ARMY RESEARCH LABORATORY

FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results

by P. Jonathon Phillips, Patrick J. Rauss, and Sandor Z. Der

ARL-TR-995 October 1996

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TR-995 **October 1996**

FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results

P. Jonathon Phillips, Patrick J. Rauss, Sandor Z. Der
Target Recognition Branch

sponsored by

DARPA
3701 N. Fairfax Drive
Arlington, VA 22203-1714

Abstract

As part of the Face Recognition Technology (FERET) program, the U.S. Army Research Laboratory (ARL) conducted supervised government tests and evaluations of automatic face recognition algorithms. The goal of the tests was to provide an independent method of evaluating algorithms and assessing the state of the art in automatic face recognition. This report describes the design and presents the results of the August 1994 and March 1995 FERET tests. Results for FERET tests administered by ARL between August 1994 and August 1996 are reported.

Contents

Appendices

Figures

Figures (cont'd)

Tables

1. Introduction

The primary mission of the Face Recognition Technology (FERET) program is to develop automatic face recognition capabilities that can be employed to assist security, intelligence, and law enforcement personnel in the performance of their duties. In order to achieve its objectives, the FERET program is conducting multiple tasks over a three-year period from September 1993. The FERET program is sponsored by the Department of Defense Counterdrug Technology Development Program through the Defense Advanced Research Projects Agency (DARPA), with the U.S. Army Research Laboratory (ARL) serving as technical agent.

The program has focused on three major tasks. The first major FERET task is the development of the technology base required for a face recognition system.

The second major task, which began at the start of the FERET program and will continue throughout the program, is collecting a large database of facial images. This database of facial images is a vital part of the overall FERET program and promises to be key to future work in face recognition, because it provides a standard database for algorithm development, test, and evaluation. The database is divided into two parts: the development portion, which is given to researchers, and the sequestered portion, which is used to test algorithms.

The third major task is government-monitored testing and evaluation of face recognition algorithms using standardized tests and test procedures. Two rounds of government tests were conducted, one at the end of Phase I (the initial development phase, ending in August 1994) and a second midway through Phase II (the continuing development phase), in March 1995. (A followup test was administered for one of the algorithms in August 1996; results are reported in app A.)

The purpose of the tests was to measure overall progress in face recognition, determine the maturity of face recognition algorithms, and have an independent means of comparing algorithms. The tests measure the ability of the algorithms to handle large databases, changes in people's appearance over time, variations in illumination, scale, and pose, and changes in the background. The algorithms tested are fully automatic, and the images presented to the algorithm are not normalized. If an algorithm requires that a face be in a particular position, then the algorithm must locate the face in the image and transform the face into the required predetermined position.

The August 1994 evaluation procedure consisted of a suite of three tests. The first test is the large gallery test. A *gallery* is the collection of images of individuals known to the algorithm, and a *probe* is an image of an unknown person presented to the algorithm. In the August 1994 test, the gallery consisted of 317 individuals, with one image per person, and in the March 1995 test, the gallery consisted of 831 individuals, with one image per person. The differences between a probe image and a gallery image of

a person include changes in time (the images were taken weeks or months apart); changes in scale; changes in illumination; and changes in pose.

Images in the FERET database were taken under semi-controlled conditions. This is in contrast to many of the algorithms in the literature, where results are reported for small databases collected under highly controlled conditions.

The second and third tests are the false-alarm and rotation tests. The goal of the false-alarm test is to see if an algorithm can successfully differentiate between probes that are in the gallery and those not in the gallery. The rotation test measures the effects of rotation on recognition performance.

As part of the FERET program, a procedure was instituted to allow researchers outside the FERET program to gain access to the FERET database (see app B for details).* Also, researchers can request to take the FERET tests. Results of future tests will be reported in supplements to this report that will be issued as needed.

Future FERET tasks will include the development of real-time systems to demonstrate face recognition in real-world situations. These demonstration systems will provide the needed large-scale performance statistics for evaluation of algorithms in real-world situations. This decision to proceed with the development of real-time systems was based in part on the results from the March 1995 test.

This report reviews algorithms developed under the FERET program and the data collection activities, and reports on the results of the August 1994 and March 1995 government-supervised tests.

**At the time of the test, the FERET database was made available to researchers in the U.S. on a case by case basis. Distribution was restricted to the U.S. because of legal issues concerning the rights of individuals to their facial images. As of May 1996, over 50 researchers had been given access to the FERET database.*

2. Overview

The object of the FERET program is to develop face recognition systems that can assist intelligence, security, and law enforcement personnel in identifying individuals electronically from a database of facial images. Face recognition technology could be useful in a number of security and law enforcement tasks:

- automated searching of mug books using surveillance photos, mug shots, artist sketches, or witness descriptions;
- controlling access to restricted facilities or equipment;
- credentialing of personnel for background and security checks;
- monitoring areas (airports, border crossings, secure manufacturing facilities, doorways, hallways, etc) for particular individuals; and
- finding and logging multiple appearances of individuals over time in surveillance videos (live or taped).

Other possible government and commercial uses of this technology could be

- verifying identity at ATM machines;
- verifying identity for the automated issue of driver's licenses; and
- searching photo ID records for fraud detection (multiple driver's licenses, multiple welfare claims, etc).

The FERET program has concentrated on two scenarios. The first is the electronic mug book, a collection of images of known individuals—in other words, a gallery. The image of an individual to be identified (a *probe*) is presented to an algorithm, which reports the closest matches from a large gallery. The performance of the algorithm is measured by its ability to correctly identify the person in the probe image. For example, an image from a surveillance photo would be a probe, and the system would display the photos of the 20 people from the gallery that most resembled the unknown individual in the surveillance photo. The final decision concerning the person's identity would be made by a trained law enforcement agent.

The second scenario is the identification of a small group of specific individuals from a large population of unknown persons. Applications for this type of system include access control and the monitoring of airports for suspected terrorists. In the access control scenario, when an individual walks up to a doorway, his or her image is captured, analyzed, and compared to the gallery of individuals approved for access. Alternatively, the system could monitor points of entry into a building, a border crossing, or perhaps an airport jetway, and search for smugglers, terrorists, or other criminals attempting to enter surreptitiously. In both situations, a large number of individuals not in the gallery would be presented to the system.

The important system performance measures here are the probabilities of false alarms and missed recognitions. A false alarm occurs when the algorithm reports that the person in a probe image is in the gallery when that person is not in fact in the gallery. A missed recognition is the reverse: the algorithm reports that the person in the probe is not in the gallery when the person is in the gallery, or identifies the person as the wrong person.

The primary emphasis of the FERET program has been to establish an understanding of the current state of the art in face recognition from frontal images and to advance it. Additionally, the program has established a baseline for the performance of recognition algorithms on rotated facial images. Later phases of the program will extend successful approaches to the task of identifying individuals when facial features are presented in any aspect from full front to full profile.

To address these tasks, a multiphase program was instituted by DARPA, with ARL as the technical agent. In Phase I (September 1993 through September 1994), five contracts were awarded for algorithm development and one contract for database collection. Phase II continued the database collection contract and exercised options on three of the algorithm development contracts.

Before the start of the FERET program, there was no way to accurately evaluate or compare the face recognition algorithms in the literature. Various researchers collected their own databases under conditions relevant to the aspects of the problems that they were examining. Most of the databases were small and consisted of images of less than 50 individuals. Notable exceptions were databases collected by three primary researchers:

(1) Alex Pentland of the Massachusetts Institute of Technology (MIT) assembled a database of ~7500 images that had been collected in a highly controlled environment with controlled illumination; all images had the eyes in a registered location, and all images were full frontal face views.

(2) Joseph Wilder of Rutgers University assembled a database of ~250 individuals collected under similarly controlled conditions.

(3) Christoph von der Malsburg of the University of Southern California (USC) and colleagues used a database of ~100 images that were of controlled size and illumination but did include some head rotation.

3. Database

A standard database of face imagery is essential for the success of this project, both to supply standard imagery to the algorithm developers and to supply a sufficient number of images to allow testing of these algorithms. Harry Wechsler at George Mason University (GMU) directed the effort to collect a database of images for development and testing (contract number DAAL01-93-K-0099).

The images of the faces are initially acquired with a 35-mm camera. The film used is color Kodak Ultra. The film is processed by Kodak and placed onto a CD-ROM via Kodak's multiresolution technique for digitizing and storing digital imagery. At GMU, the color images are retrieved from the CD-ROM and converted into 8-bit gray-scale images. After being assigned a unique file name, which includes the subject's identity number, the images become part of the database. The identity number is keyed to the person photographed, so that any future images collected on this person will have the same ID number associated with the images. The images are stored in TIFF format and as raw 8-bit data. The images are 256 pixels wide by 384 pixels high. Attempts were made to keep the interocular distance (the distance between the eyes) of each subject to between 40 and 60 pixels. The images consist primarily of an individual's head, neck, and sometimes the upper part of the shoulders.

The images are collected in a semi-controlled environment. To maintain a degree of consistency throughout the database, the same physical setup is used in each photography session. However, because the equipment must be reassembled for each session, there is some variation over collections from site to site (fig. 1).

The facial images were collected in 11 sessions from August 1993 through December 1994. Sessions were primarily conducted at GMU, with several collections done at ARL facilities. The duration of a session was one or two days, and the location and setup did not change during a session. Taking the images at different locations introduced a degree of variation in the images from one session to another session, which reflects real-world applications.

A photography session is usually performed by a photographer and two assistants. One assistant briefs each volunteer and obtains a written release form (see app C). (A release form is necessary because of the privacy laws in the United States.) The other assistant directs the subject to turn his or

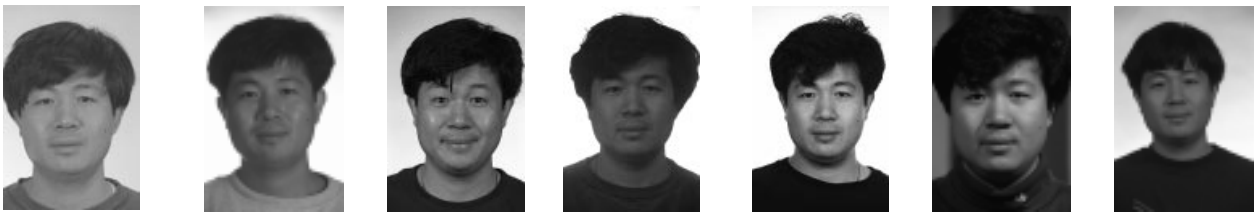


Figure 1. Examples of variations among collections.

her head to the various poses required. The images were collected at different locations, so there is some variation in illumination from one session to another. A neutral colored roll of paper was used as a standard background in the images. Subjects wearing glasses were asked to remove them.

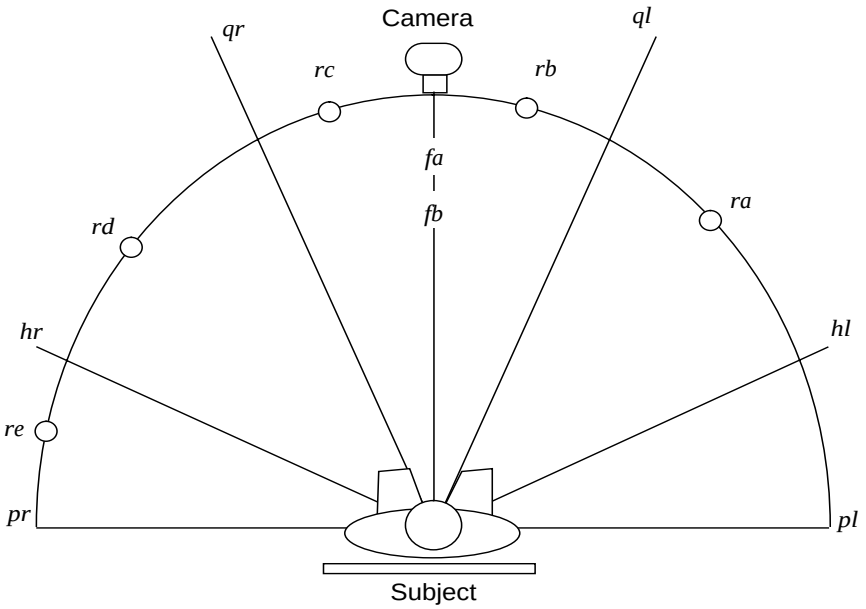
The photographs were collected under relatively unconstrained conditions. For the different poses, the subjects were asked to look at marks on the wall, where the marks corresponded to the aspects defined below.

Some questions were raised about the age, racial, and sexual distribution of the database. However, at this stage of the program, the key issue was algorithm performance on a database of a large number of individuals.

A set of images of an individual is defined as consisting of a minimum of five and often more views (see fig. 2 and 3). Two frontal views are taken, labeled *fa* and *fb*. One is the first image taken (*fa*) and the other, *fb*, usually the last. The subject is asked to present a different facial expression for the *fb* image. Images are also collected at the following head aspects: right and left profile (labeled *pr* and *pl*), right and left quarter profile (*qr*, *ql*), and right and left half profile (*hr*, *hl*). Additionally, five extra locations (*ra*, *rb*, *rc*, *rd*, and *re*), irregularly spaced among the basic images, are collected if time permits. Some subjects also are asked to put on their glasses and/or pull their hair back to add some simple but significant variation in the images.

Each individual in the database is given a unique ID number. The ID number is part of the file name for every image of that person, including images from different sets. In addition, the file name encodes head aspect, date of collection, and any other significant point about the image collected; table 1 gives a detailed description of the image name convention.

Figure 2. Possible aspects collected of subject face.



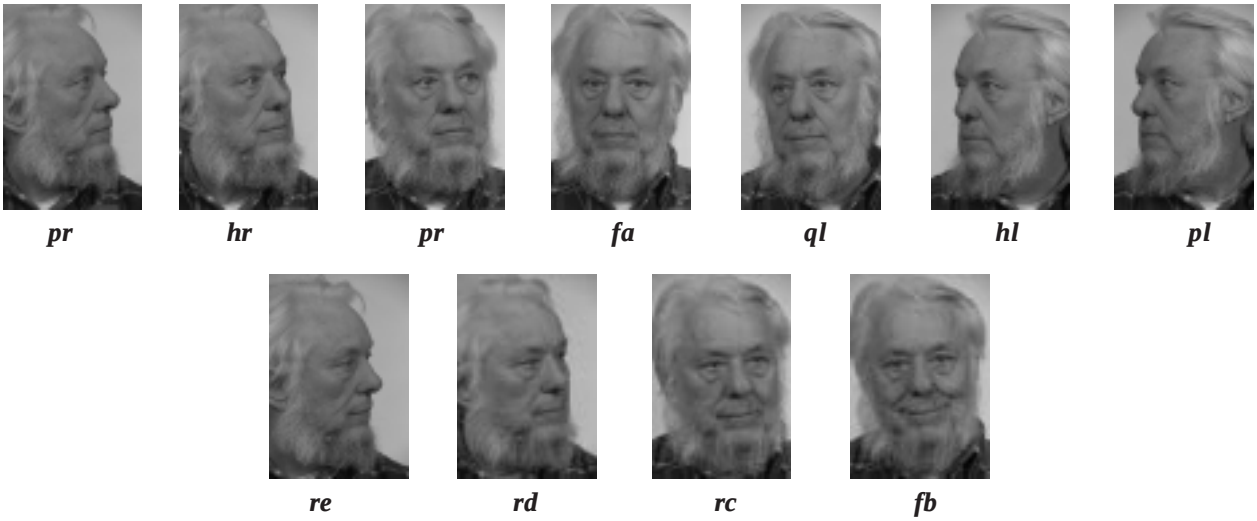


Figure 3. Typical set of images collected in one sitting.

Table 1. Image file name description.

Example file name: 00346hr001c.931230

a c d e

Segment	Category	Code		Explanation
a	ID No.	*nnnnn*		Unique for each individual.
b	Pose	fa		Full face or frontal: first shot.
		fb		Full face or frontal: last shot.
		qr, ql		Quarter profile, right and left.
		hr, hl		Half profile, right and left.
		pr, pl		Full profile, right and left.
		ra, rb, rc, rd, re		Arbitrary (random) positions (see fig. 1).
c	Special flags	(Left flag)	0	Image not releasable for publication.
			1	Image may be used for publication if authorized.
		(Right flag)	0	ASA-200 negative film used for collection.
			1	ASA-400 negative film used for collection.
		(Middle flag)	0	Image not histogram adjusted.
			1	Image histogram adjusted.
d	Special circumstances	a		Glasses worn.
		b		Duplicate with different hair length.
		c		Glasses worn and different hair length.
		d		Electronically scaled and histogram adjusted.
		e		Clothing color changed electronically.
		f		Image brightness reduced by 40%.
		g		Image brightness reduced by 60%.
		h		Image scale reduced 10%.
		i		Image scale reduced 20%.
		j		Image scale reduced 30%.
e	Date	*yymmdd*		Date image taken.

A set of images is referred to as a duplicate set if the person in the set is in a previously collected set. Some people have images in the database spanning nearly a year between their first sitting and their most recent one. A number of subjects have been photographed several times (fig. 1).

At the end of Phase I (August 1994), 673 sets of images had been collected and entered into the imagery database, resulting in over 5000 images in the database. At the time of the Phase II test (March 1995), 1109 sets of images were in the database, for 8525 total images. There were 884 individuals in the database and 225 duplicate sets of images.

The primary goal of the image collection activities in the fall of 1994 was to support the March 1995 test. Approximately 300 sets of images were given out to algorithm developers as a developmental data set, and the remaining images were sequestered by the government for testing purposes.

As an aid in the evaluation of the algorithms' robustness with respect to specific variables, the sequestered database was augmented with a set of digitally altered images. The database collectors changed the illumination levels of 40 images by using the MATLAB Image Processing Tool Box command "brighten ()," using values of –0.4 and –0.6 to create images with the illumination levels reduced by approximately 40 and 60 percent, respectively. The function that changes the illumination is nonlinear. To test sensitivity to scale changes, they electronically modified 40 images to show 10-, 20-, and 30-percent reductions of scale along each axis, using the MATLAB Image Processing Tool Box command "imresize ()." This command uses a low-pass filter on the original image to avoid aliasing, and bilinear interpolation to find each pixel density in the reduced image. This approximates obtaining the images at a greater distance from the camera. Finally, using Adobe Photoshop's paint brush tool, the database collectors electronically modified portions of clothing in several of the images to reverse the contrast. We had this done to see if any algorithms were using cues from clothing for recognition.

4. Phase I

4.1 Algorithm Development

The FERET program was initiated with an open request for proposals (RFP); 24 proposals were received and evaluated jointly by DoD and law enforcement personnel. The winning proposals were chosen based on their advanced ideas and differing approaches. In Phase I, five algorithm development contracts were awarded. The organizations and principal investigators for Phase I were

- MIT, Alex Pentland (contract DAAL01-93-K-0115);
- Rutgers University, Joseph Wilder (contract DAAL01-93-K-0119);
- The Analytic Science Company (TASC), Gale Gordon (contract DAAL01-93-K-0118);
- University of Illinois at Chicago (UIC) and University of Illinois at Urbana-Champaigne, Lewis Sadler and Thomas Huang (contract DAAL01-93-K-0114); and
- USC, Christoph von der Malsburg (contract DAAL01-93-K-0109).

Only information and results for contracts that were extended into Phase II are given in this report; for brief descriptions of the individual approaches, see appendix C.

4.2 Test Procedure

Three distinct tests were conducted, each with its own probe and gallery set. The large gallery test evaluates the algorithm performance on a large gallery of images, the false-alarm test evaluates the false-alarm performance of the algorithm, and the rotation test was designed to baseline algorithm performance on nonfrontal (rotated) images.

TASC and USC were tested on 1 to 3 August 1994, and MIT, UIC, and Rutgers on 8 to 10 August 1994. Government representatives arrived at each of the testee's sites to administer the test. The government representative brought two 8-mm computer data tapes for each test to the contractor's site. The first tape of each test contained the gallery, and the second tape contained the probe images.

All images were processed while the government representative was present. Results from the test were recorded, and the government representative took the results back to the government facilities for scoring. At the conclusion of the test, both the gallery and probe data were removed from the testee's computer system and the tapes returned to the government.

To ensure that matching was not done by file name, the government gave the gallery and probe sets random file ID numbers, and kept the links between the file name and ID number from the contractors by supplying only the ID number as the labels for the gallery and probe sets for the test.

A "pose flag" was also supplied for each image, as this information would be expected from the hypothetical "face detection" front-end that supplies the localized faces to the classification algorithm. The pose flag tells the pose of the face in the image at the time of collection. The flags are *fa*, *ql*, *qr*, *hl*, *hr*, *pl*, and *pr*—the same pose flags as in the FERET database.

The computation time of the algorithms was not measured or considered as a basis for evaluation. However, the algorithms had to be able to perform the tests on a few standard workstation-type computers over three days. The rationale for this restriction was to ensure that an algorithm was not so computationally intensive as to preclude it being implemented in a real-time system.

4.3 Test Design

The August 1994 FERET evaluation procedure consisted of a suite of three tests designed to evaluate face recognition algorithms under different conditions. The results from the suite of tests present a robust view of an algorithm and allow us to avoid judging algorithm performance by one statistic.

The first test, the large gallery test, measures performance against large databases. The main purpose of this test was to baseline how algorithms performed against a database when the algorithm had not been developed and tuned with a majority of the images in the gallery and probe sets.

The second test, the false-alarm test, measures performance when the gallery is significantly smaller than the probe set. This test models monitoring an airport or port of entry for suspected terrorists where the occurrence of the suspects is rare.

The third test, the rotation test, baselines performance of the algorithm when the images of an individual in the gallery and probe set have different poses. Although difficult, this is a requirement for numerous applications. This test was used only to establish a baseline for future comparisons, because the rotation problem was out of the scope of the FERET program.

The algorithms tested are fully automatic. The processing of the gallery and the probe images is done without human intervention. The input to the algorithms for both the gallery and the probe is a list of image names along with the nominal pose of the face in the image. The images in the gallery and probe sets are from both the developmental and sequestered portions of the FERET database. Only images from the FERET database are included in the test. Algorithm developers were not prohibited from using images outside the FERET database to develop their algorithms or tune parameters in their algorithms. The faces in the images were not placed in

a predetermined position or normalized. If required, prepositioning or normalization must be performed by the face recognition system.

