197. E-mail: rauss@erim-it.com.*

Manuscript received 2 Nov. 1998; revised 24 Sept. 1999; accepted 19 May 2000.
Recommended for acceptance by D.J. Kriegman.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 108156.

The FERET tests model the following face recognition applications: identification from large law enforcement databases and verification from biometric signatures stored on smart cards. For both applications, there are a limited number of facial images per person and the face representation is learned (or decided) prior to people being enrolled in the system.

In the Federal Bureau of Investigation's (FBI) Integrated Automated Fingerprint Identification System (IAFIS), the only required mugshot is a full frontal image [2]. The IAFIS stores digital fingerprints and mugshots and will be the main depository of criminal fingerprints and mugshots in the United States. Other examples of large databases with one image per person are photographs from drivers licenses, passports, and visas.

When the IAFIS is fully operational, it is expected to receive 5,000 mugshots per day (1,800,000 per year). Because of the large number of mugshots, it is not practical to continually update the representation. Updating the representation would require training from millions of faces and updating millions of database records.

For verification applications where biometric signatures are stored on smart card, a user inserts a smart card into an electronic reader and provides a new biometric signature to the system. The system then reads the biometric signature stored on the smart card and compares it with the new signature. Based on the comparison, the claimed identity is either accepted or rejected. Because of the limited amount of storage space, a facial image cannot be stored on a smart card and a representation of the face must be stored. Thus, once the first person is enrolled in the system, it is not possible to update the facial representation. Also, because of limited storage space, the representation of only one facial image is stored on a smart card.

The FERET was a general evaluation designed to measure performance of laboratory algorithms on the FERET database. The main goals of the FERET evaluation were to assess the state-of-the-art and the feasibility of automatic face recognition. Thus, the FERET test did not

explicitly measure the effect on performance of individual components of an algorithm nor did the test measure performance under operational scenarios. An operational test evaluates algorithms in an orderly and scientific manner under all conditions in which a system will operate.

To obtain a robust assessment of performance, algorithms were evaluated against different categories of images. The categories were broken out by a lighting change, people wearing glasses, and the time between the acquisition date of the database image and the image presented to the algorithm. By listing performance in 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 assess which applications can be successfully addressed.

All face recognition algorithms known to the authors consist of two parts: 1) face localization and normalization and 2) face identification. We use the term face localization and normalization to differentiate it from face detection. In detection, the task is to find all faces in an image, where there can be multiple or no faces in the image. In the FERET evaluation, there is one face in an image. In the first part of an algorithm, the face is located in an image and then the face is normalized into a standard position for the recognition portion of the algorithm. Usually, normalization requires that a set of facial features is actually located to within a couple of pixels.

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*. (A glossary of terms is in the Appendix.) The September 1996 test evaluated both fully and partially automatic algorithms. Partially automatic algorithms are given a facial image and the coordinates of the centers 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 percent 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 reports 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 also on the aggregate performance of all algorithms tested. Through this type of assessment, the community learns in an 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 and Rauss [6], Phillips et al. [7], and Rauss et al. [8]). 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*. Since there is only one face in an image, sometimes "probe" refers to the identity of the person in a probe image. 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, which 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 probes. A *duplicate* probe is usually an image of a person whose corresponding gallery image was taken on a different day. (Technically, the probe and gallery images were from different image sets; see description of the FERET database below.)

The FERET database is designed to advance the state of the art in face recognition, with the collected images 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 (Fig. 1).

Images of an individual were acquired in sets of 5 to 11 images. 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

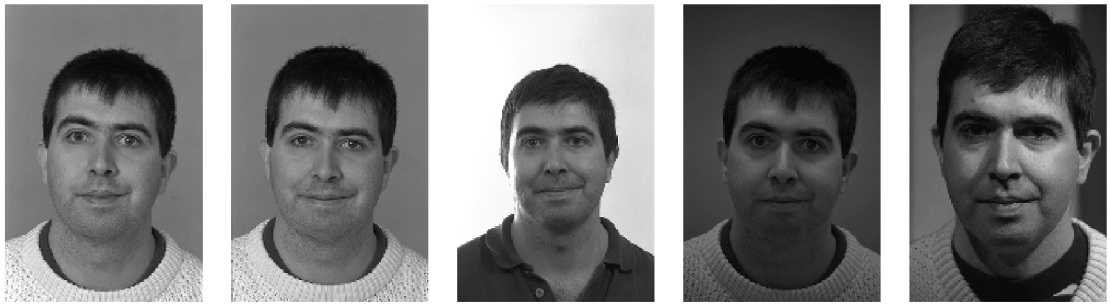

Fig. 1. Examples of different categories of probes (image). The duplicate I image was taken within one year of the **fa** image and the duplicate II and **fa** images were taken at least one year apart.

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, 1,564 sets of images were in the database, consisting of 14,126 total images. The database contains 1,199 individuals and 365 duplicate sets of images. For some people, more than two years elapsed between their first and most recent sittings, with some subjects being photographed multiple times (Fig. 1). The development portion of the database consisted of 503 sets of images and was released to researchers. The remaining images were sequestered.

3 Test Design

3.1 Test Design Principles

The FERET September 1996 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 evaluation design must solve the *three bears problem*. The test cannot be 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 used in the evaluation and the evaluation protocol. Tests are administered using an evaluation 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. The design principles directly relate the evaluation to the face recognition problem being evaluated. For FERET, the applications are searching large databases and verifying identities stored on smart cards. Stating the design principles 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. For all results in this paper, the images in the galleries and probe sets were distinct. The target set is given to the algorithm as the set of known facial images. The images in the query set consist 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 $s_i(k)$.

Multiple galleries and probe sets can be constructed from the target and query sets. A gallery $\mathcal{G}$ is a subset of the target set. Similarly, a probe set $\mathcal{P}$ is a subset of the query set. 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 $t_k \in \mathcal{G}$.