The large gallery test examines recognition rates from as large a database as was available at the time. The probe set consists of all the individuals in the gallery, as well as individuals not in the gallery. For this test, the gallery consisted of 317 frontal images (one per person), and the probe set consisted of 770 faces; table 2 gives a breakdown of the gallery and probe images by category.

Each set of facial images includes two frontal images (*fa* and *fb* images), as shown in figure 3. One of these images is placed in the gallery and referred to as the FA image. The frontal image that is not placed in the gallery is placed in the probe set and called the FB image. The image (*fa* or *fb*) to be designated the FA image can be selected manually or randomly. In the August 1994 test, all the *fa* images were selected to be the FA images. In the March 1995 test, the process was random, with a 50/50 chance of the *fa* or *fb* image being selected as the FA image.

For diagnostic purposes, 48 FA images were placed in the probe set. For these images, the algorithms should produce exact matches with their copies in the gallery. Some probe images were not in the gallery, by which we mean that the person whose image was in the probe was not in one of the gallery images. Duplicate images are images of people in the gallery taken from a duplicate set of images of that person (see sect. 3 for a definition and description of duplicate sets of images). All the duplicates are frontal images. Quarter and half rotations are those images with head rotation as indicated (*hl*, *hr*, *ql*, and *qr*, as shown in fig. 2 and 3). The remaining categories consist of the electronically altered frontal images discussed in section 3.

Table 2. Type and number of images used in gallery and probe set for large gallery test.

Image category	Number
Gallery images:	
FA frontal images	317
Probe images:	
FA frontal images	48
FB frontal images	316
Frontal probes not in gallery	50
Duplicates	60
Quarter rotations	26
Half rotations	48
40% change in illumination	40
60% change in illumination	40
10% reduction in scale	40
20% reduction in scale	40
30% reduction in scale	40
Contrast-reversed clothes	22
Total probes	770

The false-alarm test evaluates the false-alarm performance of the algorithms. The system is presented with a small gallery and a large probe set, with many individuals unmatched in the gallery. All images for this test were full frontal face images. For this test, a gallery of 25 frontal faces (one image per person) was supplied. The probe set consisted of 305 images; table 3 gives the type and number of the various images.

We conducted the rotation test to examine algorithm robustness under head rotations. A gallery of 40 quarter-rotated (*qr* or *ql* images) and 40 half-rotated (*hl* or *hr*) images (one per person) was supplied and tested with the probe set defined in table 4.

Because the approach that TASC uses requires matched face / profile pairs (see app C), TASC could not use the same test gallery and probe sets. Therefore, a special test set was generated for evaluating the performance of the TASC approach. For the large gallery test, the gallery consisted of 266 image pairs, with the probe set defined in table 5. For the August 1994 test, the reporting of confidence values was optional, and TASC elected not to report the confidence scores. Thus, it was not possible to construct a receiver operator curve (ROC) for TASC, and results are not reported for the false-alarm test. (The decision to construct an ROC was made after TASC took the test.) Because the TASC algorithm required frontal / profile pairs, it could not be tested for rotation. Hence, the rotation test was not taken.

Table 3. Type and number of images used in gallery and probe set for false-alarm test.

Image category	Number
Gallery images:	
FA frontal images	25
Probe images:	
FB frontal images	25
Frontal probe images not in gallery	204
40% change in illumination	10
60% change in illumination	9
10% reduction in scale	19
20% reduction in scale	19
Contrast-reversed clothes	19
Total probes	305

Table 4. Type and number of images used in gallery and probe set for rotation test.

Image category	Number
Gallery images:	
Quarter rotations	40
Half rotations	40
Total gallery	80
Probe images:	
Quarter rotations (*qr*,*ql*)	85
Probes not in gallery (*fa*,*fb*,*qr*,*ql*,*hl*,*hr*)	50
Intermediate rotations (*fa*,*fb*,*hl*,*hr*)	90
Total probes	225

Table 5. Type and number of images used in gallery and probe set in large gallery test for TASC.

Image category	Number
Gallery images:	
FA frontal profile image pairs	266
Probe images:	
Frontal profile image pairs	249
FB frontal profile pairs not in gallery	25
40% change in illumination	10
60% change in illumination	8
10% reduction in scale	14
20% reduction in scale	14
30% reduction in scale	28
Total probes	378

4.4 Output Format

The contractors were requested to supply the test results in an ASCII file in the following format: the probe ID number being tested, a rank counter, the gallery ID number of a match, and a false-alarm flag that indicates whether the algorithm determined that the probe was in the gallery or not (1 if the algorithm reported that the probe was in the gallery and 0 if the probe was reported as not in the gallery). Also requested was the confidence score of the match; see table 6 for an example of an output file. The score of the match is a number that measures the similarity between a probe and an image in the gallery. Each algorithm used a different measure of similarity, and it is not possible to directly compare similarity measures between different algorithms. Reporting the similarity measure was optional on the August 1994 test. All algorithm developers except for TASC reported this number. For the August 1994 large gallery test, all algorithm developers reported the top 50 gallery matches in ranked order for each probe. For the false-alarm test, the top 25 (the size of the gallery) were reported, and in the rotation test, the top 25 were reported.

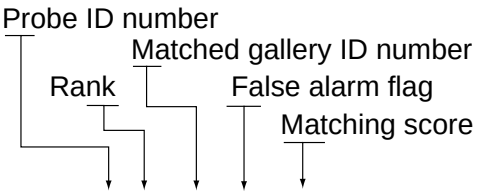
No testing was done to determine how the algorithms would respond to a face-like piece of clutter that might be forwarded to the recognition algorithm from the face detection front-end. Tests of this nature will have to wait until detection and recognition algorithms are interfaced together in a full demonstration system.

4.5 Calculation of Scores

The results for the FERET phase I and II tests are reported by two sets of performance statistics. One is the cumulative matched versus rank (cumulative match) and the other is the receiver operator curve (ROC). Both scores are computed from the output files provided by the algorithm developers (sect. 4.4). The selection of which score is computed depends on the test and analysis being performed.

The performance results for the large gallery test and the rotation test are reported by a graph of the cumulative match score. Performance scores are

Table 6. Example of a results file.



1	3	45	1	87.34
1	2	45	1	75.45
1	3	111	1	67.23
⋮				
1	50	231	0	11.56

reported for a number of subsets of the probe set. It is not possible to compute the cumulative match score for the entire probe set, because the probe set contains probes that are not in the gallery. For the large gallery test, we report the cumulative match score for the subset of all probes that have a corresponding match in the gallery and for all categories listed in table 2 (sect. 4.3), except the FA versus FA category. Probes not in the gallery are not counted towards the cumulative score.

In the large gallery test, each algorithm reports the top 50 matches for each probe, provided in a rank-ordered list (table 6). From this list one can determine if the correct answer of a particular probe is in the top 50, and if it is, how far down the list is the correct match. For example, for probe 1, if the correct match is with gallery image 22, and the match between probe 1 and gallery image 22 is ranked number 10 (the algorithm being tested reports that there are nine other gallery images that are better matches than gallery image 22), then we say that the correct answer for probe 1 is rank 10.

For a probe set we can find for how many probes the correct answer is ranked 5 or less. In the previous example, probe 1 would not be counted. The figures in this report show the percentage of probes that are of a particular rank or less. The horizontal axis is the rank, and the vertical axis the percentage correct. For example, for the MIT curve in figure 4 (sect. 4.6), the first box indicates that the correct answer was rank 1 for 80 percent of the probes, the box at position 2 indicates that the correct answer was rank 1 or 2 for ~82 percent of the probe images, that ~87 percent of the probes were of rank 10 or less, etc.

The following formula is used to compute scores for a given category. To make the explanation concrete, we use the class of duplicate images in the large gallery test. Let P be a subset of probe images in the probe set; e.g., P is the set of duplicate images in the large gallery test for USC. The number of images in P is denoted by $|P|$; in this example $|P|$ is 50. Let R_k be the number of probes in P that are ranked k or less; e.g., if $k = 10$, then $R_k = 43$. Thus, the percentage of probes that are rank k or less is R_k/P, or in the example case, $R_{10}/|P| = 43/50 = 0.86$ (fig. 6, sect. 4.6).

For the false-alarm test, an ROC is used to evaluate the algorithms. The ROC allows one to assess the trade-off between the probability of false

alarm and the probability of correct identification. In the false-alarm test, there are two primary categories of probes. The first are probes not in the gallery that generate false alarms. A false alarm occurs when an algorithm reports that one of these probes is in the gallery. The false-alarm rate is the percentage of probes not in the gallery that are falsely reported as being in the gallery. The false-alarm rate is denoted by P_F. The second category of probes is the set that is in the gallery. This set, characterized by the percentage of these probes that are correctly identified, is denoted by P_I. The pair of values P_I and P_F describe the operation of a system in an open universe; in an open universe, not every probe is in the gallery.

There is a trade-off between P_F and P_I. If every probe is tagged as a false alarm, then $P_F = 0$ and $P_I = 0$. At the other extreme, if no probes are declared to be false alarms, then $P_F = 1$ and P_I is the percentage of probes in the gallery with a rank 1. For an algorithm, performance is not characterized by a single pair of statistics (P_I,P_F) but rather by all pairs (P_I,P_F), and this set of values is an ROC (see fig. 16, sect 4.6.2: the horizontal axis is the false-alarm rate and the vertical axis the probability of correct identification). From the ROC it is possible to compare algorithms.

Say we are given algorithm **A** and algorithm **B**, along with a false-alarm rate for each, P_F^A and P_F^B, and a probability of correct identification for each, P_I^A and P_I^B. Algorithms **A** and **B** cannot be compared from the performance points (P_I^A, P_F^A) and (P_I^B, P_F^B). This is especially true if (P_I^A, P_F^A) and (P_I^B, P_F^B) are not close in value. The two systems may be operating at different points on the same ROC, or, for different values of P_F or P_I, one algorithm could have better performance.

For each P_F or P_I, an optimal decision rule could be constructed to maximize performance for the other parameter. For testing and evaluating algorithms, it is not practical to construct an ROC in this manner, and an approximation is used. For each probe, the algorithm reports the person in the gallery with which the probe is most similar, along with a confidence score. The test scorer obtains this information from the results file by reading the information about the highest ranked gallery image. Assume that a high confidence score implies greater likelihood that images are of the same person. Apply a threshold to the confidence score. The algorithm reports that the probe is not in the gallery if the confidence score is below the threshold. If the match score is greater than or equal to the threshold, then estimate the identity of the probe as the gallery image with the highest confidence score. A false alarm is a probe whose match score is greater than or equal to the threshold and is not in the gallery. Let $\hat{F}$ denote the number of false alarms. The probability of a false alarm is $P_F = \hat{F}/F^*$, where F^* is the number of probes in the probe set that are not in the gallery. A probe in the gallery is correctly identified if the algorithm reports the correct identity, and the match score is greater than or equal to the threshold. The probability of correct identification is $P_I = \hat{I}/I^*$, where $\hat{I}$ is the number of probes correctly identified, and I^* is the number of probes in the probe set that are in the gallery.

We generated the ROC by varying the threshold and recomputing P_F and P_I for each threshold. Initially, the threshold is set higher than the highest match score. This will generate the point $P_F = 0$ and $P_I = 0$. The threshold is incrementally lowered, and for each value, P_F and P_I are computed. The process of lowering the threshold will sweep out the ROC, and P_F and P_I will monotonically increase.

4.6 Results*

4.6.1 Large Gallery Test Performance

The results for the large gallery test are reported as cumulative match versus rank. Scores are presented for overall performance and for a number of different categories of probe images. Table 7 shows the categories corresponding to the figures presenting these results (fig. 4 to 15).

Figure 4 reports overall performance, where the probe set consisted of all probes for which there was a gallery image of the person in the probe. This includes the FA, FB, duplicate, rotation, and electronically altered images. The figure indicates the number of probe images scored for this category: e.g., for MIT there were 770 probes in the overall category, and for TASC there were 378 probes. This information is provided for all the figures. All scores in figures 4 and 6 to 15 were adjusted to take into account an error in the construction of the test set: 180 images that did not meet the requirements for the Phase 1 effort were mistakenly included in the gallery and had to be removed from all the scored results; in these images, the face took up much less of the field of view than had been specified. The annota-

Table 7. Figures reporting results for large gallery test.

Figure no.	Category title	Description of category
4	Adjusted overall match	Score for all probes in gallery, adjusted for 180 probes placed by mistake in probe set.
5	Unadjusted overall match	Score for all probes in gallery including 180 probes placed by mistake in probe set.
6	Duplicate match	Given a duplicate frontal image, find frontal match.
7	FA versus FB match	Given FB frontal image, find frontal match from same set.
8	Quarter match	Given quarter profile, find frontal match.
9	Half match	Given half profile, find frontal match.
10	10% scale match	Given an image reduced by 10%, find frontal match.
11	20% scale match	Given an image reduced by 20%, find frontal match.
12	30% scale match	Given an image reduced by 30%, find frontal match.
13	40% illumination match	Given an image with brightness reduced to 40%, find frontal match.
14	60% illumination match	Given an image with brightness reduced to 60%, find frontal match.
15 a	Clothes change—dark	Given an image with clothing contrast changed darker than original, find match.
15 b	Clothes change—light	Given an image with clothing contrast changed lighter than original, find match.

**Results are presented only for contractors whose funding was continued into Phase II.*

tion “adjusted” in the figures indicates that the scores were adjusted for this reason. However, MIT and USC voluntarily took the test with these more difficult images. Figure 5 shows a comparison of the overall performance on the uncorrected set of images, along with that for the adjusted set of probes. Figure 6 shows the performance on the duplicate frontal images. These scores are also adjusted for images that were unreadable because of computer media damage. Figure 7 shows the performance on the FB frontal images.

Figures 8 to 15 show performance for each of the remaining categories from table 2, except for the FA images and probes that are not in the gallery.

4.6.2 False-Alarm Test Performance

Figure 16 shows the ROC generated from the false-alarm test. We adjusted these values also to remove images that were unreadable because of computer media damage. We report only overall performance results for the entire probe set.

4.6.3 Rotated Gallery Test Performance

Figure 17 shows the results for the test examining the algorithms’ robustness under nonfrontal images in the gallery (also adjusted to omit unreadable images). We report only overall performance results for the entire probe set.

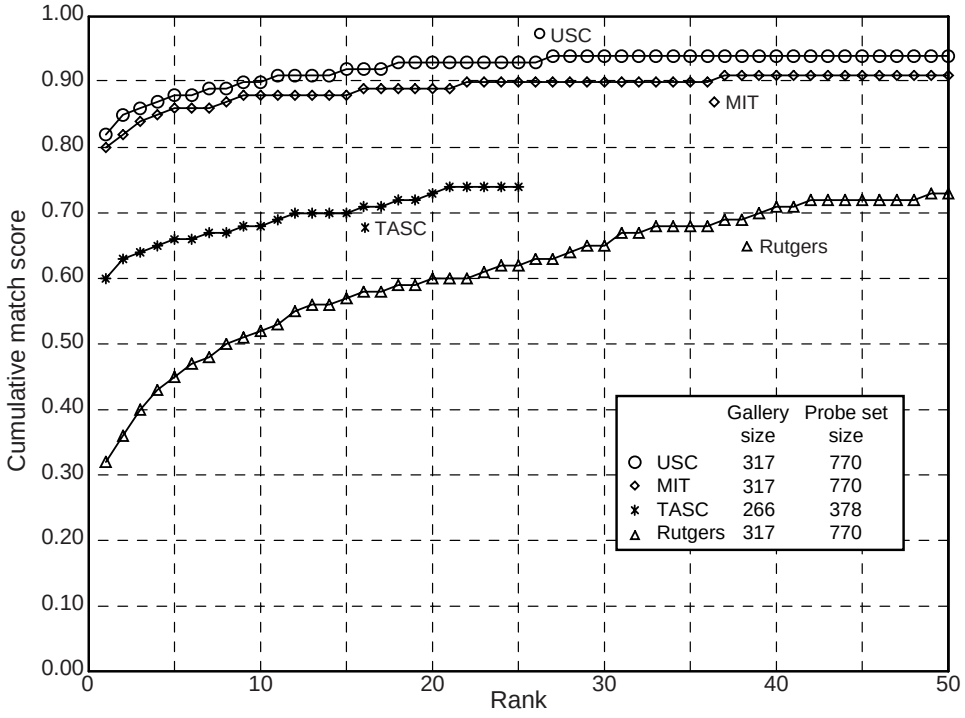


Figure 4. Large gallery test: overall scores, adjusted (August 1994).

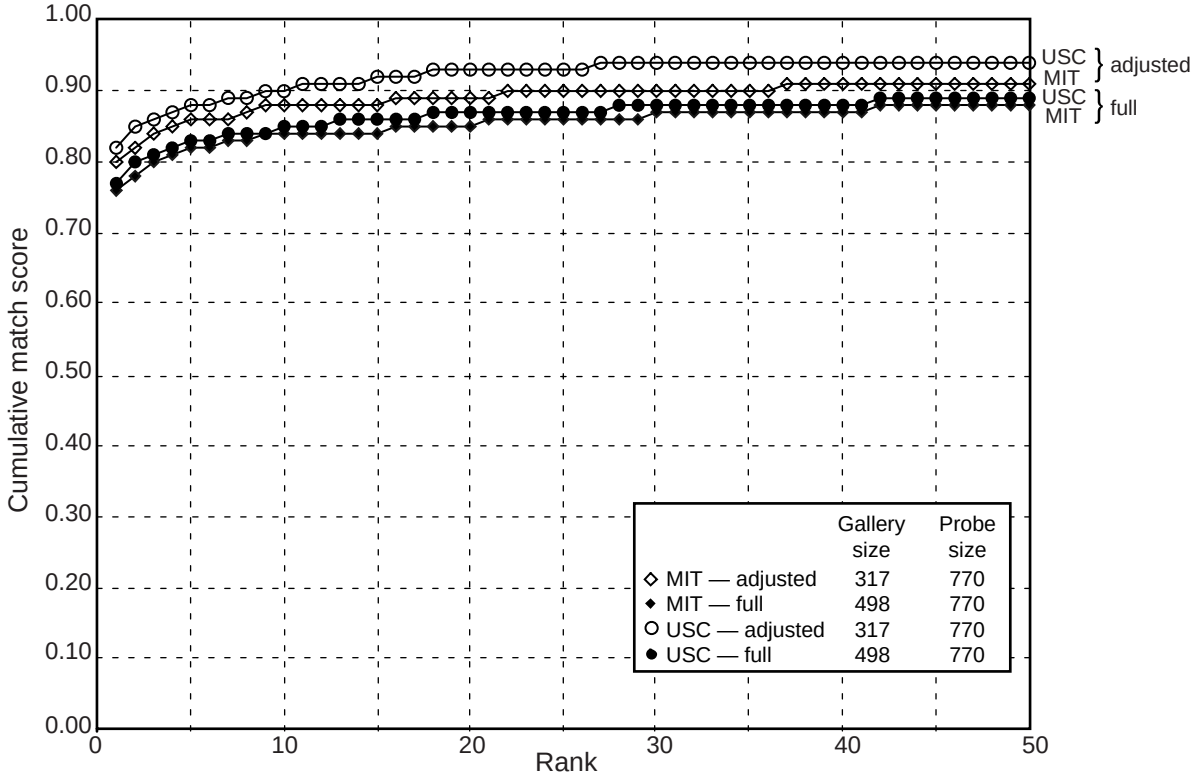


Figure 5. Large gallery test: overall scores: full set versus corrected set (August 1994).

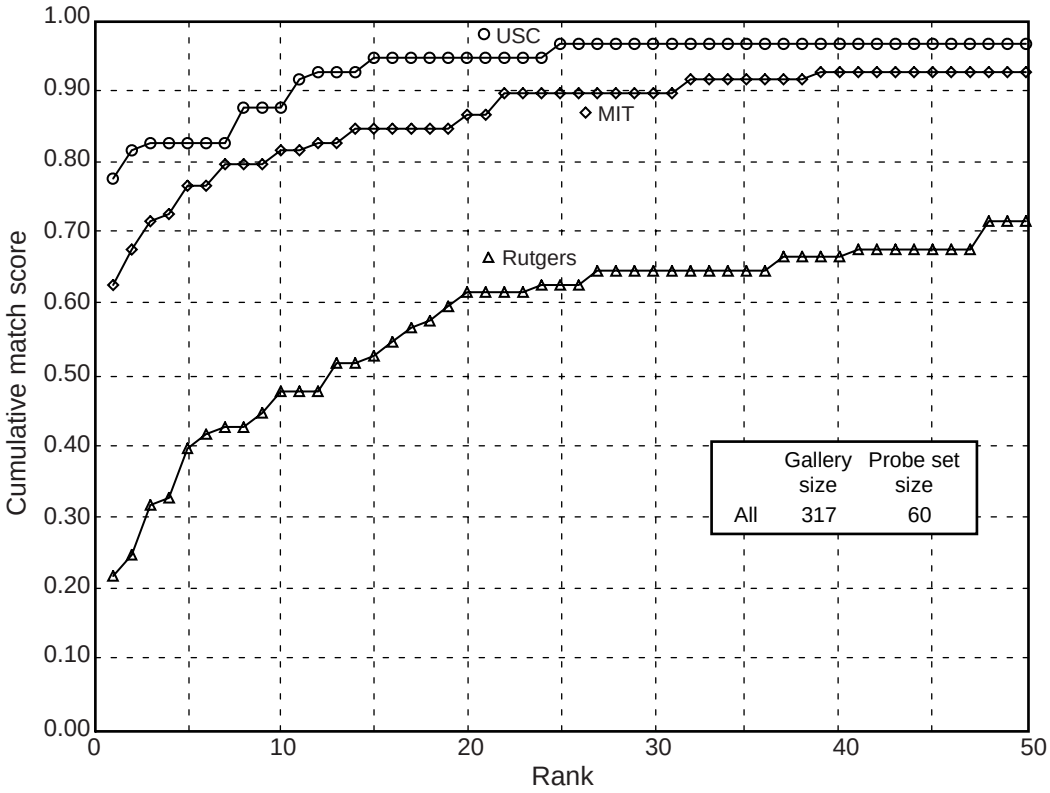


Figure 6. Large gallery test: duplicate scores: adjusted (August 1994).

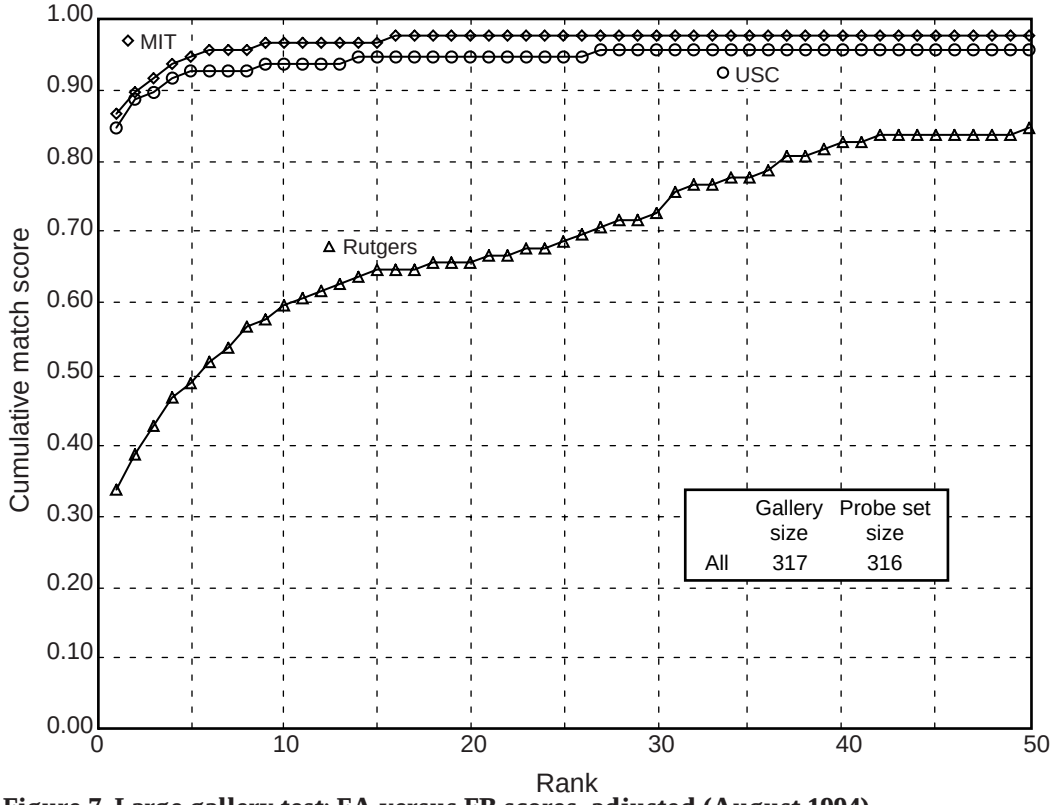


Figure 7. Large gallery test: FA versus FB scores, adjusted (August 1994).

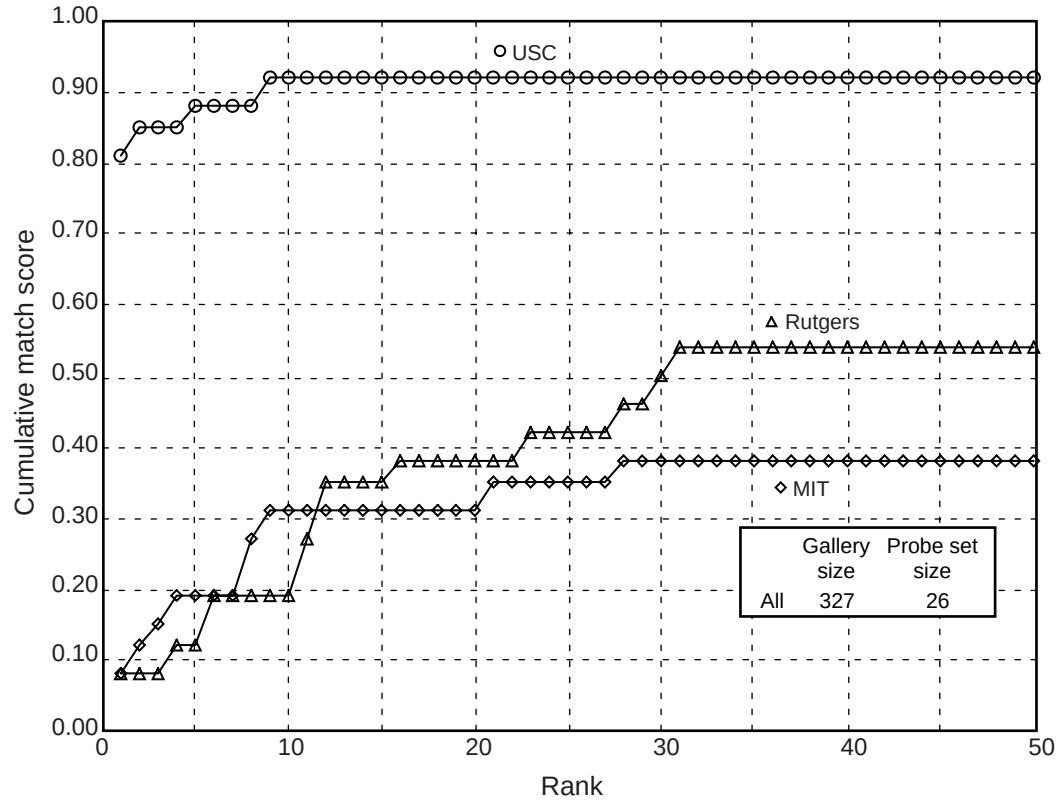


Figure 8. Large gallery test: quarter profile scores, adjusted (August 1994).

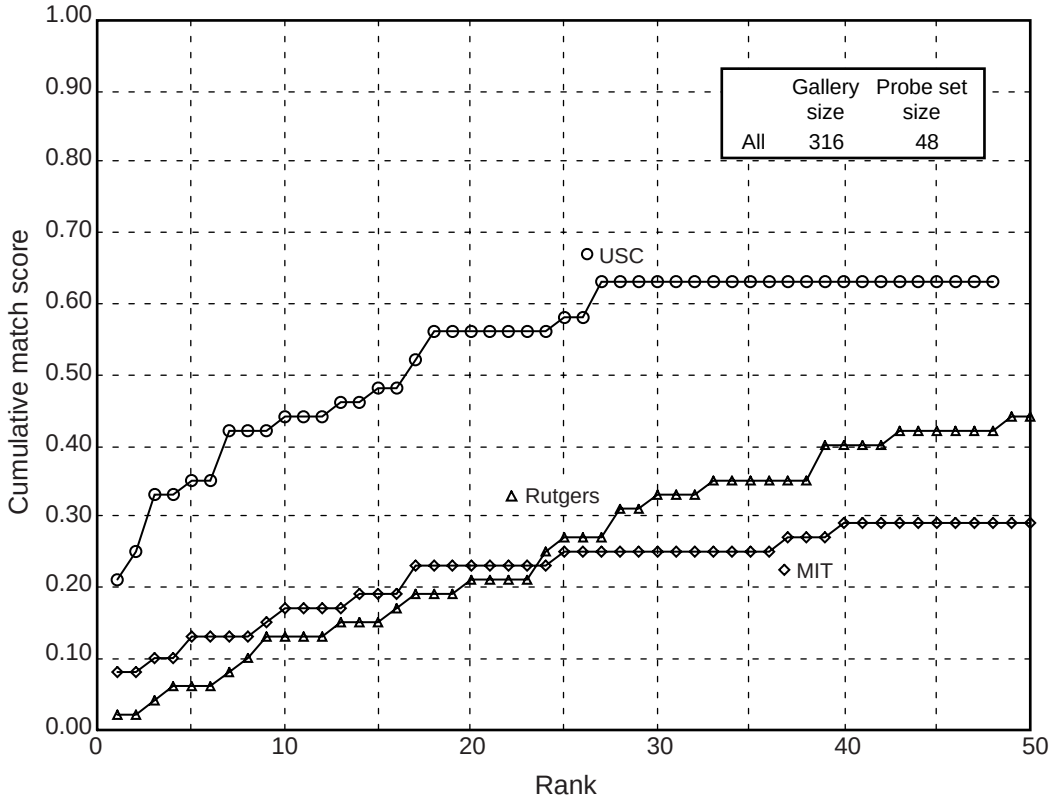


Figure 9. Large gallery test: half profile scores, adjusted (August 1994).

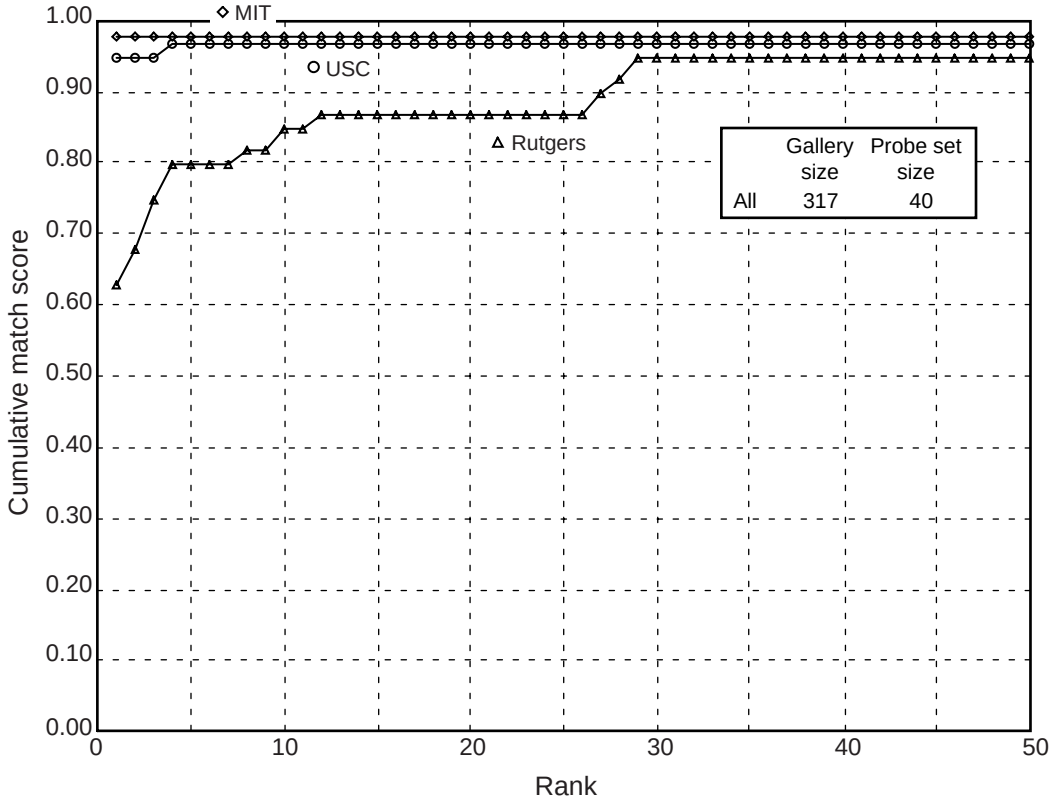


Figure 10. Large gallery test: 10% scale reduction scores, adjusted (August 1994).

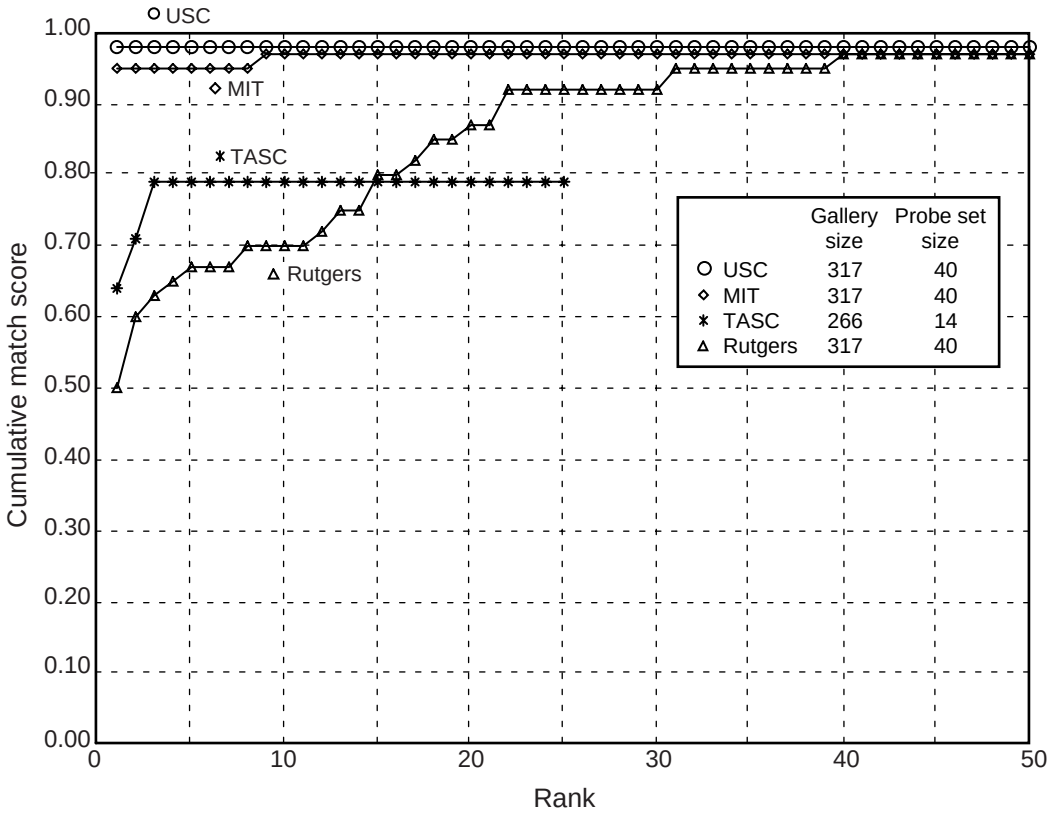


Figure 11. Large gallery test: 20% scale reduction scores, adjusted (August 1994).

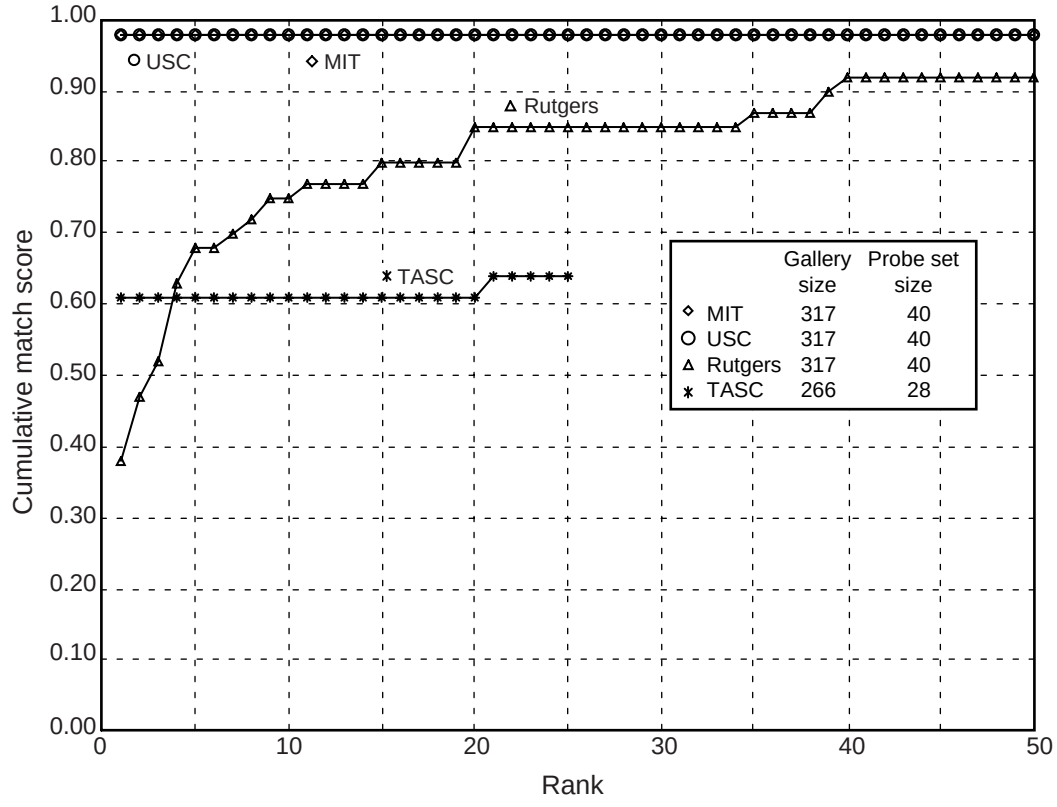


Figure 12. Large gallery test: 30% scale reduction scores, adjusted (August 1994).

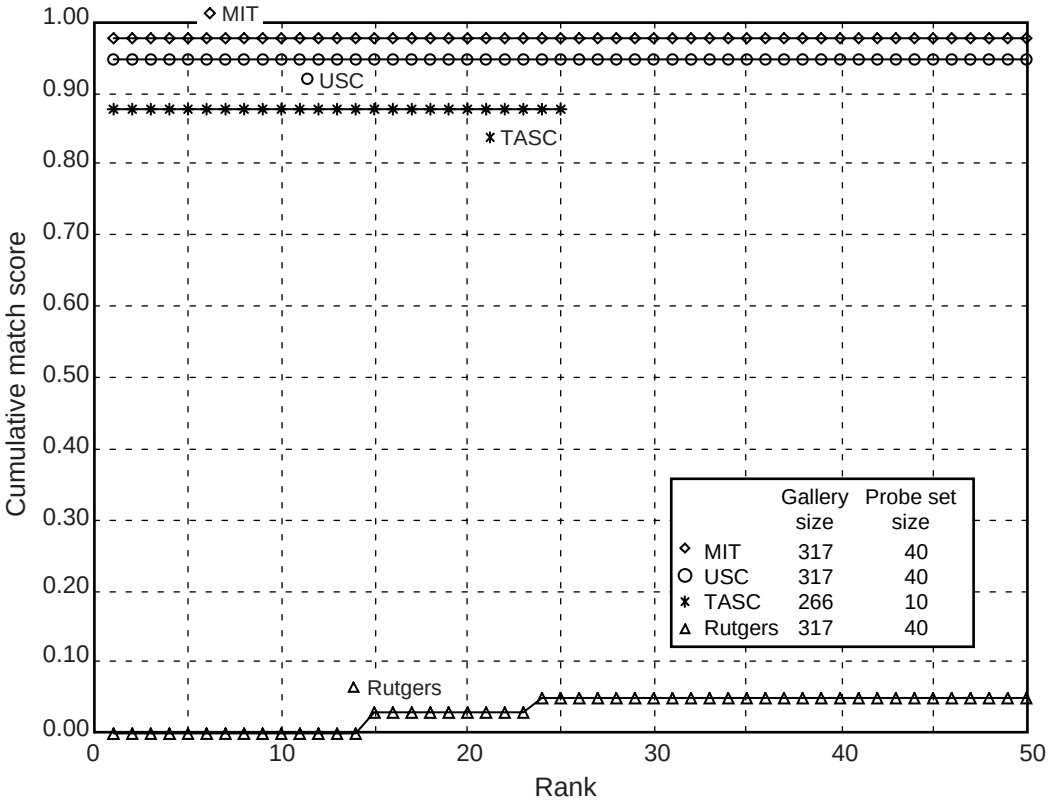


Figure 13. Large gallery test: 40% of illumination scores, adjusted (August 1994).