Using target and query sets allows us to compute performance for different categories of images. Possible probe categories include: 1) gallery and probe images taken on the same day, 2) duplicates taken within a week of the gallery image, and 3) duplicates where the time between the images is at least one year. 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

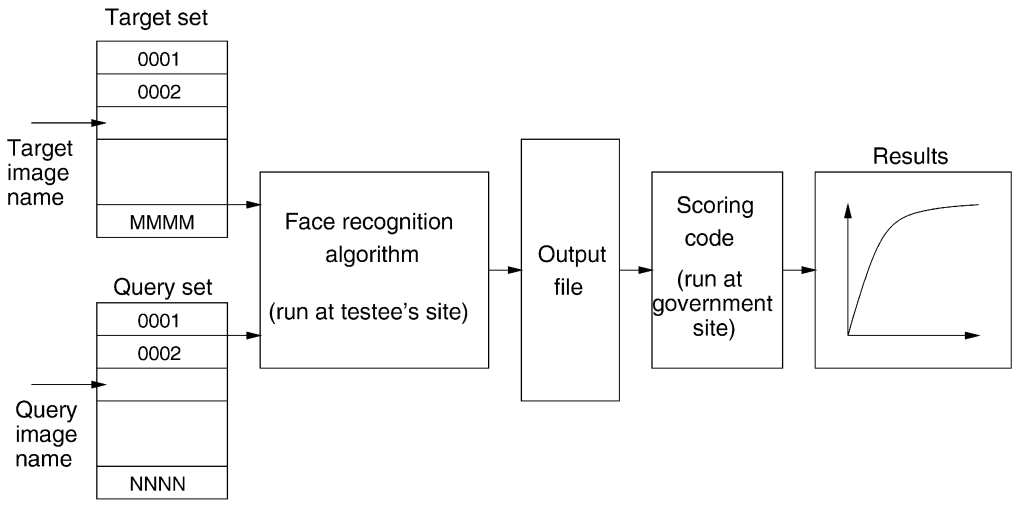

Fig. 2. Schematic of the FERET testing procedure.

then create galleries of $200, 300, \ldots, 1,000$ 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.

We now list the three design principles. First, all faces in the target set are treated as unique faces. This allows us to construct multiple galleries with one image per person. In practice, this condition is enforced by giving every image in the target and query set a unique random identification.

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. The third design rule is that all algorithms compute a similarity measure between all combinations of images from the target and query sets.

3.2 Test Details

In the September 1996 FERET test, the target set contained 3,323 images and the query set 3,816 images. The target set consists of **fa** and **fb** frontal images. The query set consisted of all the images in the target set plus the **fc**, rotated images, and digitally modified images. The digitally modified images in the query set were designed to test the effects of illumination and scale. (Results from the rotated and digitally modified images are not reported here.) All the results reported in this article are generated from galleries that are subsets of this target set and probe sets that are subsets of this query set. 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. (Note: Having the target set as subset of the query set does not constitute training and testing on the same images. This is because the face representation is learned prior to the start of the test.)

There were two versions of the September 1996 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 centers 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 target and query 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 and were beyond the scope of the September 1996 FERET test.

Fig. 2 presents a schematic of the testing procedure. To ensure that matching was not done by file name, we gave the images random names. A rough estimate of the pose of each face was provided to each testee. Example pose estimates provided were: frontal, and quarter and half right.

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

TABLE 1
Representation and Similarity Metric for Algorithms Evaluated

Algorithm	Represenation	Similarity measure
Excalibur Co.	Unknown	Unknown
MIT Media Lab 95	PCA	L_2
MIT Media Lab 96	PCA-difference space	MAP Bayesian Statistic
Michigan St. U.	Fischer discriminant	L_2
Rutgers U.	Greyscale projection	Weighted L_1
U. of So. CA.	Dynamic Link Architecture (Gabor Jets)	Elastic graph matching
U. of MD 96	Fischer discriminant	L_2
U. of MD 97	Fischer discriminant	Weighted L_2
Baseline	PCA	L_1
Baseline	Correlation	Angle

different and important aspects of face-recognition algorithms and report different performance statistics. The open universe model is used to evaluate verification applications. The FERET scoring procedures for verification is given in Rizvi et al. [10].

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, which are plotted on a graph. The horizontal axis of the graph is rank and the vertical axis is the probability of identification (P_I) (or percentage of correct matches).

The computation of an identification score is quite simple. Let $\mathcal{P}$ be a probe set and $|\mathcal{P}|$ be the size of $\mathcal{P}$. We score probe set $\mathcal{P}$ against gallery $\mathcal{G}$, where $\mathcal{G} = \{g_1, \ldots, g_M\}$ and $\mathcal{P} = \{p_1, \ldots, p_N\}$, 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. 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 score for $g_k \in \mathcal{G}$. A probe p_i is in the top n if $s_i(id(i))$ is one of the nth smallest scores $s_i(\cdot)$ for gallery $\mathcal{G}$. Let R_n denote the number of probes in the top n. We reported $R_n/|\mathcal{P}|$, the fraction of probes in the top n.

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 September 1996 FERET test was designed to measure algorithm performance for identification and verification tasks. In this article, we report identification results. Verification results are reported in Rizvi et al. [9], [10]. We report results for 12 algorithms that include 10 partially automatic algorithms and two fully automatic algorithms. The test was administered in September 1996 and March 1997 (see Table 1 for the representation and similarity metric for each algorithm and Table 2 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 [3], [4]. Algorithms were also tested from Excalibur Corporation (Carlsbad, California), Michigan State University (MSU) [11], [16], Rutgers University [13], the University of Southern California (USC) [14], and two from the University of Maryland (UMD) [1], [15], [16]. 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 the 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 [5], [12]. 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 centers of the eyes were placed on specific pixels, 2) faces were masked to remove background and hair, and 3) the nonmasked 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 centers of the eyes were placed on specific pixels and 2) faces were masked to remove background and hair.

TABLE 2
List of Groups That Took the September 1996 Test Broken Out by Versions Taken and Dates Administered (the 2 by MIT Indicates that Two Algorithms were Tested)

		Test Date		
Version of test	Group	September 1996	March 1997	Baseline
Fully Automatic	MIT Media Lab [3,4]	•		
	U. of So. California [14]		•	
Eye Coordinates Given	Baseline PCA [5,12]			•
	Baseline Correlation			•
	Excalibur Corp.	•		
	MIT Media Lab	2		
	Michigan State U. [11,16]	•		
	Rutgers U. [13]	•		
	U. Maryland [1,15,16]	•	•	
	USC		•	

We report identification scores for four categories of probes. For three of the probe categories, performance was computed using the same gallery. For the fourth category, a subset of the first gallery was used. The first gallery consisted of images of 1,196 people with one image per person. For the 1,196 people, the target and query sets contain **fa** and **fb** images from the same set. (The FERET images were collected in sets and, in each session, there are two frontal images, **fa** and **fb**, see Section 2.) One of these images was placed in the gallery and the other was placed in the **FB** probe set. The **FB** probes were the first probe category. (This category is denoted by **FB** to differentiate it from the **fb** images in the FERET database.) (Note: the query set contained all the images in the target set, so the probe set is a subset of the query set.) Also, none of the faces in the gallery images wore glasses. Thus, the **FB** probe set consisted of probe images taken on the same day and under the same illumination conditions as the corresponding gallery image.