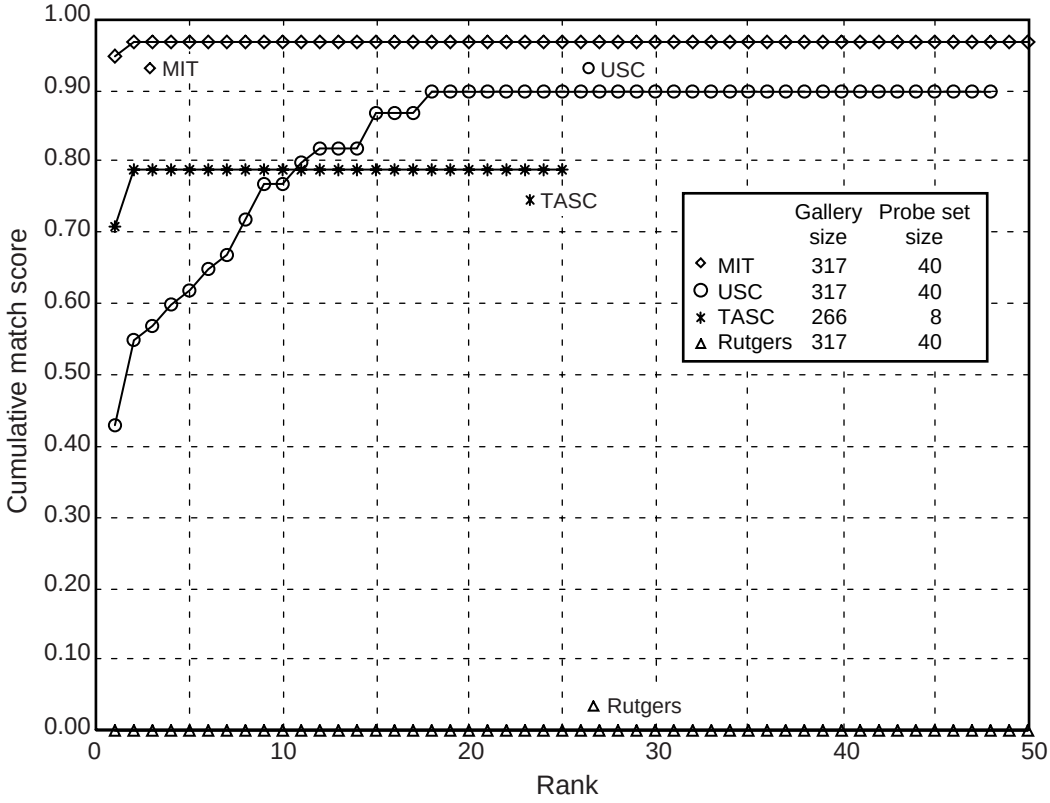
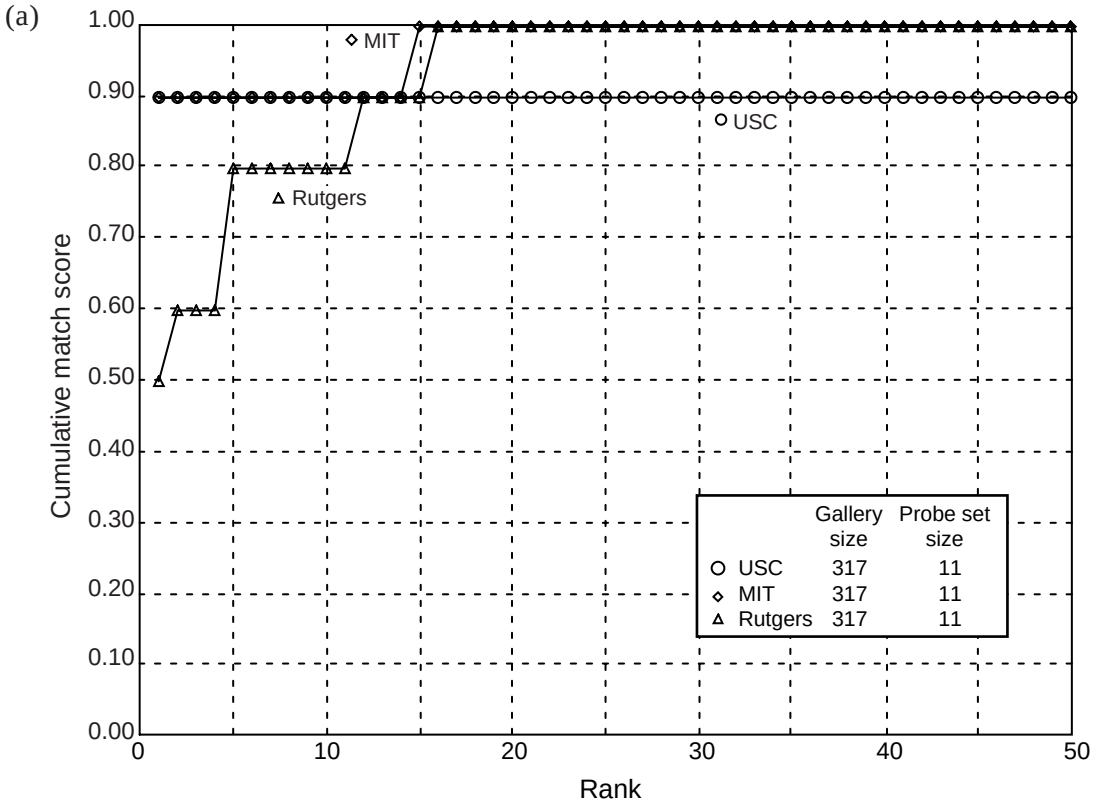


Figure 14. Large gallery test: 60% of illumination scores, adjusted (August 1994).



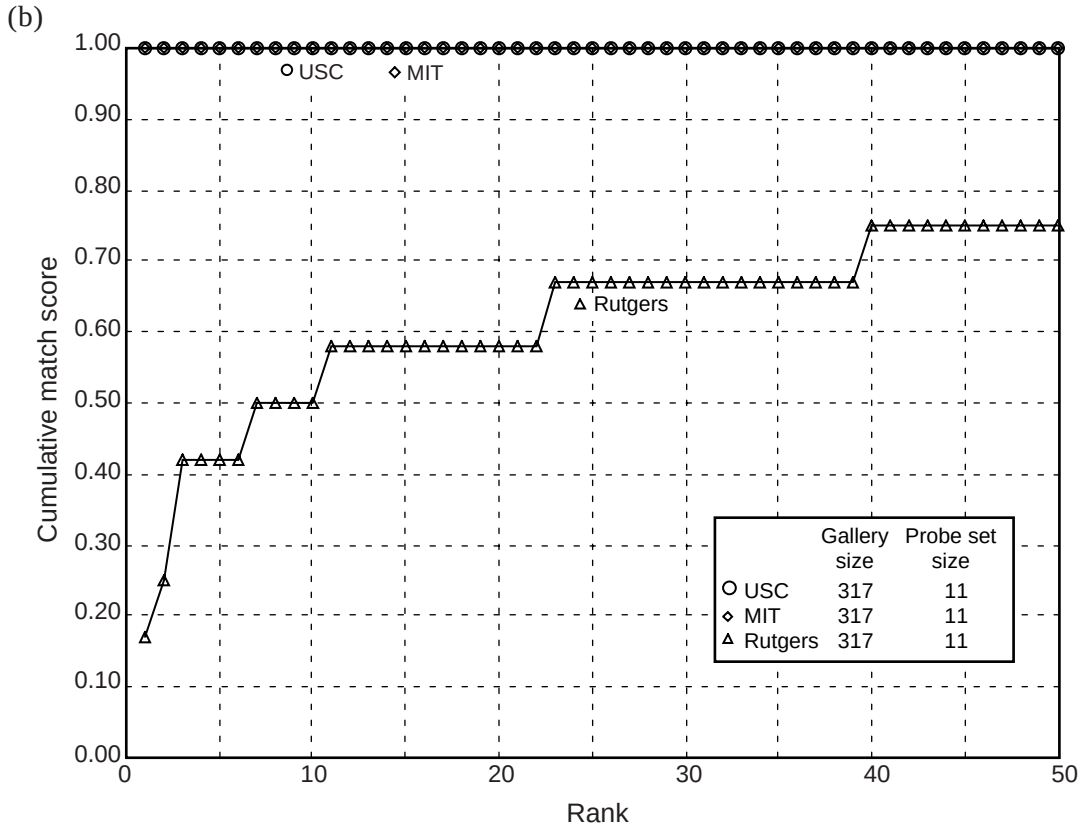


Figure 15. Large gallery test: (a) clothing color darkened scores, adjusted; (b) clothing color lightened scores, adjusted (August 1994).

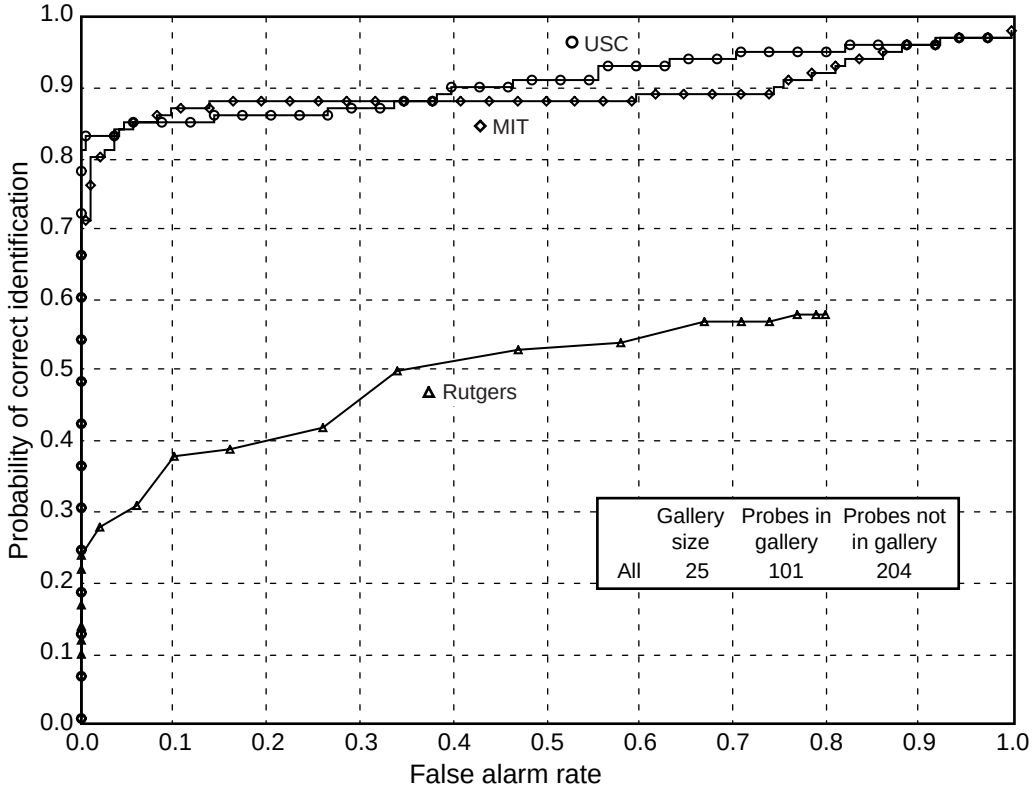


Figure 16. False-alarm test: ROC (August 1994).

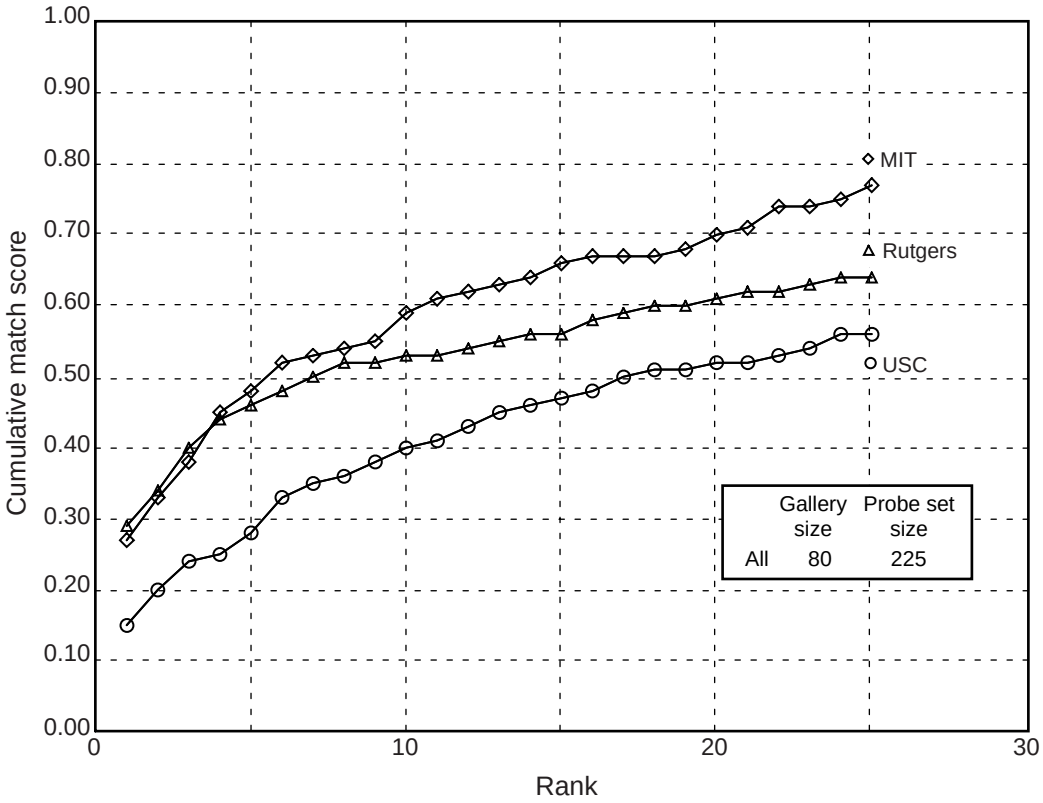


Figure 17. Rotation test: overall scores (August 1994).

4.7 Analysis

Performance of the algorithms falls roughly into three categories. In the first category are the algorithms of MIT and USC; both these algorithms perform comparably on the large gallery test and on the false-alarm test. The second category consists of the TASC algorithm, and the third category is the Rutgers algorithm. As a rule there is a noticeable difference in performance between each category. It is harder to draw definite conclusions about performance within the category, because there is no estimate of the variance of the recognition scores; e.g., we do not know how the performance score would change if we moved the FB images to the gallery and the FA images to the probe set.

The graphs show that the MIT, USC, and TASC approaches consistently outperform the Rutgers approach. The testing sets for TASC are different from the others, so the TASC results can be compared only roughly; an exact comparison was not possible from these test results, because of the need for different test sets.

Comparison of figures 4 and 8 shows that the Rutgers and MIT algorithms are very sensitive to changes in profile, particularly MIT. The USC algorithm maintains high performance for quarter-profile images, but performance drops considerably for half profiles (fig. 9). Most of the algorithms show little if any degradation under scale reduction up to 30 percent (fig. 10 to 12). Likewise, USC and TASC show greater sensitivity to illumination than the other algorithms (fig. 13 and 14). Examination of figure 5 shows that the mistakenly included gallery images are indeed harder to use, as both the MIT and USC algorithms show an 8 to 9 percent drop in performance when these images are included in the gallery.

The false-alarm test (fig. 16) shows the same breakout in performance groups as the large gallery test: MIT and USC are comparable across the entire ROC, and they outperform Rutgers.

The rotation test confirms the finding from the large gallery test that rotation is a hard problem and was beyond the scope of phase I of the FERET program. On the rotation test, MIT and Rutgers had comparable performance and outperformed USC. This is in contrast to the large gallery test, where USC outperformed MIT and Rutgers on the rotation categories.

The conclusion drawn from the phase I test was that the next step in the development of face recognition algorithms was to concentrate on larger galleries and on recognizing faces in duplicate images. The large gallery test established a baseline for algorithm performance. The algorithms tested demonstrated a level of maturity that allows them to automatically process a gallery of 316 images and a probe set of 770 images. The results on all categories of probes were well above chance, and the algorithms demonstrated various degrees of invariance to changes in illumination, scale, and clothing color.

The decision to concentrate on larger galleries and duplicates was driven by real-world considerations. All applications require algorithms to recognize people from images taken on different days, and many users require the algorithms to work on databases of over 10,000 individuals. The other hard problem identified by the test was recognizing faces when the probe and gallery image have different poses. It was decided to delay working on this problem to avoid spreading the research effort too thinly. Also, solving the duplicate problem is a prerequisite to the rotation probe. Real-world applications will use rotated images taken at different times.

5. Phase II

In Phase II, TASC, MIT, and USC continued development of their approaches. The MIT and USC teams continued work on developing face recognition algorithms from still images. The TASC effort switched to developing an algorithm for recognizing faces from video. The emphasis was to estimate the three-dimensional shape of the face from motion and recognize the face based on its shape. In phase II, Rutgers performed a study comparing and assessing the relative merits of long-wave infrared images and visible images for face recognition and detection. Their results are not reported here. Since the Rutgers and TASC efforts pursued different avenues, it was not appropriate for their algorithms to take the phase II test.

Phase I of the FERET program established a baseline for face recognition algorithms; the goal of phase II was to improve the performance of the algorithms to the point that they could be ported to a real-time experimental/demonstration system. An experimental/demonstration system would enable one to collect performance statistics over a longer time period than is possible with a laboratory test.

One of the conclusions from the phase I test was that greater improvement was needed in the ability of algorithms to recognize faces when the probe and gallery images were taken weeks, months, or years apart (duplicate images). Another major concern was how algorithm performance would scale as the size of the gallery increased. In phase II, both the MIT and USC teams concentrated on these two issues. As a measure of progress, both MIT and USC took the March 1995 phase II FERET test. The data collection activities in phase II were designed to support the March 1995 test.

The March 1995 test consisted of one test that was an enlarged version of the large gallery test of August 1994. The main difference is that the gallery consisted of 831 individuals, and there were 463 duplicate images in the probe set. The designation of the *fa* or *fb* frontal image as FA was determined randomly. Only 780 out of the 831 FB images were placed in the probe set. The breakout of the images in the test is given in table 8.

The testing procedure for March 1995 was the same as for the August 1994 test. The test was administered at MIT on 1 to 2 March 1995 and at USC on 6 to 8 March 1995. The time limit for taking the test was three days.

In phase II, the MIT team developed two versions of their face recognition algorithm. In the “original” version, the feature locator module passed the top location for each feature to the identification module, and in the “hierarchical” version, the top three locations were passed to the identification module. Both versions of the algorithm were tested.

5.1 Results

The contractors were requested to supply the test results in the same format as the earlier Phase I test, as shown in table 6, except that the ranked list was to include the top 100 matches instead of the top 50.

The scoring protocol for this test is the same as the large gallery test from phase I, and the results are scored and reported in the same manner. Table 9 shows the categories of images corresponding to the figures presenting the results (fig. 18 to 28).

Table 8. Number and types of images used in March 1995 test.

Image category	Number
Gallery images:	
FA frontal images	831
Probe images:	
FA Frontal images (*fa*)	71
FB frontal images	780
Probes not in gallery (frontal images)	45
Duplicate frontal images	463
Quarter rotations	33
Half rotations	48
40% change in illumination	40
60% change in illumination	40
10% reduction in scale	40
20% reduction in scale	40
30% reduction in scale	40
Contrast-reversed clothes	40
Total probes	1680

Table 9. Figures reporting results for March 1995 test.

Figure no.	Category title	Description of category
18	Overall match	Given any probe aspect, find correct ID.
19	FA versus FB match	Match FB frontal images from same set.
20	Duplicate match	Match frontals collected on different dates.
21	Quarter match	Given quarter profile, find frontal match.
22	Half match	Given half profile, find frontal match.
23	60% illumination match	Given an image with brightness reduced to 60%, find frontal match.
24	40% illumination match	Given an image with brightness reduced to 40%, find frontal match.
25	10% scale match	Given an image reduced by 10%, find frontal match.
26	20% scale match	Given an image reduced by 20%, find frontal match.
27	30% scale match	Given an image reduced by 30%, find frontal match.
28	Clothes change	Given an image with clothes contrast changed, find match.

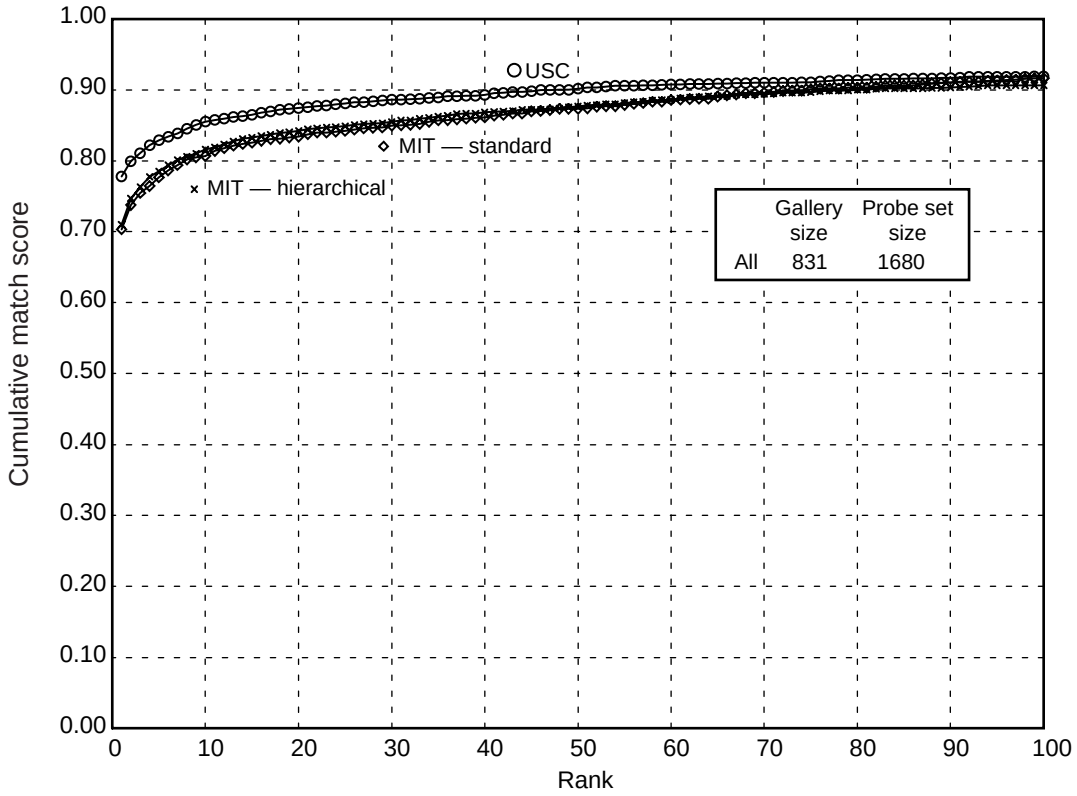


Figure 18. Large gallery test: overall scores (March 1995).

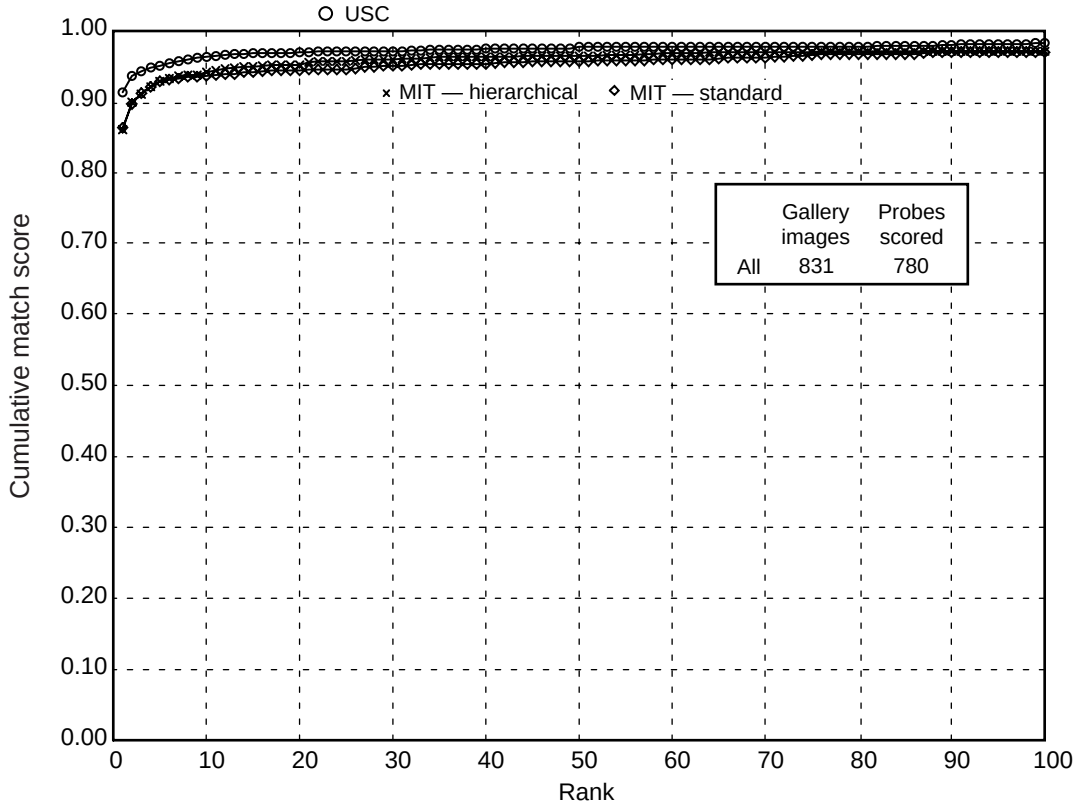


Figure 19. Large gallery test: FA versus FB (March 1995).

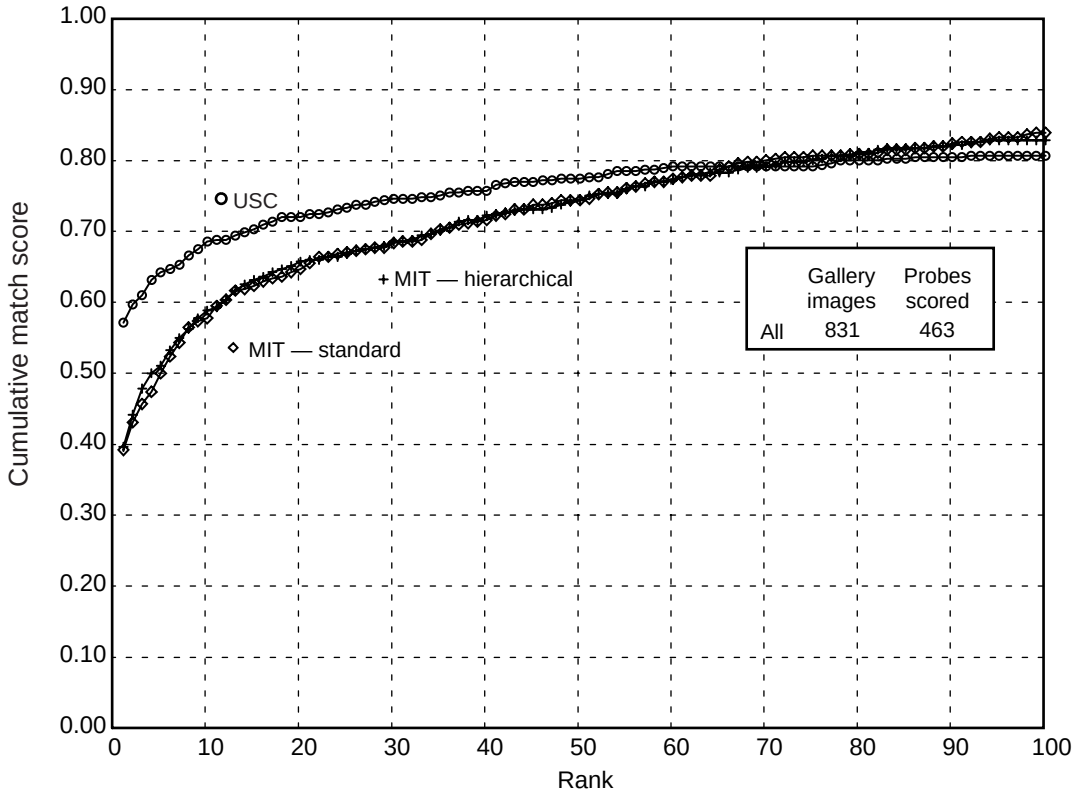


Figure 20. Large gallery test: duplicate scores (March 1995).

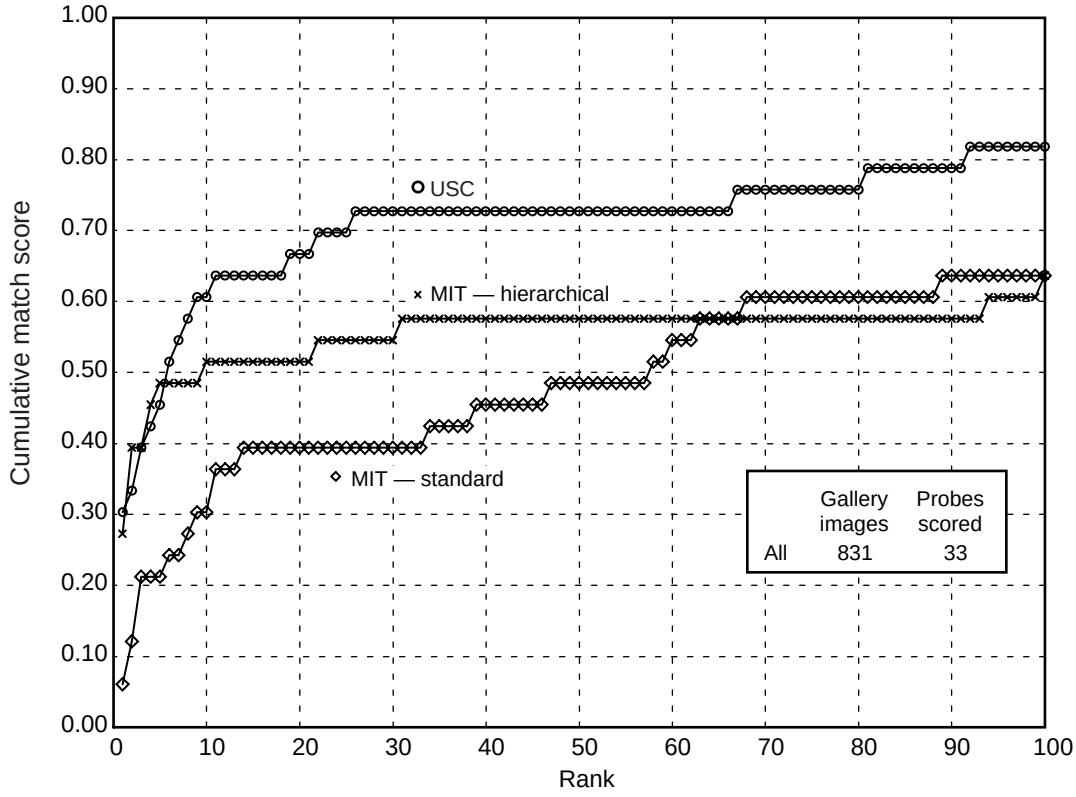


Figure 21. Large gallery test: quarter rotation (March 1995).

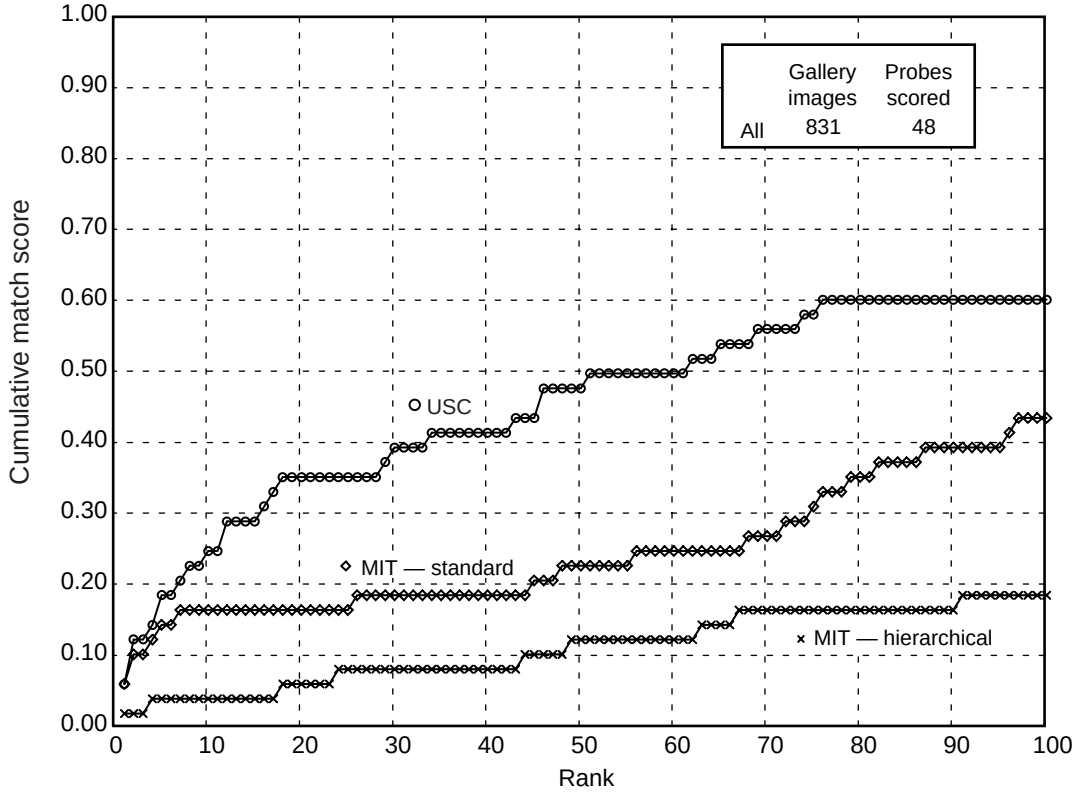


Figure 22. Large gallery test: half rotation (March 1995).

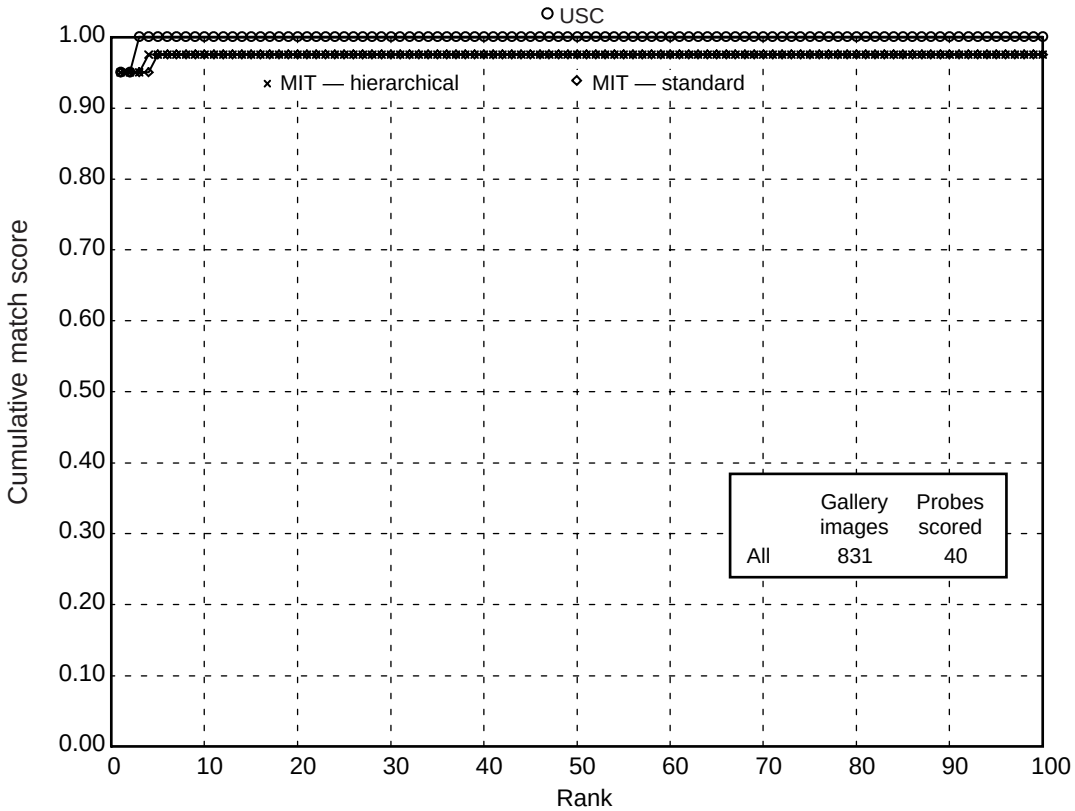


Figure 23. Large gallery test: 60% original illumination (March 1995).

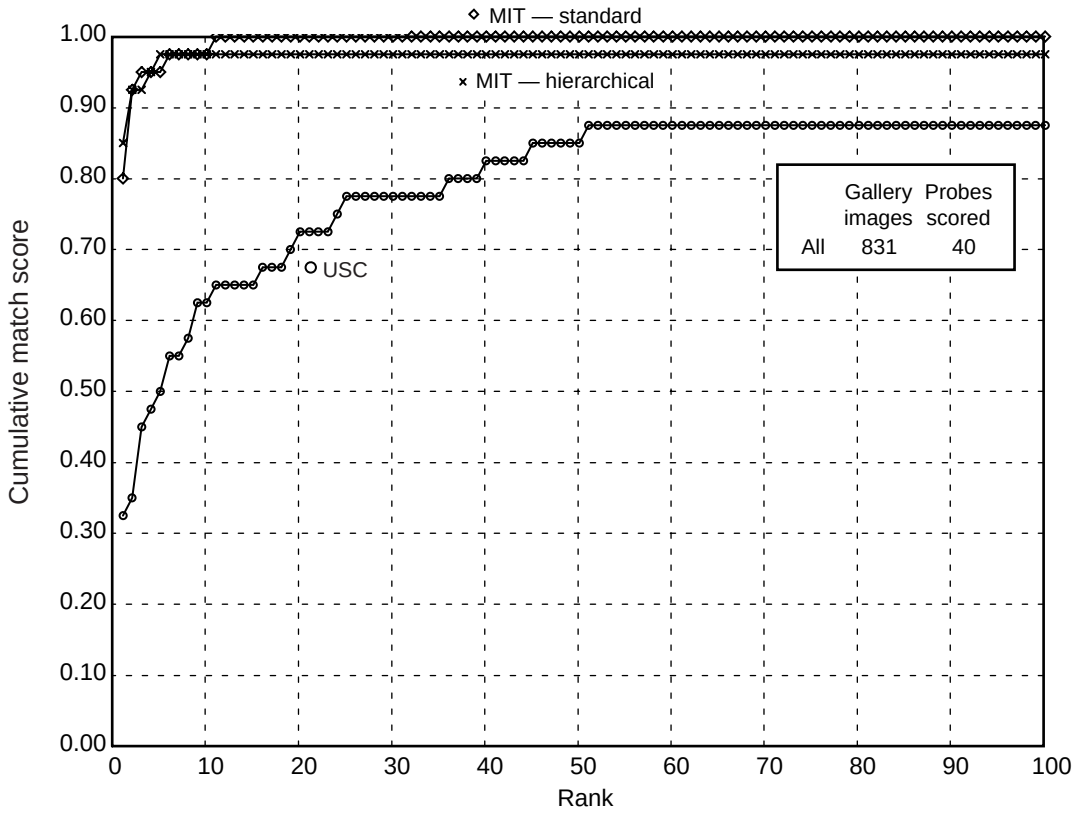


Figure 24. Large gallery test: 40% original illumination (March 1995).

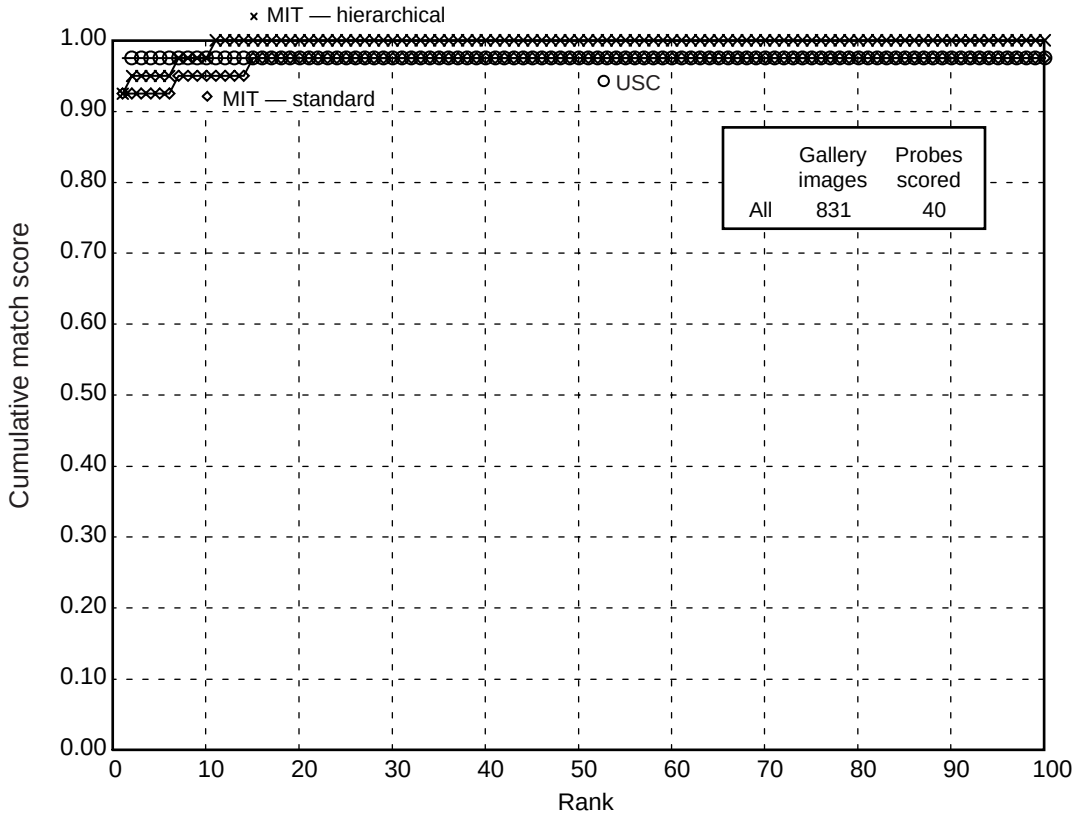


Figure 25. Large gallery test: 10% reduced image size (March 1995).

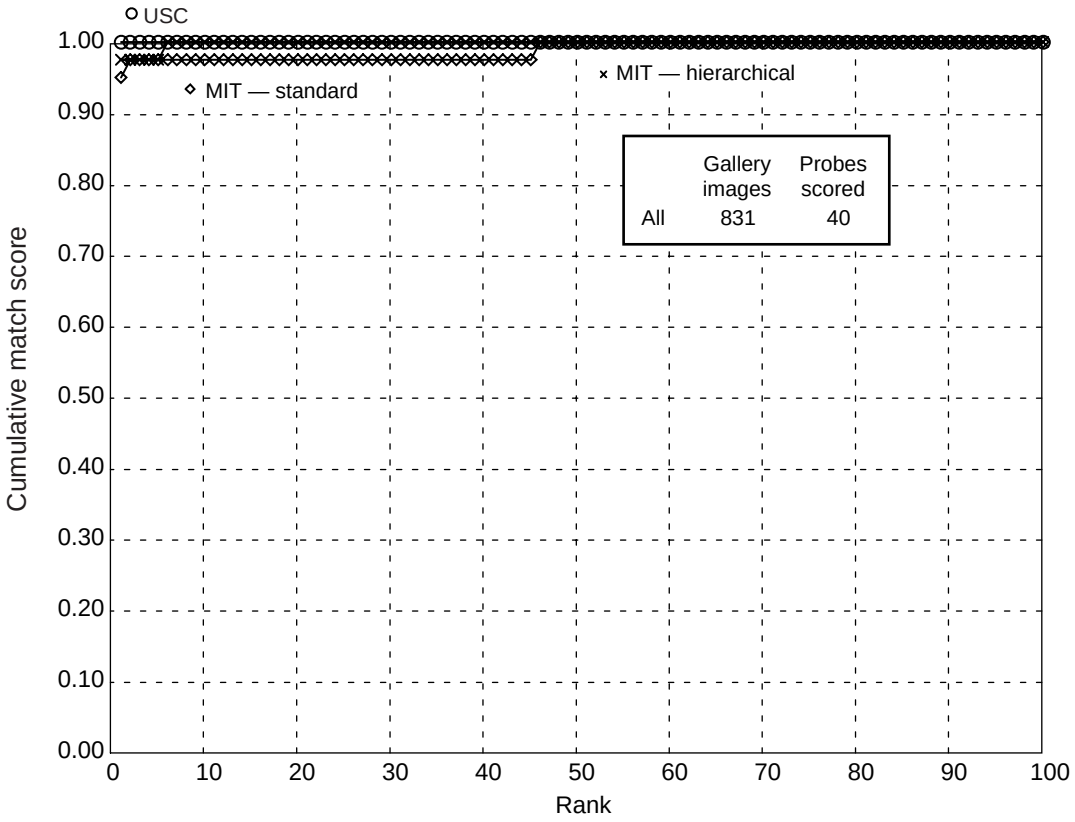


Figure 26. Large gallery test: 20% reduced image size (March 1995).