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** probes (images taken the same day as the corresponding gallery image, but with a different camera and lighting). The fourth category consisted of duplicates where there was at least one year between the acquisition of the probe image and corresponding gallery image, i.e., the gallery images were acquired before January 1995 and the probe images were acquired after January 1996. We refer to this category as the duplicate II probes. The gallery for the **FB**, duplicate I, and **fc** probes was the same. The gallery for duplicate II probes was a subset of 864 images from the gallery for the other categories.

5.1 Partially Automatic Algorithms

In this section, we report results for the partially automatic algorithms. Table 3 shows the categories corresponding to the figures presenting the results, type of results, and size of the gallery and probe sets (Figs. 3, 4, 5, and 6). The results for each probe category are presented on two graphs. One graph shows performance for algorithms tested in September 1996 and the baseline algorithms. The other shows performance for algorithms tested in March 1997, the baseline algorithms, and the UMD algorithm tested in September 1996 (this shows improvement between tests). (The results are reported as cumulative match scores.)

In Figs. 7 and 8, we compare the difficulty of different probe sets. Whereas Figs. 4, 5, and 6 report identification performance for each algorithm, Fig. 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 Fig. 4. In Fig. 8, we presented current upper bound for performance on partially automatic algorithms for each probe category. For each category of probe, Fig. 8 plots the algorithm with the highest top rank score (R_1). Figs. 7 and 8 report performance of four categories of probes, FB, duplicate I, **fc**, and duplicate II.

5.2 Fully Automatic Performance

In this section, 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 from both institutions. Fig. 9 shows performance for **FB** probes and Fig. 10 shows performance for duplicate I probes. (The gallery and probe sets are the same as in Section 5.1.)

TABLE 3
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

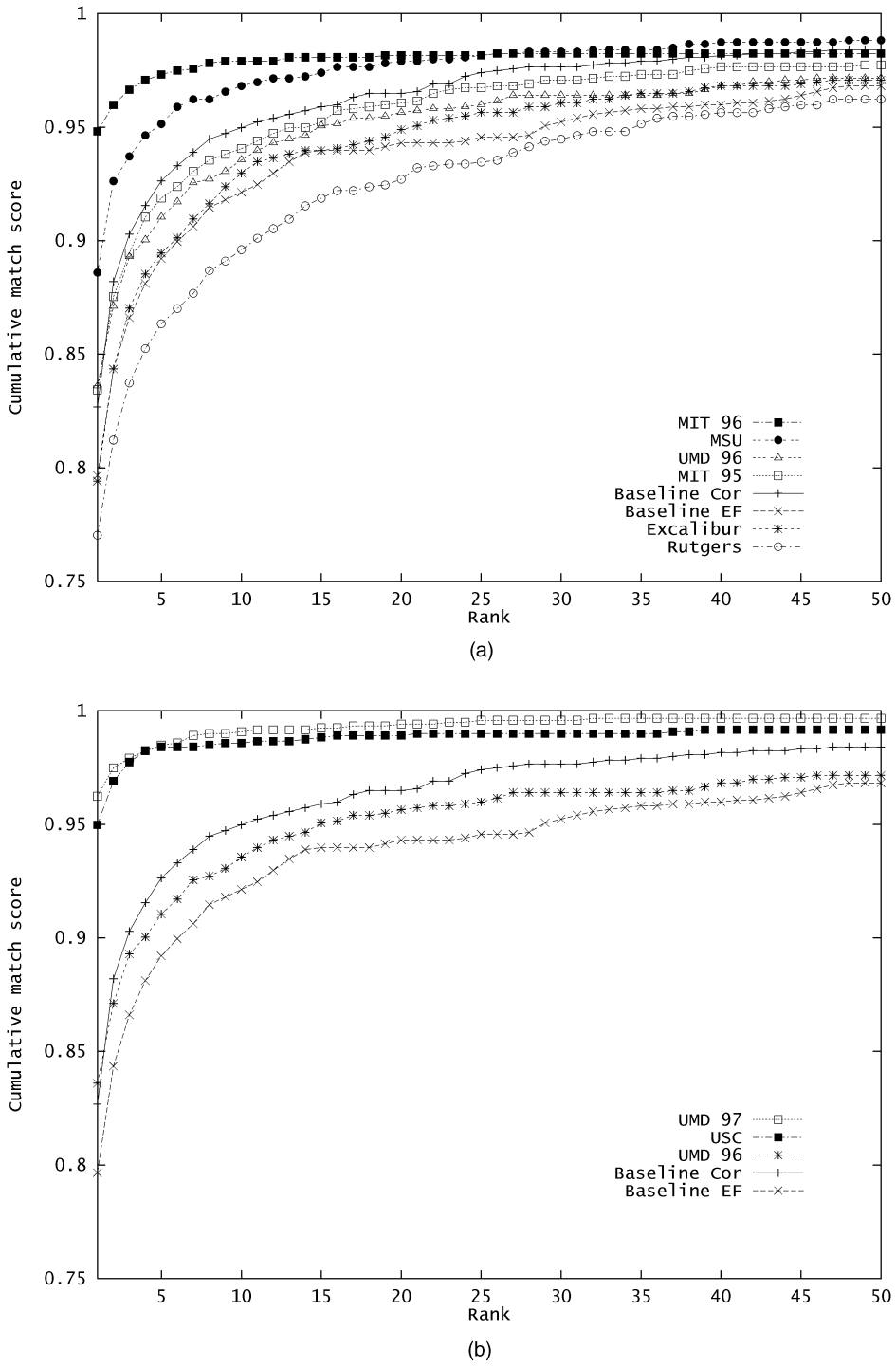

Fig. 3. Identification performance against **FB** probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.