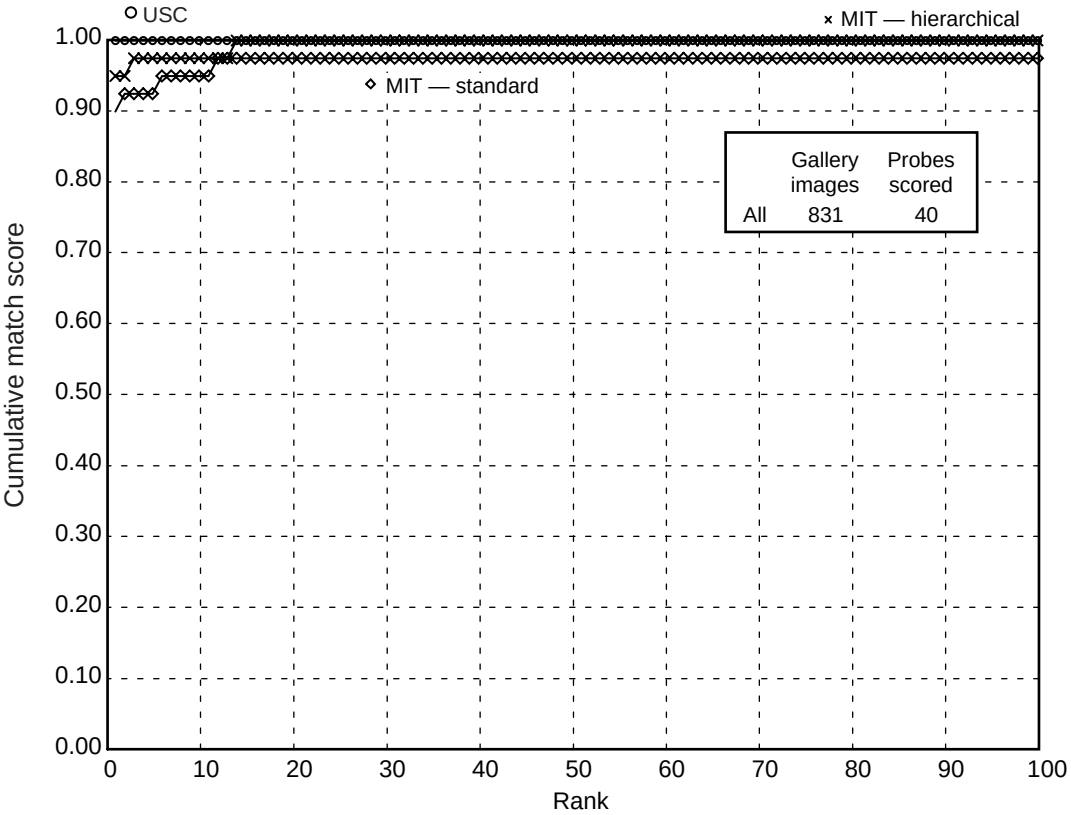
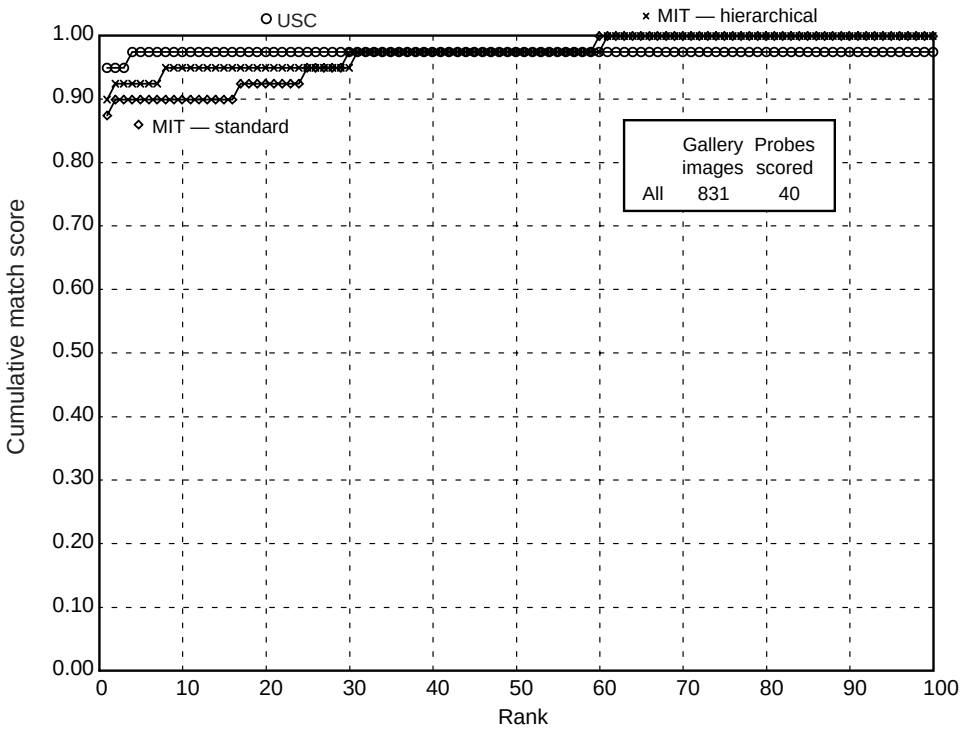


Figure 27. Large gallery test: 30% reduced image size (March 1995).

Figure 28. Large gallery test: clothes contrast change (March 1995).



5.2 Analysis

Analysis of figure 18 shows that the USC and the two MIT algorithms performed well on the test set, with the USC algorithm showing slightly better results. Figure 19 shows that for frontal images taken on the same date, the algorithms give virtually identical results. All the algorithms show a marked decrease in performance when the test images were taken on different dates from those of the gallery images (fig. 20), with the MIT algorithms showing a greater decrease in performance. Figures 21 and 22 show that all the algorithms are still sensitive to the angle of the face to be recognized, especially the MIT algorithms. The MIT algorithms show almost no decrease in performance due to reduced illumination (fig. 23 to 24). The USC algorithm exhibits degraded performance after illumination is reduced to 40 percent of original. All the algorithms demonstrate insensitivity to reduced image size up to 30 percent (fig. 25 to 27). The algorithms were not “tested to failure” by continual reductions in image size, because the research groups were told that variations in scale would not exceed a factor of two. The algorithms also do not degrade significantly when the clothes contrast changes (fig. 28), suggesting that the algorithms have been successful in using the face features for recognition.

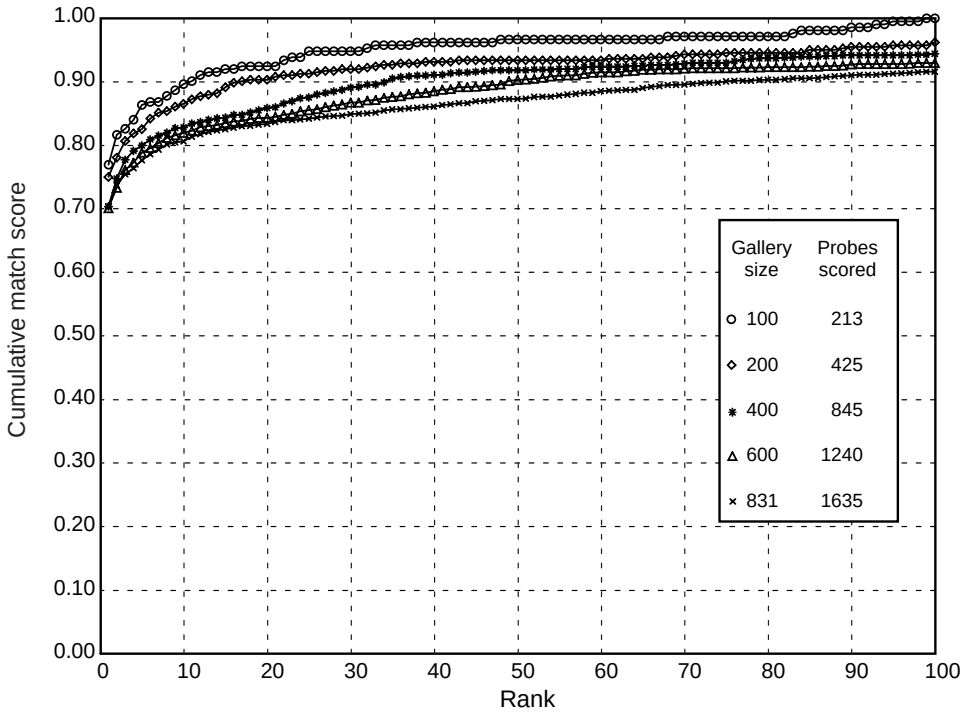
The MIT modification for hierarchical searching for features has little impact on the recognition of probe images if the image is frontal face, as can be seen in figures 18 to 20 and 23 to 26. It did improve the performance slightly on images with the largest scale change (fig. 27). The most notable difference in performance between the hierarchical approach and the standard approach can be seen in the rotated images (fig. 21 and 22). The hierarchical approach shows a significant improvement in performance on the

quarter-profile images and a modest decrease in performance on the half-profile images. This indicates that the hierarchical approach does improve robustness on images where the face is not full frontal but most of the face is presented. The loss of performance on the half-profile images may be due to difficulties in locating the eye farthest from the camera: notice in figure 3 the differences between the *ql* and *hl* and between the *qr* and *hr* images. Only in the quarter images can both eyes be fully seen.

As a means of assessing the effect of gallery size on performance, the MIT standard and algorithm was tested on a series of galleries of increasing size: the graduated gallery study. Gallery sizes of 100, 200, 400, 600, and 831 were used by the MIT team to test the capacity versus performance of their system. Figures 29 to 34 show the size of the gallery and number of probes scored. These galleries were a subset of the original 831-person gallery, and for each run of this experiment, the original probe set of 1680 was used. In computing the scores, the appropriate subset of probes was used: i.e., in the gallery of 100 people, the FA versus FB results involved only FB images in the probe set that were in this gallery.

Figures 29 through 34 show the MIT algorithm's performance for overall, duplicate, and FB images with galleries of increasing size. These figures show the expected decline in performance as the gallery becomes larger. Figures 31 and 34 show that for duplicates (frontal images taken on a different date from that of the gallery image), going from a gallery of 100 individuals to one of 831 individuals causes more than a 10-percent reduction in performance.

Figure 29. Graduated gallery study: overall scores (March 1995).



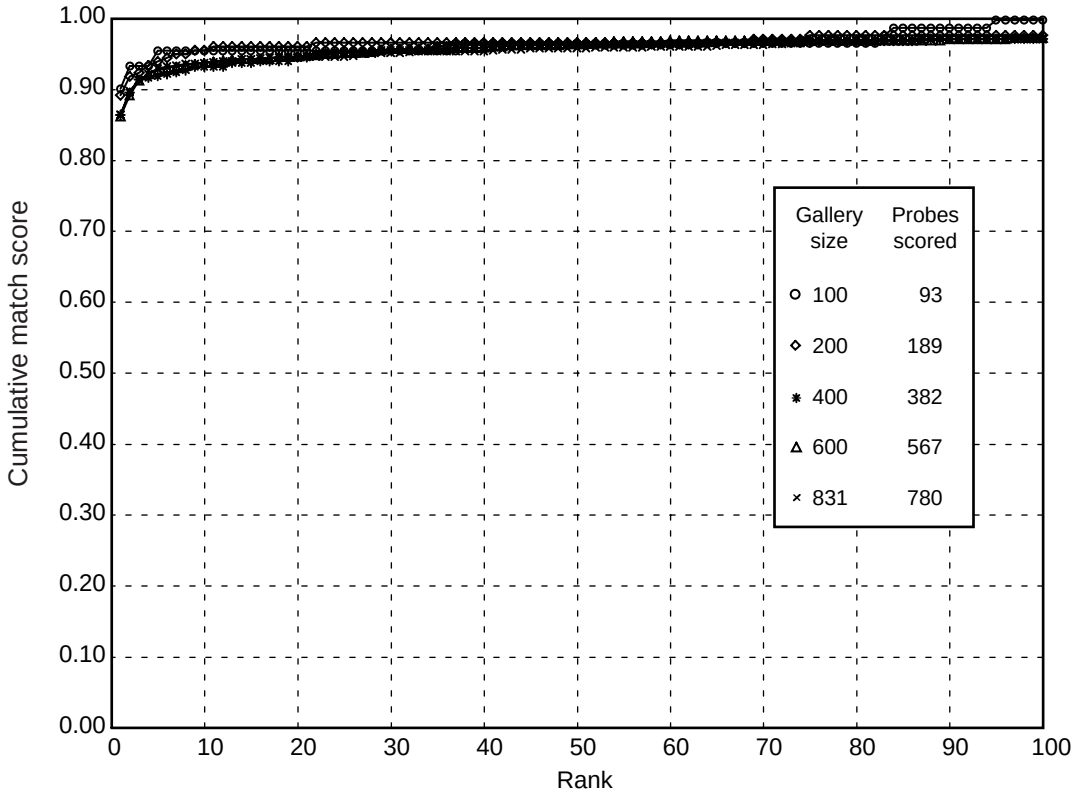


Figure 30. Graduated gallery study: FA versus FB scores (March 1995).

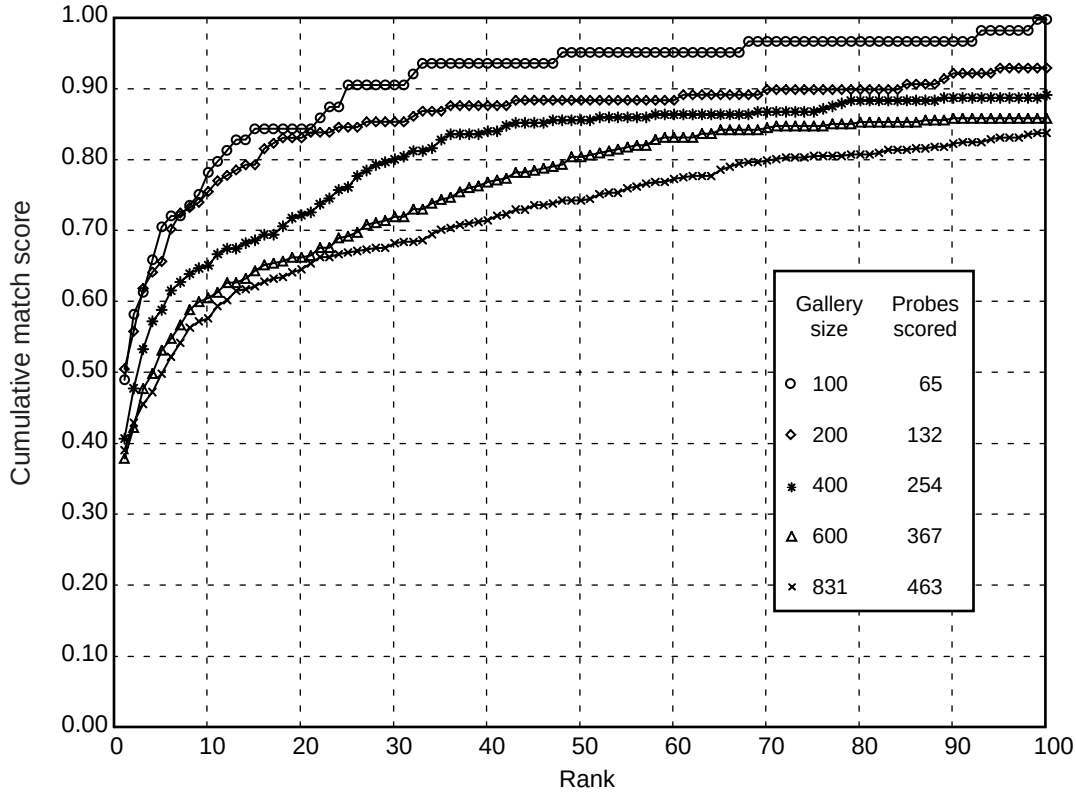


Figure 31. Graduated gallery study: duplicate scores (March 1995).

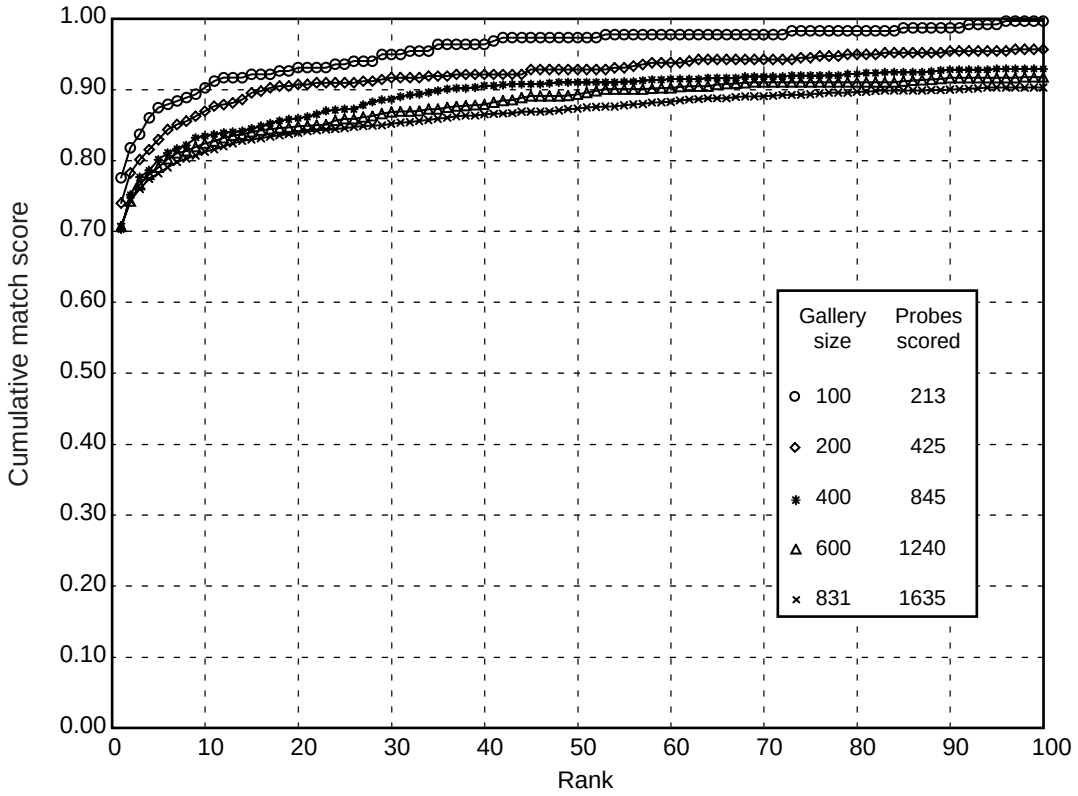


Figure 32. Graduated gallery study: overall scores (March 1995).

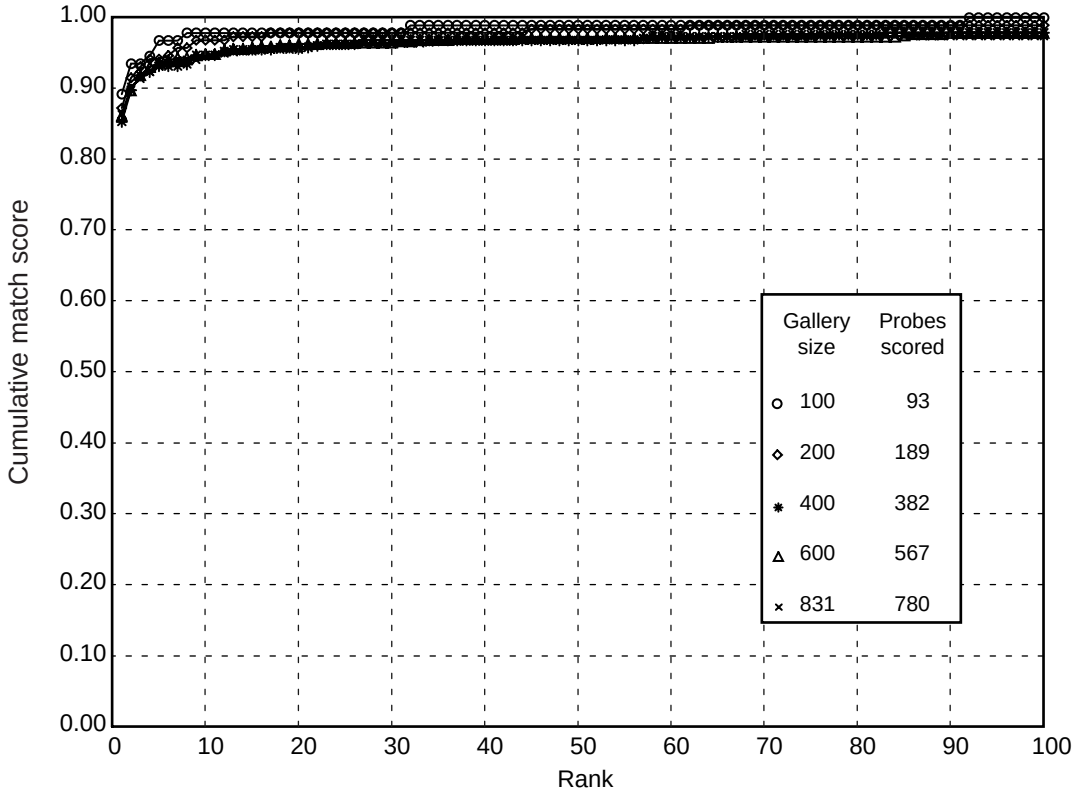


Figure 33. Graduated gallery study: FA versus FB scores (March 1995).

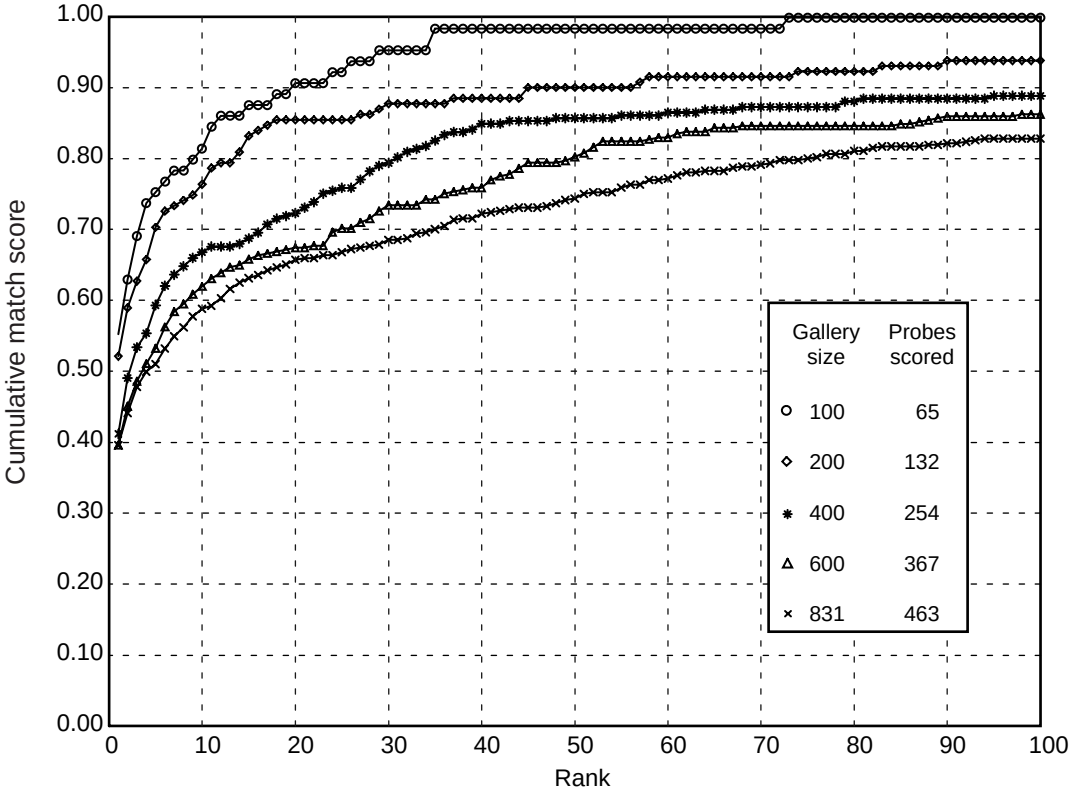


Figure 34. Graduated gallery study: duplicate scores (March 1995).