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 change 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 4 and 5, 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 4 and 5.) 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 4, the UMD March 1997 algorithm scored highest on gallery 1 and the baseline PCA and correlation tied for ninth place. Also included in this table is average

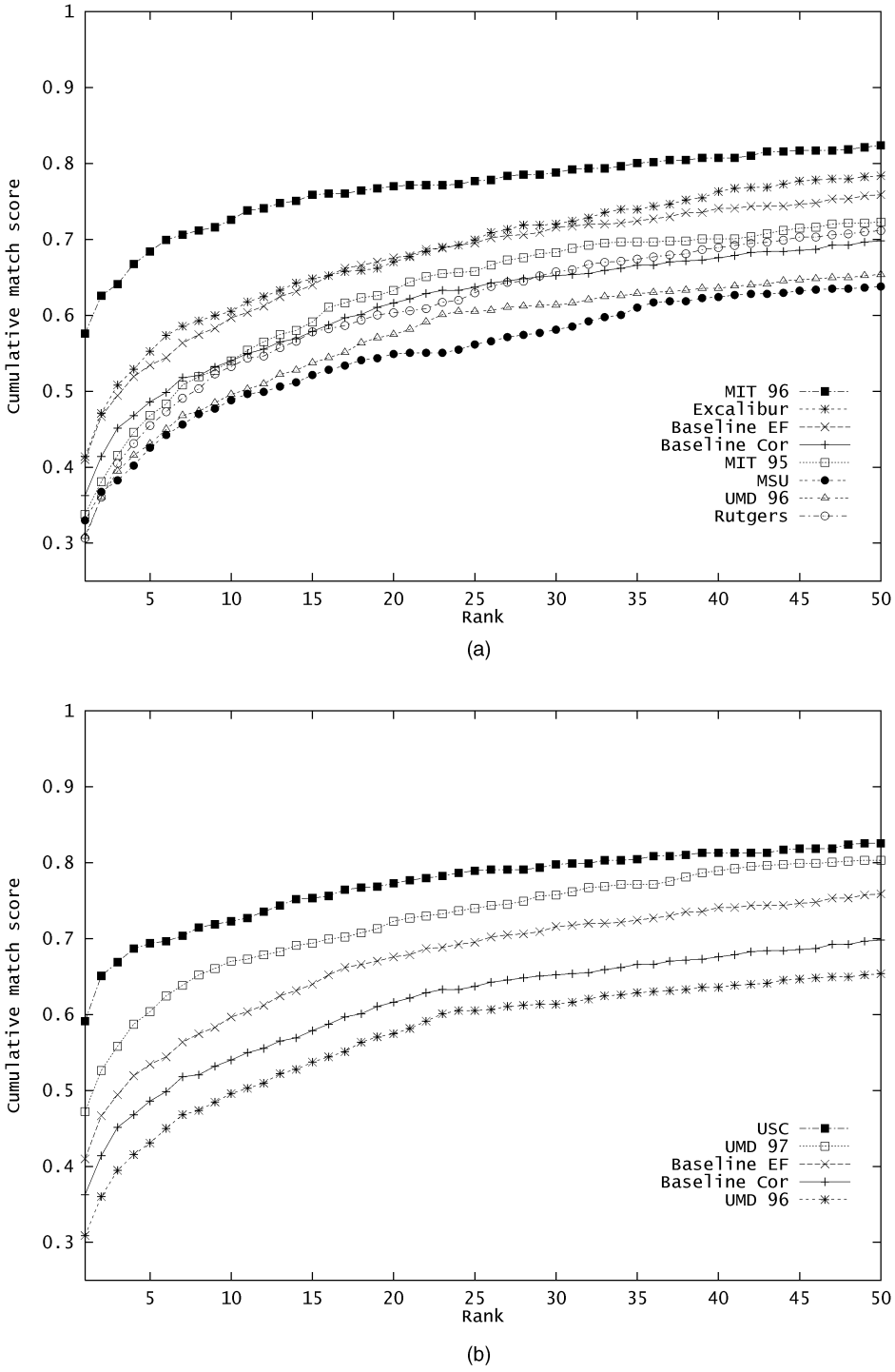

Fig. 4. Identification performance against all duplicate I probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.

performance for all algorithms. Table 4 reports results for **FB** probes. Table 5 is organized in the same manner as Table 4, except that duplicate I probes are scored. Tables 4 and 5 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 sixth gallery contains the most recent images collected). In Table 5, 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 5 for gallery 6 because there are no duplicates for this gallery.

6 Discussion and Conclusion

In this paper, we presented the September 1996 FERET evaluation protocol for face recognition algorithms. The protocol was designed so that performance can be measured on different galleries and probe sets and on