6. Comparison of August 1994 and March 1995 Test Performance

The principal objective for the August 1994 test was to evaluate each algorithm against a common baseline so that we could quantitatively measure each algorithm's performance and compare it to other algorithms on a common test set. In addition, during Phase I, we evaluated each algorithm to determine its potential for solving or at least contributing to solving the more complex face recognition problems of the future. Finally, the overall results of Phase I were considered in the selection of three research groups to continue algorithm development (out of the original five).

In contrast, the principal objectives of the March 1995 evaluation were to assess the maturity of the two algorithms tested and to determine if either or both were mature enough to be used in a demonstration system. This included testing against a more demanding and difficult test, including a larger database and more duplicate images. In addition, the March 1995 test was used to measure the performance improvements of recent modifications to both algorithms. Although the performance numbers decrease, the actual performance of both algorithms was judged to have improved, because they were successful despite increases in the number of images, in the number of duplicates, and in the difficulty of the test. Because of these factors, any comparison of the August 1994 and March 1995 results is very difficult.

However, one test in particular can be compared. The FA versus FB test, which identifies the alternative frontal images from the same collection date, is not affected by the presence of duplicate images. It is, therefore, reasonable to compare these test results. The March 1995 testing provides greater insight into the effects of an increased database as reflected by the increased gallery size. Figure 35 shows that the absolute performance increased as the gallery size increased for the USC algorithm, but no significant change was observed for the MIT standard algorithm.

One of the primary investigations of the March 1995 test studied the effect of duplicate images on performance. This test was of key importance to the FERET program and is also one of the most difficult problems to be addressed by any face recognition algorithm. The March 1995 test provided a 10× increase in duplicates and a 2.5× increase in gallery size over the August 1994 test.

Comparing the effects of duplicate images on the August 1994 and March 1995 test results, we determined that the correct recognition of individuals had declined, in the absolute sense. However, the March 1995 test provided a more stringent evaluation of each algorithm's performance by providing a more robust and diverse database. Therefore, we view the decline in performance as minimal, given the nature of the problem and the significant increase in the number of duplicate images used in testing. This result, combined with comparable FA versus FB scores against a larger gallery, leads us to conclude that the MIT and USC algorithms performed better in the March 1995 test.

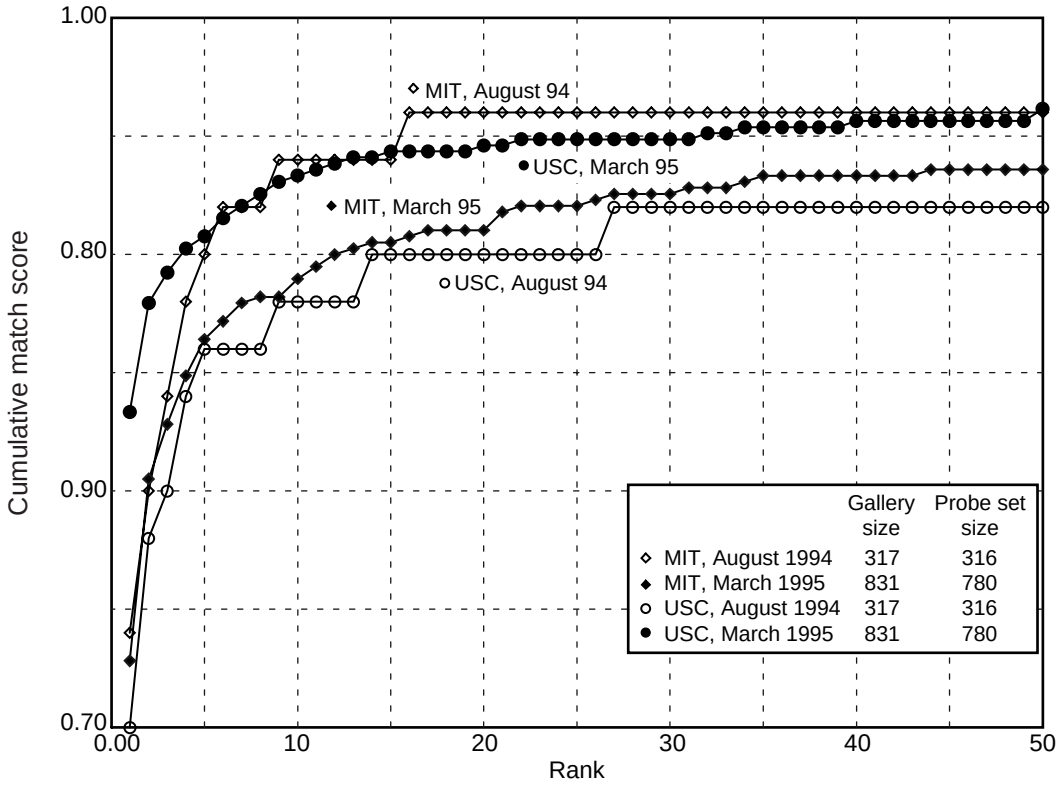


Figure 35. Large gallery tests: comparison of FA versus FB scores from phase I and phase II.

7. Tests on Algorithms Outside FERET Program

At the time of this report, only one other organization had submitted an algorithm for government testing. Joseph Atick, head of the Laboratory of Computational Neuroscience at Rockefeller University, New York, requested a government test of the Rockefeller algorithm. This algorithm was tested with the large gallery test of March 1995 and the false-alarm test of August 1994 at the Rockefeller site on 6 to 8 November 1995, under the same constraints as the previous tests. This report contains no information on the algorithmic approach, as these details were not revealed to us.

The Rockefeller algorithm performs quite well. Figures 36 to 39 show the Rockefeller results plotted with the MIT and USC results from the Phase II test. The algorithm performs significantly better than any tested algorithm on the quarter-rotated images (fig. 39). Figures 40 to 45 show the Rockefeller algorithm performance under the remaining test conditions. It performs comparably to the USC and MIT algorithms under these conditions.

In addition, the Rockefeller algorithm took the false-alarm test from Phase I. Figure 46 shows the results for Rockefeller along with the MIT and USC results. Note that the USC and MIT results are from August 1994, as a false-alarm test was not included in the March 1995 test.

It is anticipated that other algorithms will be submitted for testing in the future. Results from these tests will be published under separate covers as the need arises.

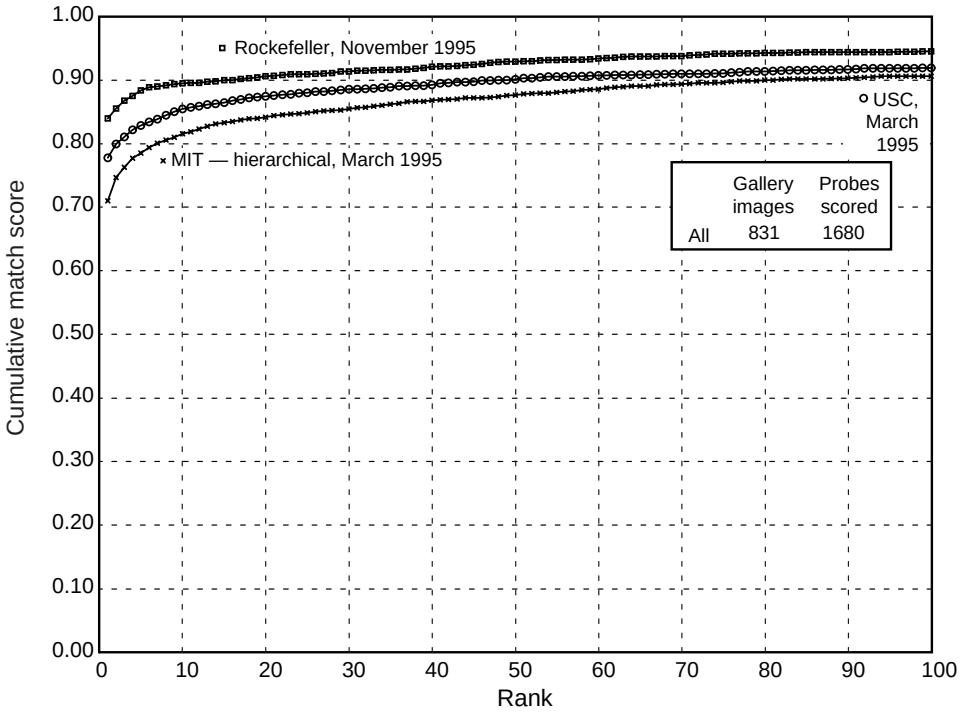


Figure 36. Large gallery tests: overall scores (November 1995).

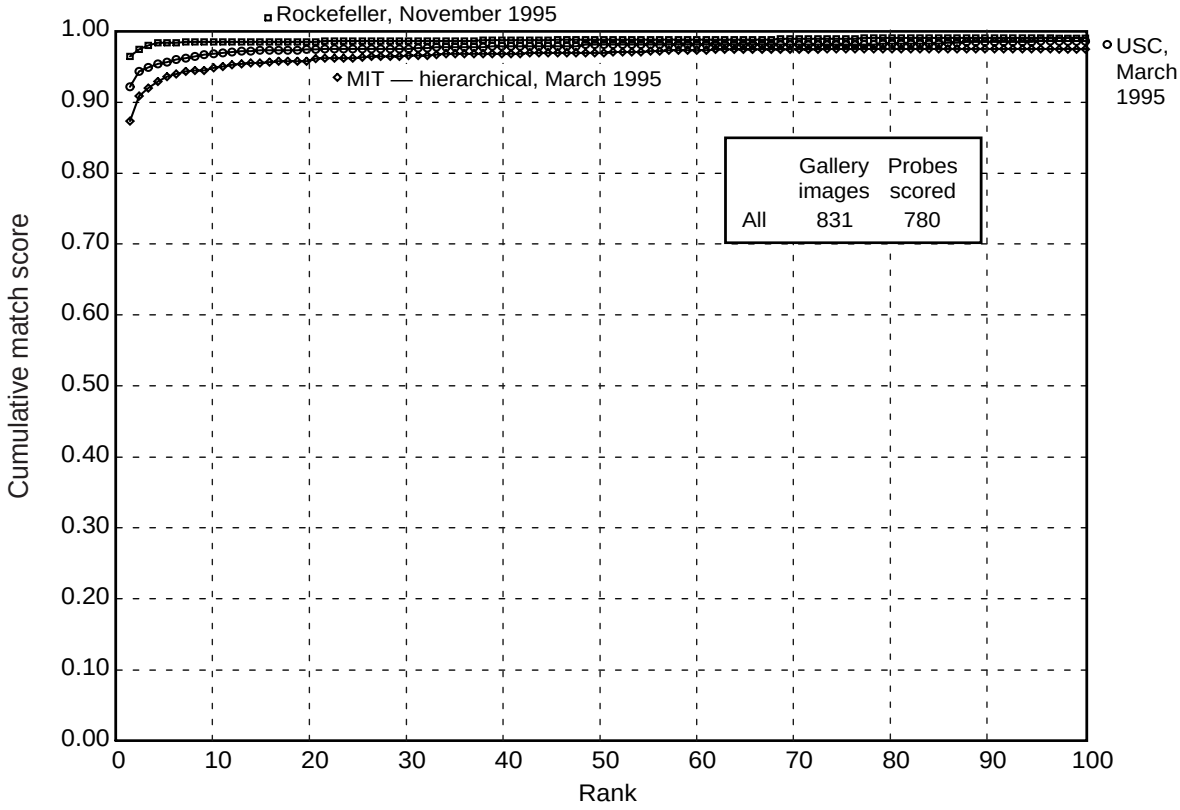


Figure 37. Large gallery tests: FA versus FB scores (November 1995).

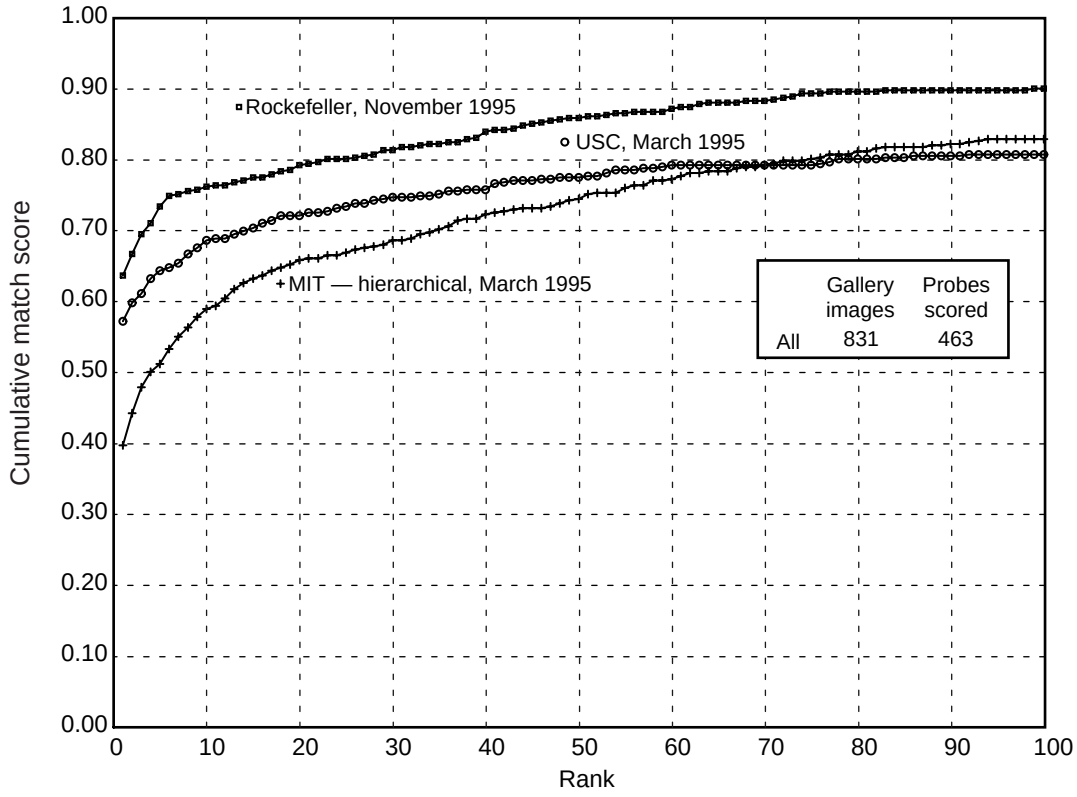


Figure 38. Large gallery tests: duplicate scores (November 1995).

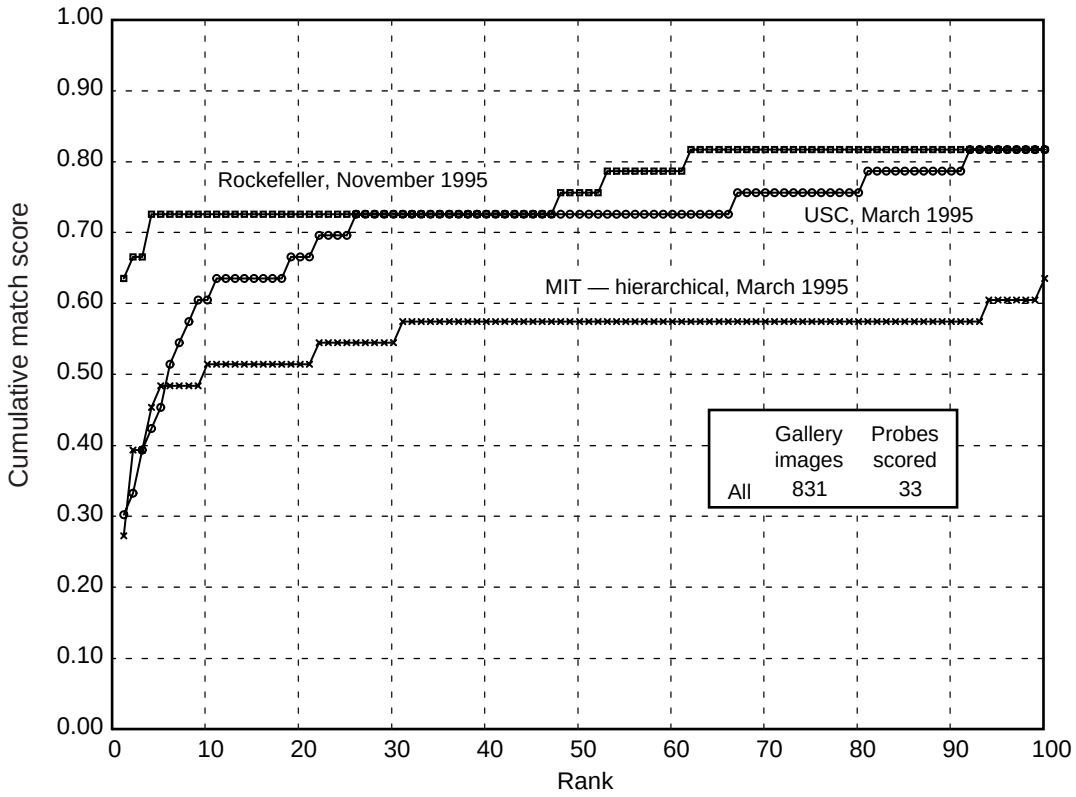


Figure 39. Large gallery tests: quarter rotation scores (November 1995).

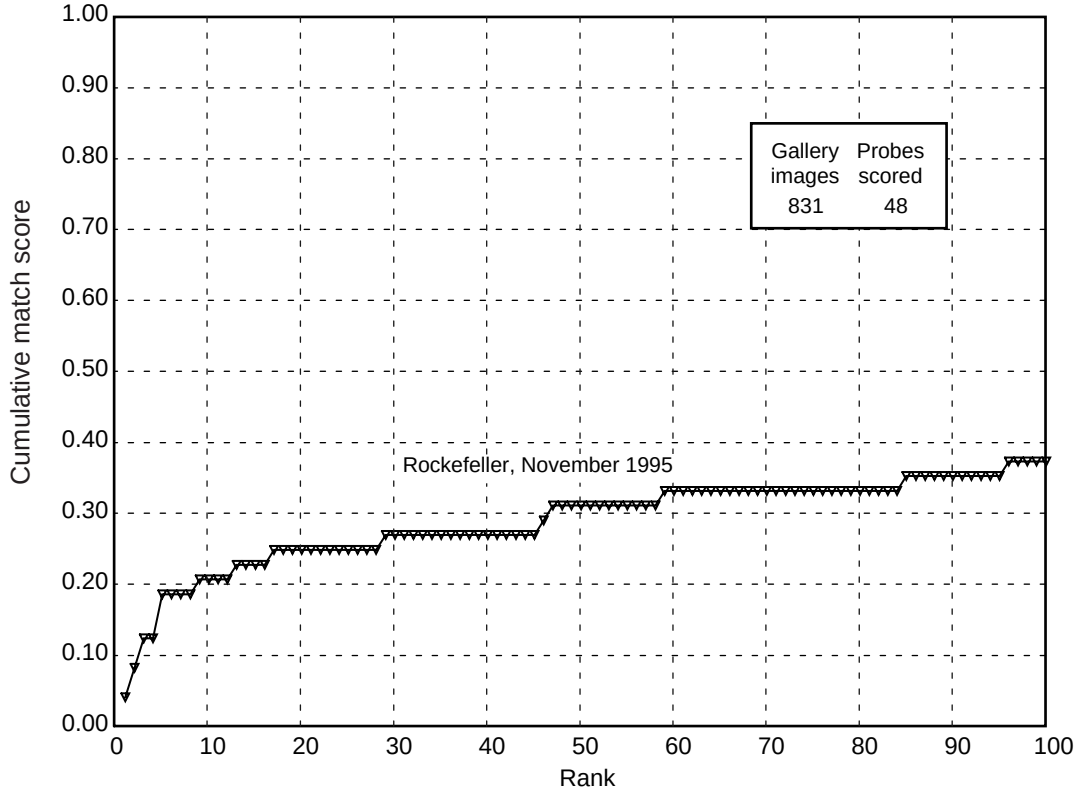


Figure 40. Large gallery tests: half rotation scores (November 1995).

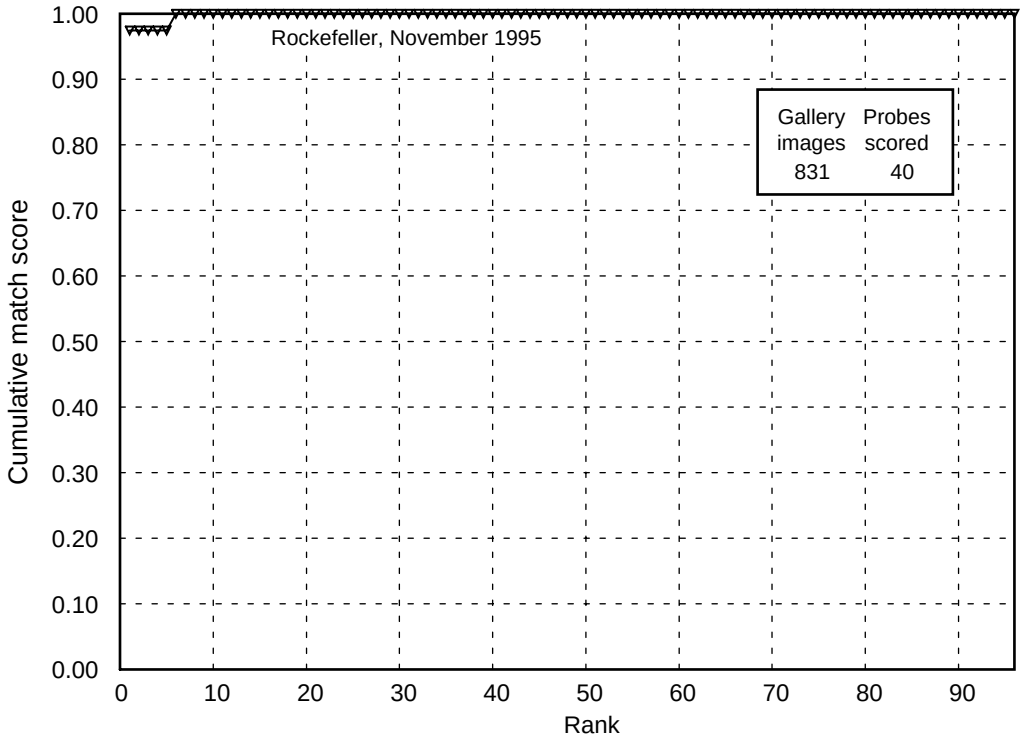


Figure 41. Large gallery test: 60% illumination reduction scores (November 1995).