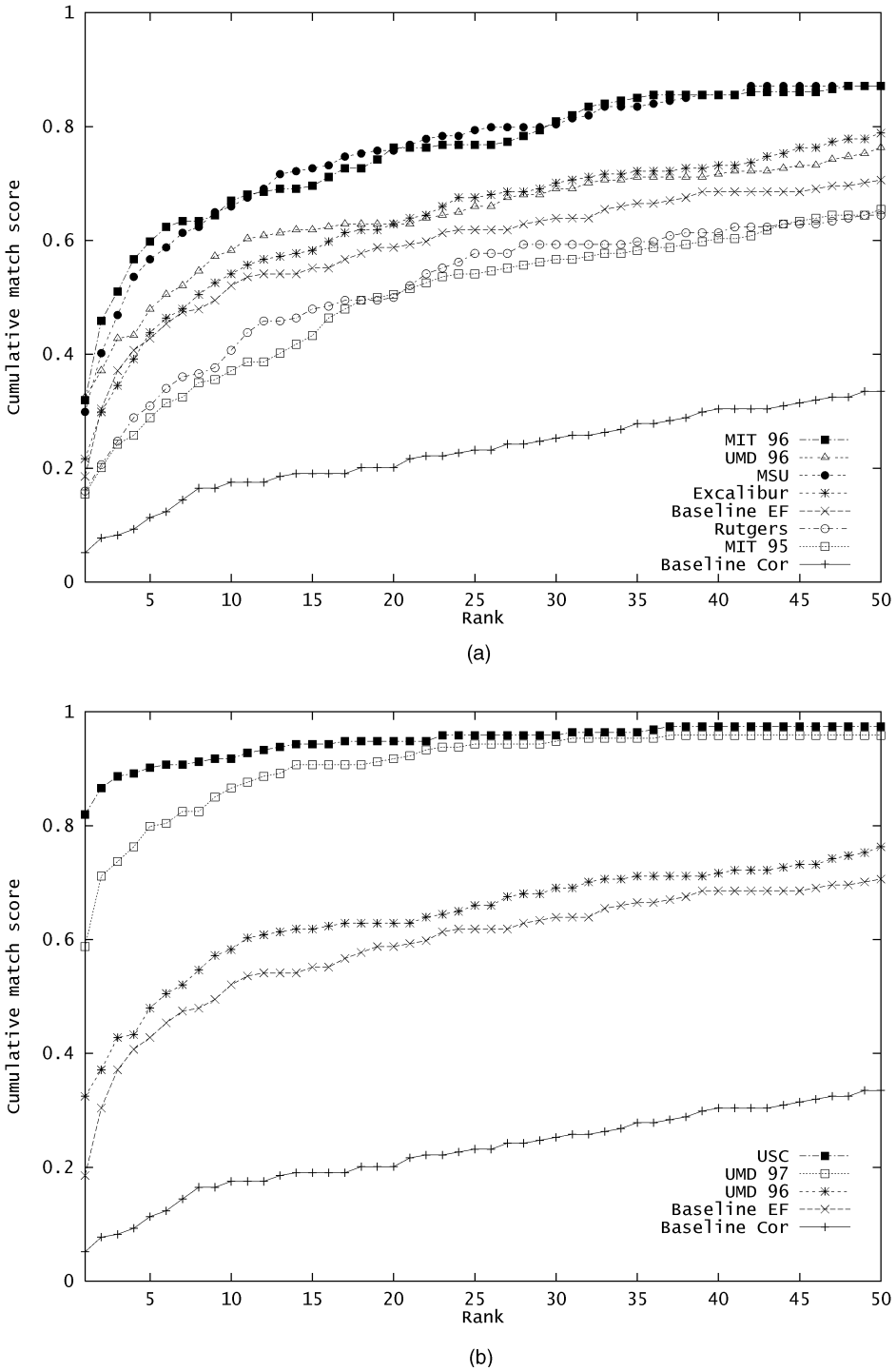

Fig. 5. Identification performance against **fc** probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.

identification and verification tasks. (Verification results mentioned in this section are from Rizvi et al. [9], [10].)

The September 1996 test was the latest FERET evaluation (the others were the August 1994 and March 1995 tests [7]). One of the main goals of the FERET evaluations was to encourage and measure improvements in the performance of face recognition algorithms, which is seen in the September 1996 FERET test. The first case is the improvement in performance of the MIT Media Lab September 1996 algorithm over 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 evaluations, 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 evaluations only evaluated fully automatic algorithms). The best top rank score for **fb** probes on the August 1994 evaluation was 78 percent on a gallery of 317 individuals and, for

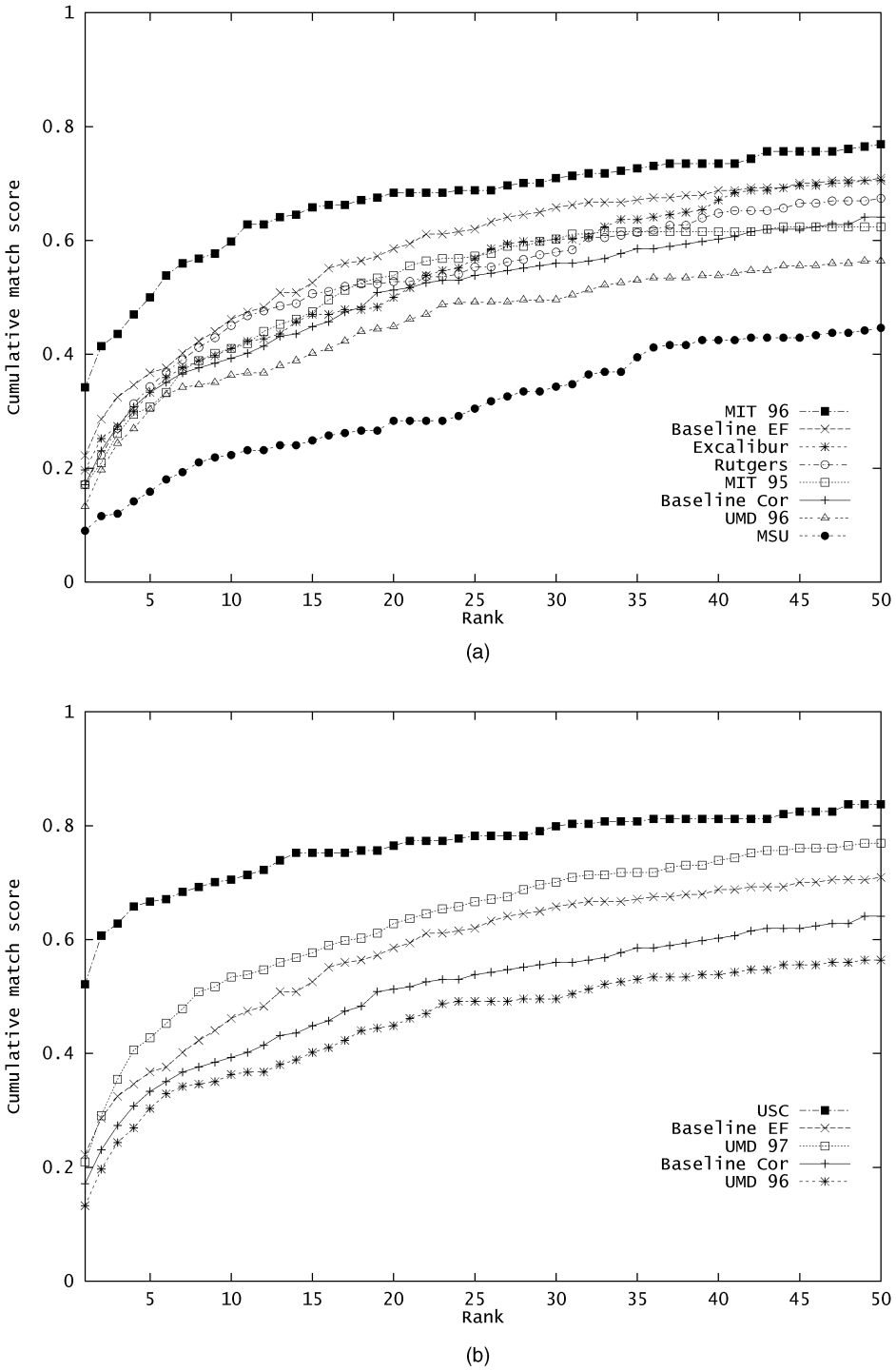

Fig. 6. Identification performance against duplicate II probes. (a) Partially automatic algorithms tested in September 1996. (b) Partially automatic algorithms tested in March 1997.

March 1995, the top score was 93 percent on a gallery of 831 individuals [7]. This compares to 87 pecent in September 1996 and 95 percent in March 1997 (gallery of 1,196 individuals). This method shows, that over the course of the FERET evaluations, 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 percent in March 1995.

On duplicate I probes, MIT Media Lab improved from 39 percent (March 1995) to 51 percent (September 1996); USC's performance remained approximately the same at 57-58 percent 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 September 1996 duplicate I probe set was more difficult to process than the March 1995 set. The September 1996 duplicate I probe set contained the duplicate II probes and the March 1995

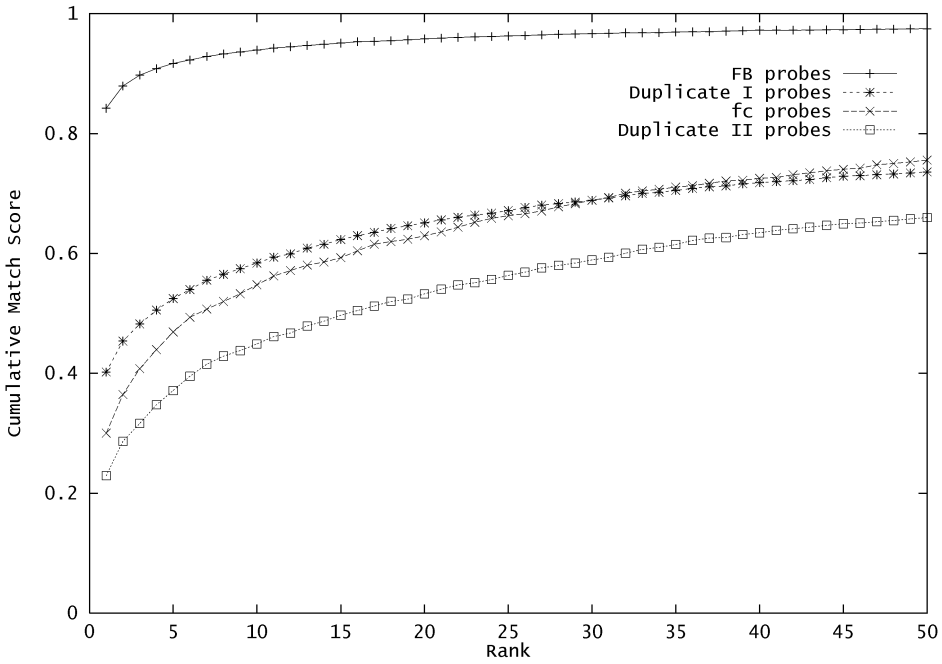

Fig. 7. Average identification performance of partially automatic algorithms on each probe category.

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 evaluations is to identify areas of strengths and weaknesses in the field of face recognition. We addressed this issue by reporting performance for multiple galleries and probe sets and different probe categories. 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 4 and 5. For **fb** probes, the range for performance is 80 percent to 94 percent; for duplicate I probes, the range is 24 percent to 69 percent. Equally important, Tables 4 and 5 show the variability in relative performance levels. For example, in Table 5, UMD September 1996 duplicate performance varies between number three and nine, while at the same time there are algorithms that consistently outperform other algorithms. Of the algorithms tested in September 1996, the

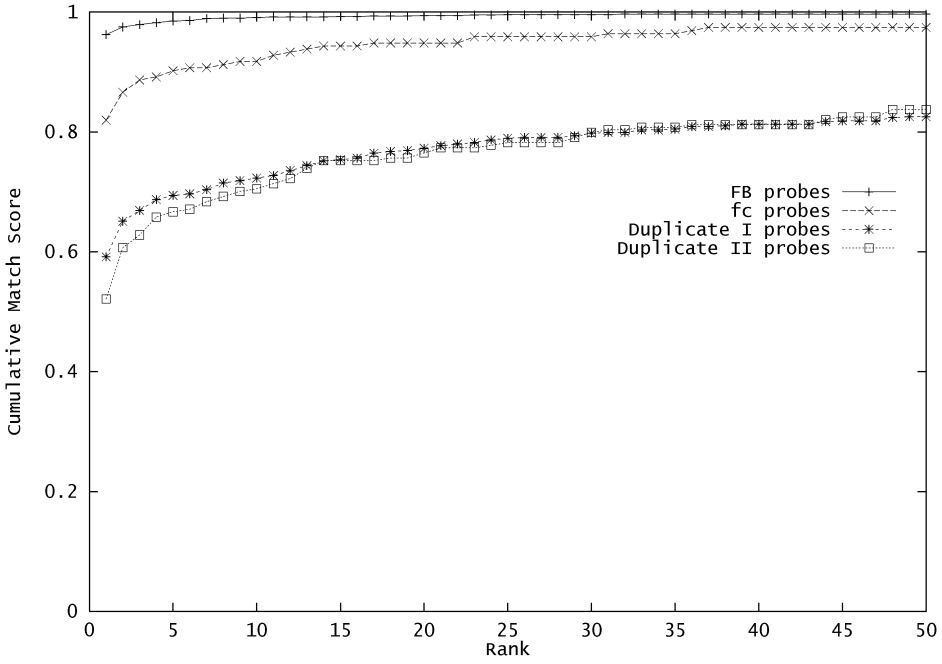

Fig. 8. Current upper bound identification performance of partially automatic algorithm for each probe category.

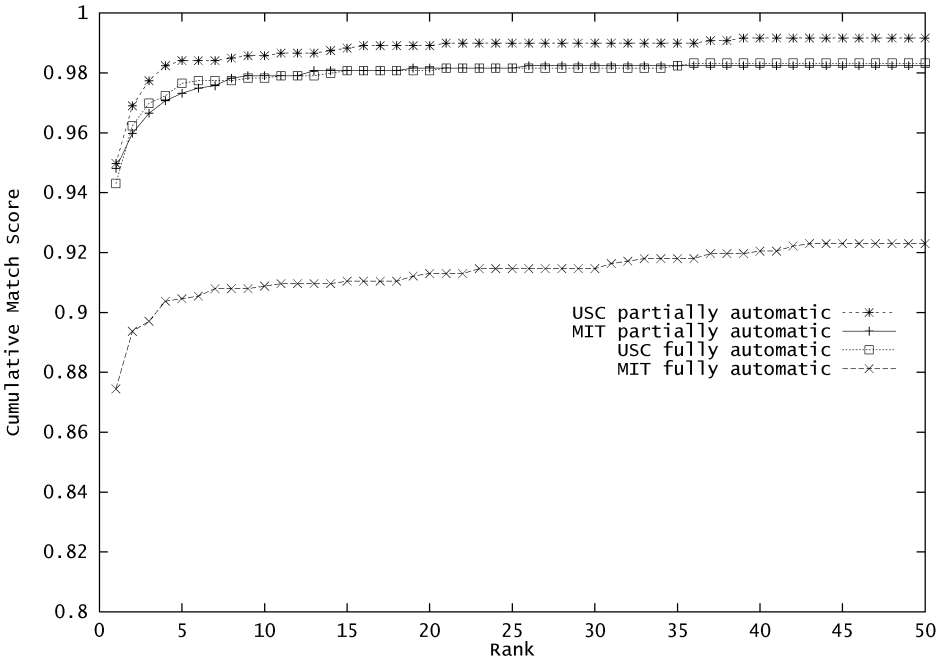

Fig. 9. Identification performance of fully automatic algorithms against partially automatic algorithms for **FB** probes.

September 1996 MIT algorithm clearly outperformed the other algorithms. In addition, the September 1996 MIT algorithms and the algorithms tested in March 1997 (UMD March 1997 and USC) outperformed the other algorithms tested. This shows that, despite the overall variation in performance, definite conclusions about algorithm performance can be made. These conclusions are consistent with Figs. 4, 5, and 6.

The variation in Tables 4 and 5 is because traditional method of calculating error bars and confidence regions do not apply to face recognition. These traditional methods require that each run of the decision problem be made with the same classes, i.e., character recognition with the 26 letters in the English alphabet. However, in face recognition, changing the people in the gallery changes the underlying classification problem. (Remember, each person is a different class.) Computing error bars with different people in the gallery is equivalent to computing error bars for a character recognition system using performance from different sets of characters.

Similar results were found in Moon and Phillips [5] in their study of principal component analysis-based face

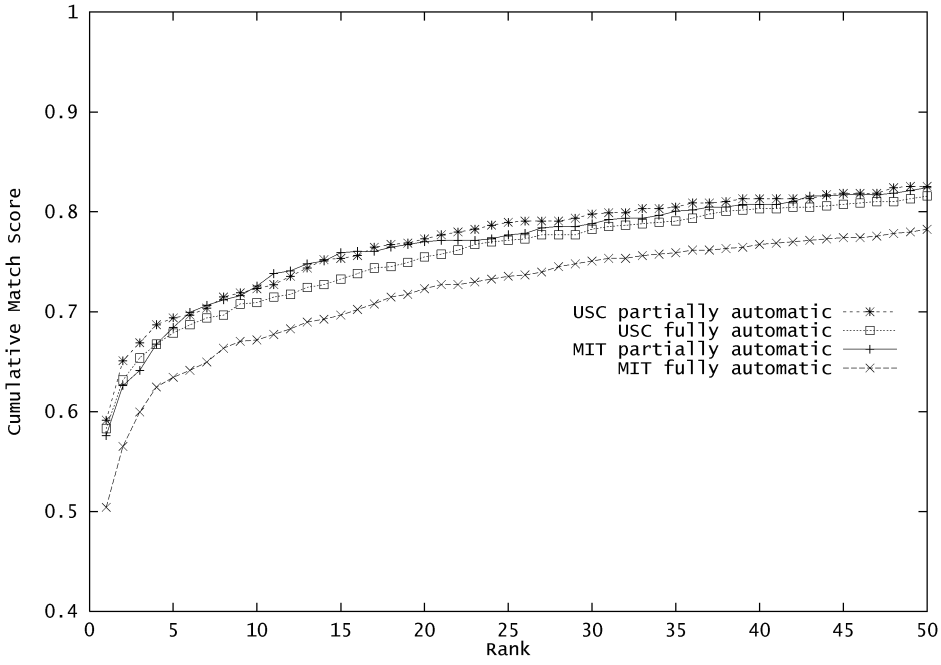

Fig. 10. Identification performance of fully automatic algorithms against partially automatic algorithms for duplicate I probes.

TABLE 4
Variations in Identification Performance on Six Different Galleries on FB Probes
Images in Each of the Galleries 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

recognition algorithms. This shows that an area of future research is measuring the effect of changing galleries and probe sets and statistical measures that characterize these variations.

Figs. 7 and 8 show 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 include 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 necessarily predictive of performance on a different task.

The September 1996 FERET evaluation 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 evaluations has 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

TABLE 5
Variations in Identification Performance on Five Different Galleries on Duplicate Probes
Images in Each of the Galleries do 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

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 September 1996 test.

APPENDIX

GLOSSARY OF TECHNICAL TERMS

Duplicate. A probe image of a person whose corresponding gallery image was taken from a different image set. Usually, a duplicate is taken on a different day than the corresponding gallery image.

Duplicate I probes. Set of duplicate probes for a gallery.

Duplicate II probes. Set of duplicate probes where there is at least one year between the acquisition of the corresponding probe and gallery images.

FB probes. Probes taken from the same image set as the corresponding gallery images.

fc probes. Probes taken on the same day, but with different illumination from the corresponding gallery images.

Fully automatic algorithm. An algorithm that can locate a face in an image and recognize the face.

Gallery. In computing performance scores, images of the set of known individuals. The gallery is used in computing performance after a FERET test is administered. A gallery is a subset of a target set. A target set can generate multiple galleries.

Probe. Image containing the face of an unknown individual that is presented to an algorithm to be recognized. Probe can also refer to the identity of the person in a probe image.

Partially automatic algorithm. An algorithm that requires that the centers of the eyes are provided prior to recognizing a face.

Probe set. A set of probe images used in computing algorithm performance. The probe set is used in computing performance after the FERET test is administered. A probe set is a subset of a query set. A query set can generate multiple probe sets.

Query set. The set of unknown images presented to the algorithm when a test is administered. See probe set.

Target set. The set of known images presented to the algorithm when a test is administered. See gallery.

ACKNOWLEDGMENTS

The work reported here is part of the Face Recognition Technology (FERET) program, which is sponsored by the US Department of Defense Counterdrug Technology Development Program. Portions of this work was done while P.J. Phillips was at the US Army Research Laboratory (ARL). P.J. Phillips would like to acknowledge the support of the National Institute of Justice.

REFERENCES

[1] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *J. Optical Soc. Am. A,* vol. 14, pp. 1,724-1,733, Aug. 1997.

[2] R.M. McCabe, "Best Practice Recommendation for the Capture of Mugshots Version," 1997. 2.0.http://www. nist. gov/itl/div894/894.03/face/face.html.

[3] B. Moghaddam, C. Nastar, and A. Pentland, "Bayesian Face Recognition Using Deformable Intensity Surfaces," *Proc. Computer Vision and Pattern Recognition '96,* pp. 638-645, 1996.

[4] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 7, pp. 696-710, July 1997.

[5] H. Moon and P.J. Phillips, "Analysis of PCA-Based Face Recognition Algorithms," *Empirical Evaluation Techniques in Computer Vision,* K.W. Bowyer and P.J. Phillips, eds., pp. 57-71, Los Alamitos, Calif.: IEEE CS Press, 1998.

[6] P.J. Phillips and P. Rauss, "The Face Recognition Technology (FERET) Program," *Proc. Office of Nat'l Drug Control Policy, CTAC Int'l Technology Symp.,* pp. 8-11, Aug. 1997.

[7] P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET Database and Evaluation Procedure for Face-Recognition Algorithms," *Image and Vision Computing J.,* vol. 16, no. 5, pp. 295-306, 1998.

[8] P. Rauss, P.J. Phillips, A.T. DePersia, and M. Hamilton, "The FERET (Face Recognition Technology) Program," *Surveillance and Assessment Technology for Law Enforcement, SPIE,* vol. 2,935, pp. 2-11, 1996.

[9] S. Rizvi, P.J. Phillips, and H. Moon, "The FERET Verification Testing Protocol for Face Recognition Algorithms," Technical Report NISTIR 6,281, Nat'l Inst. Standards and Technology, http://www.nist.gov/itl/div894/894.03/pubs.html#face. 1998.

[10] S. Rizvi, P.J. Phillips, and H. Moon, "The FERET Verification Testing Protocol for Face Recognition Algorithms," *Image and Vision Computing J.,* to appear.

[11] D. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, Aug. 1996.

[12] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[13] J. Wilder, "Face Recognition Using Transform Coding of Grayscale Projection Projections and the Neural Tree Network," *Artifical Neural Networks with Applications in Speech and Vision,* R.J. Mammone, ed., pp. 520-536, Chapman Hall, 1994.

[14] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 7, pp. 775-779, July 1997.

[15] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant Analysis of Principal Components for Face Recognition," *Proc. Third Int'l Conf. Automatic Face and Gesture Recognition,* pp. 336-341, 1998.

[16] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng, "Discriminant Analysis of Principal Components for Face Recognition," *Face Recognition: From Theory to Applications,* H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie, and T.S. Huang, eds., pp. 73-85, Berlin: Springer-Verlag, 1998.

P. Jonathon Phillips received the BS degree in mathematics in 1983 and the MS in electronic and computer engineering in 1985 from George Mason University and the PhD degree in operations research from Rutgers University in 1996. He is a leading technologist in the fields of computer vision, biometrics, face recognition, and human identification. He works at the National Institute of Standards and Technology (NIST), where he is currently detailed to the Defense Advanced Projects Agency to manage the Human Identification at a Distance (HumanID) program. Prior to this, he served as project leader for the Visual Image Processing Group's Human Identification project. His current research interests include computer vision, identifying humans from visual imagery, face recognition, biometrics, digital video processing, developing methods for evaluating biometric algorithms, and computational psycho-physics. Prior to joining NIST, he directed the Face Recognition Technology (FERET) program at the US Army Research Laboratory. He developed and designed the FERET database collection and FERET evaluations, which are the de facto standards for the face recognition community. Also, he has conducted research in face recognition, biomedical imaging, computational psychophysics, and autonomous target recognition. Dr. Phillips was codirector of the NATO Advanced Study Institute on Face Recognition: From Theory to Applications, coorganizer of the First and Second Workshops on Empirical Evaluation Methods for Computer Vision Algorithms, and coprogram chair of the Second International Conference on Audio and Video-Based Biometric Authentication. He is guest coeditor of the special section on empirical evaluation of computer algorithms in the *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* and the special issue of *Computer Vision and Image Understanding (CVIU)* on empirical evaluation. He has coedited two books. The first, *Face Recognition: From Theory to Applications*, was coedited with Harry Wechsler, Vicki Bruce, Francoise Fogelman-Soulie, and Thomas Huang. The second, *Empirical Evaluation Techniques in Computer Vision*, was coedited with Kevin Bowyer. He is a member of the IEEE.

Hyeonjoon Moon received the BS degree in electronics and computer engineering from Korea University, Seoul, in 1990, the MS, and PhD degrees in electrical and computer engineering from the State University of New York at Buffalo, in 1992 and 1999, respectively. From 1993 to 1994, he was a systems engineer at Samsung Data Systems in Seoul, Korea. From 1996 to 1999, he was a research associate at the US Army Research Laboratory in Adelphi, Maryland. Currently, he is a senior research scientist at Lau Technologies in Littleton, Massachusetts. His research interests include image processing, neural networks, computer vision, and pattern recognition. He is a member of the IEEE.

Syed A. Rizvi (S'92-M'96) received the BSc degree (honors) from the University of Engineering and Technology, Lahore, Pakistan, the MS, and PhD degrees from the State University of New York (SUNY) at Buffalo, in 1990, 1993, and 1996, respectively, all in electrical engineering. From May 1995 to July 1996, he was a research associate with the US Army Research Laboratory, Adelphi, Maryland, where he developed coding and automatic target recognition algorithms for FLIR imagery. Since September 1996, he has been an assistant professor with the Department of Engineering Science and Physics at the College of Staten Island of the City University of New York. From 1996 to 1998, he was a consultant with the US Army Research Laboratory, collaborating in research for image compression and automatic target recognition. His current research interests include image and video coding, applications of artificial neural networks to image processing, and automatic target recognition. He has published more than 60 technical articles in his area of research. He is a member of SPIE and the IEEE.

Patrick J. Rauss received the BS degree in engineering physics from Lehigh University in 1987. He received the MEng degree in applied remote sensing and geo-iInformation systems from the University of Michigan in 2000. He has worked as a civilian researcher for the US Army since 1988, first with the Night Vision and Electro-Optics Directorate and, since 1992 with the Army Research Laboratory's EO-IR Image Processing Branch. He worked closely with Dr. Phillips on the FERET program from 1995 to 1997. His current research interests are automated processing of hyperspectral imagery for material classification and using supervised, adaptive learning techniques for hyperspectral and computer vision applications. Over the years, he has developed signal and image processing tools and techniques for a wide range of sensors including X-ray fluorescent spectrometers, midwave and long-wave infrared radiometers, Fourier transform infrared spectrometers, forward looking infrared imagers, and hyperspectral imagers.