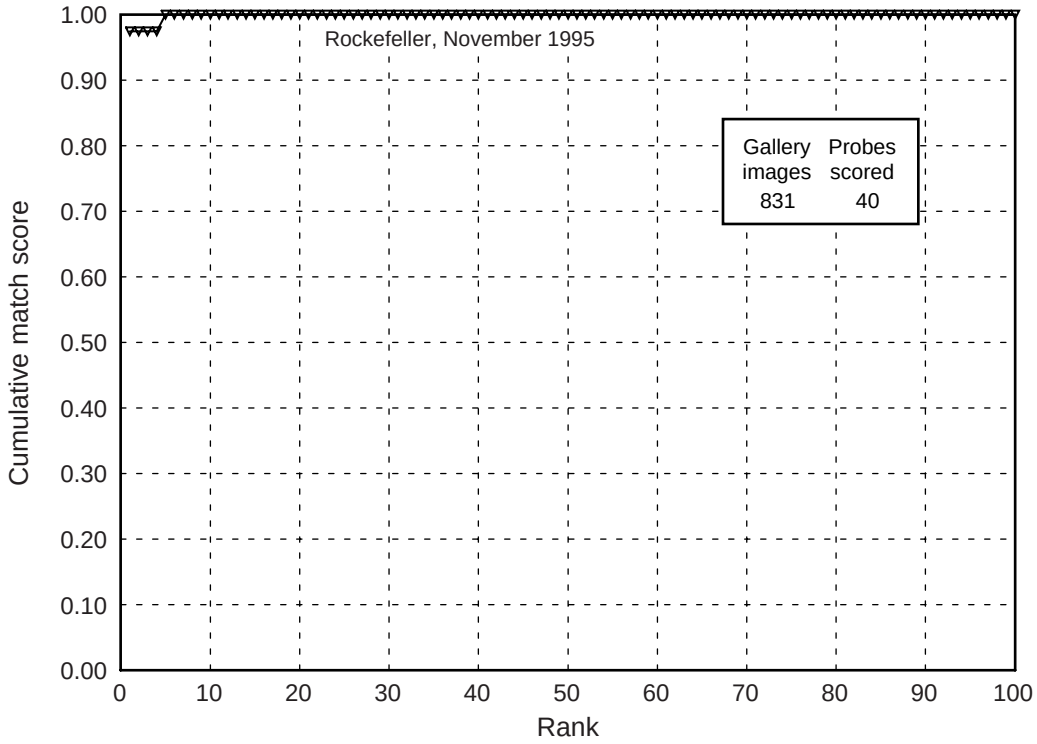


Figure 42. Large gallery test: 40% illumination reduction scores (November 1995).

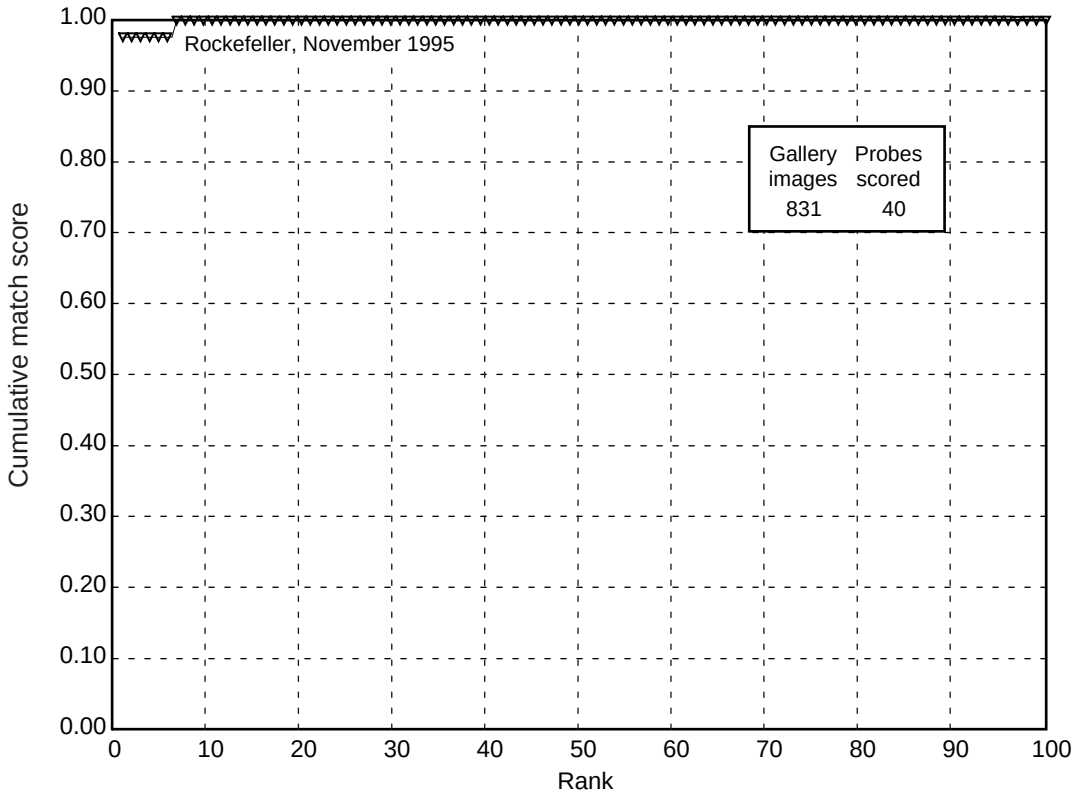


Figure 43. Large gallery test: 10% reduced image size scores (November 1995).

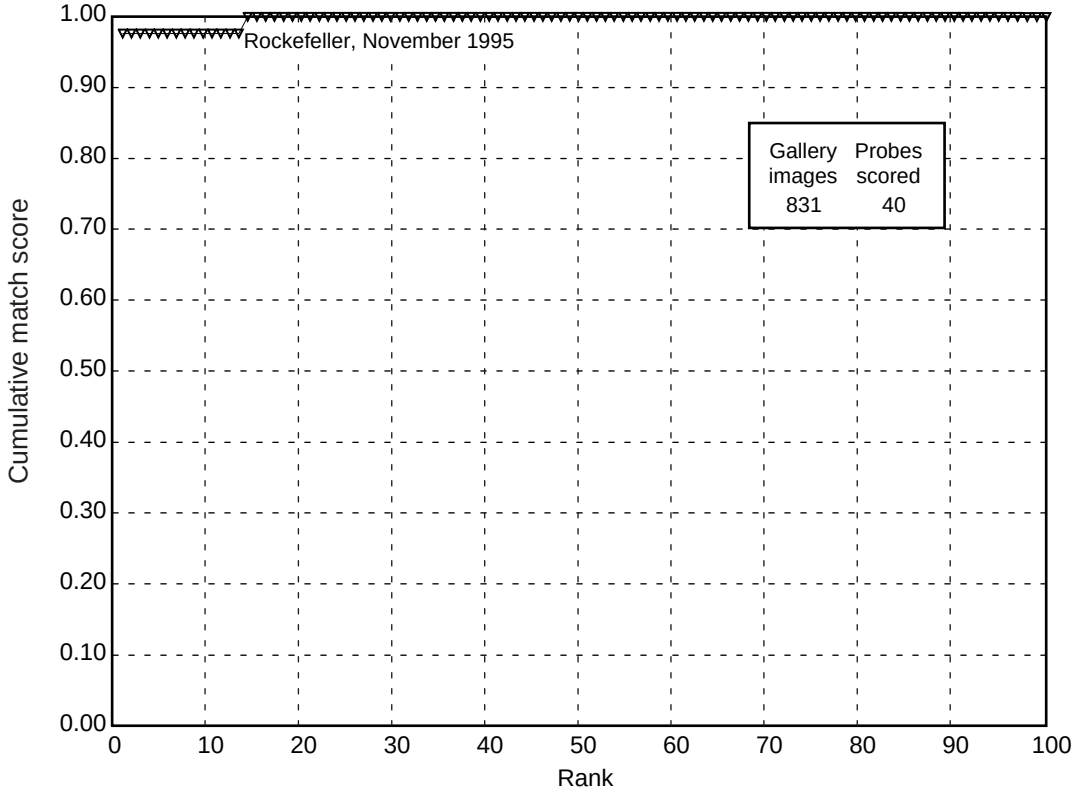


Figure 44. Large gallery test: 20% reduced image size scores (November 1995).

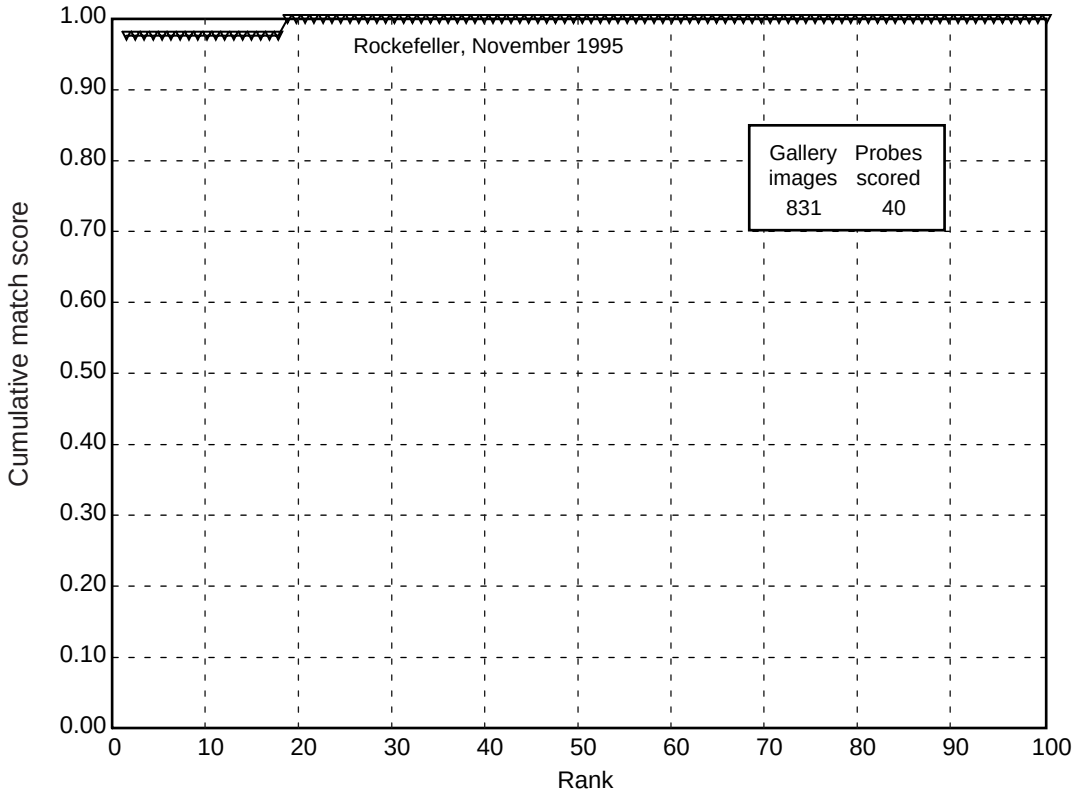


Figure 45. Large gallery test: 30% reduced image size scores (November 1995).

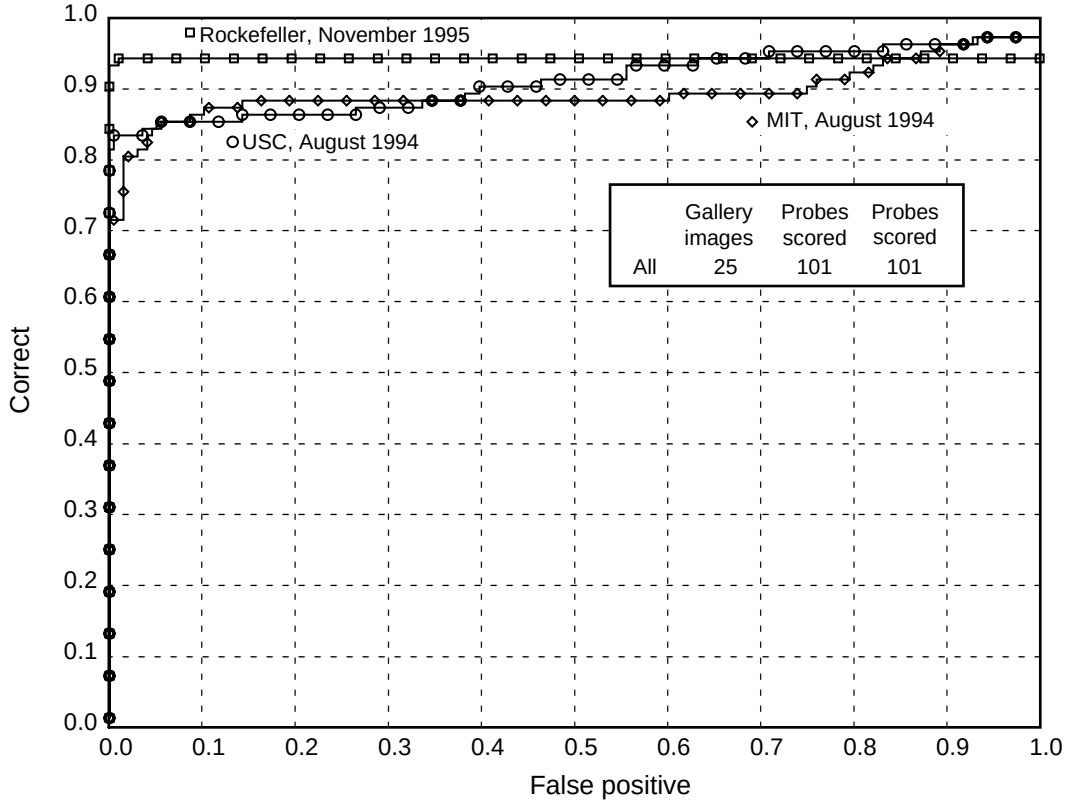


Figure 46. False-alarm test comparison.

8. Summary

Under the sponsorship of DARPA, ARL is conducting the algorithm development and facial database development portions of the FERET program. This program addresses the complex issues of facial recognition that have direct and daily applications to the intelligence and law enforcement communities. The FERET program is currently investigating techniques and technologies that show significant promise in the area of face recognition. The long-term goal of the FERET program is to transition one or more of these algorithms into a fieldable face recognition system.

Face recognition is a very difficult problem that is further complicated by the fact that there are billions of people in the world, but researchers have images of only a few thousand individuals and only a small number of images for each individual. To a human observer, the large number of variations in personal appearance that occur naturally appear normal, but for the developers of face recognition algorithms, these produce large discrepancies and, therefore, problems for the algorithms. It is this overall problem of facial recognition that the FERET program is addressing.

The basic goal of the Phase I test was to baseline algorithm performance on a known database so that we can gauge performance and understand the technical roadblocks to a viable, fielded system. Before the FERET program, most research efforts that addressed the issue of facial recognition used database images that were carefully registered when collected. Since the FERET database was collected to address a real-world problem, it was created to be more realistic, although still providing some control over the type and nature of the images collected.

In support of the Phase I test, a database of over 5000 images was collected. This required numerous collection activities and a large-scale effort to catalogue the images into a database. This database has been requested by and distributed to at least 50 different research groups, greatly assisting researchers in the development and performance evaluation of their algorithms.

The first phase of the FERET program, which included the August 1994 test and evaluation effort, was judged to be very successful. Accomplishments during Phase I included the following:

1. For the first time in face-recognition development, the performance of several algorithms was established against a common baseline.

2. The state of the art was significantly advanced in the area of face recognition. At the start of the program, algorithms worked on either a small database or on databases of images collected under highly controlled conditions. At the end of Phase I, algorithms were working with databases of up to 500 individuals collected under semi-controlled conditions.

3. A database of facial images was established that models real-world conditions.

4. Areas for future research were identified:

    - Increase the size of the database.
    - Increase the number of duplicate images (images of the same person taken at different times).

    Partly based on the results of the first phase of the FERET program, MIT, TASC, and USC were chosen to continue their research efforts in Phase II. Accomplishments during Phase II included the following:

1. Face recognition algorithms were developed that were sufficiently mature that they can be ported to real-time experimental / demonstration systems.
2. The size of the FERET database was increased to 1109 sets of images and 8525 images. This included 225 duplicate sets.
3. TASC proceeded with developing algorithms to extract shape from motion in video sequences.

    From the results of the Phase II test, we concluded that the overall performance for face recognition algorithms had reached a level of maturity that they should be ported to a real-time experimental / demonstration system. The goals of this system will be to

1. develop large-scale performance statistics (this requires long runs over a period of weeks or months in a controlled real-world scenario; an example is detecting and recognizing people as they walk through a door or portal);
2. demonstrate the capabilities of the system to potential end users; and
3. identify weaknesses that cannot be determined in laboratory development efforts or represented in databases collected under the current image acquisition protocol.

    In the future, ARL will continue to address the research being conducted by assisting in the development of a larger and more varied facial database, testing and evaluating new face recognition algorithms being developed, supporting algorithm research and development, and establishing baselines for human performance.

    Future research into facial recognition will require tests that are more robust in design and content. Tests relating to various hair styles, the wearing of glasses, increased variation in rotational angle, and inclination / declination of the face are only a few of the areas where future research is needed. Future test designs will require larger databases consisting of images having a larger range of human variability, such as that obtained over many weeks of observation. Future areas of growth in the collection of database images will include

1. images of individuals taken over an extended period of time,
2. images with a variety of features (e.g., glasses, facial hair, disguises, etc),

3. images of faces at different rotational poses,

4. images with various vertical head positions (inclination and declination of head up to 4°), and

5. video sequences with subjects moving through the field of view.

The performance of face recognition algorithms will probably continue to improve. This was reflected when MIT retook the March 1995 test in August 1996. The results are presented in appendix A.

Acknowledgments

The authors would like to thank the other ARL personnel who have assisted on this program: Michael Lander, Dennis Cook, Quochien (Henry) Vuong, Barbara Collier, and the Technical Publishing Branch; we also thank Frank Shields of SAIC.

Bibliography

Coward, W. M., and D. A. McConathy, *A Monte Carlo study of the inferential properties of three methods of shape comparison*, Am. J. Phys. Anthropol. **99**, No. 3 (1994), 369–377.

DePersia, A. T., P. J. Phillips, and M. K. Hamilton, *The FERET Program: Overview and Accomplishments*. U.S. Army Research Laboratory (1994) (unpublished paper).

Gordon, G. G., *Face recognition from frontal and profile views*, Proc. International Workshop on Automatic Face and Gesture Recognition, M. Bushell, editor (1995).

Lades, M., J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen, *Distortion invariant object recognition in the dynamic link architecture*, IEEE Trans. Computers **42** (1993).

Maurer, T., and C. von der Malsburg, *Single-view based recognition of faces rotated in depth*, Proc. International Workshop on Automatic Face and Gesture Recognition, M. Bushell, editor (1995).

Moghaddam, B., and A. Pentland, *Face recognition using view-based and modular eigenspaces*, Conference on Automatic Systems for the Identification and Inspection of Humans, Proc. SPIE **2277**, San Diego, CA (1994).

Pentland, A., B. Moghaddam, and T. Starner, *View-based and modular Eigenspaces for face recognition*. Proc. Computer Vision and Pattern Recognition **94** (1994).

Turk, M., and A. Pentland, *Eigenfaces for recognition*, J. Cognitive Neurosci. **3**, No. 1 (1991).

University of Illinois at Chicago, *Army Project on Comparing Human Faces*, Final Report for Contract DAAL01-94-K-0114 (1994).

Weng, J., *SHOSLIF: The hierarchical optimal subspace learning and inference framework*, Michigan State University, Technical Report CPS-94-15 (March 1994).

Wilder, J., and R. J. Mammone, *Projection based face recognition*, Conference on Automatic Systems for the Identification and Inspection of Humans, Proc. SPIE **2277**, San Diego, CA (1994).

Wiskott, L., J. M. Fellous, N. Kruger, and C. von der Malsburg, *Face recognition and gender determination*, International Workshop on Automatic Face and Gesture Recognition (1995).

Appendix A. Further Testing at MIT

The development of face recognition algorithms is a dynamic process; today's performance statistics soon become outdated, as old algorithms are improved and new ones developed. After the March 1995 test, the Massachusetts Institute of Technology (MIT) Media Laboratory group continued development of their algorithm and asked to retake the March 1995 test with the new algorithm.[1] The request was granted, and on 13 August 1996, the test was administered.

To support further research in face recognition, after the groups took the March 1995 test, Army Research Laboratory (ARL) released additional images to those groups. The performance in this appendix reflects the MIT group's use of these additional data in developing the algorithm, and the results are compared only with the results obtained with the MIT algorithm tested in March 1995. Figures A-1 to A-3 compare the performance of the March 1995 and August 1996 algorithms: overall scores, scores on FA versus FB images (alternative frontal images), and scores on duplicate images.

The results show a substantial improvement on the duplicate images and reflect a conserted effort to develop algorithms to address the issue of duplicate images. Similar increases in performance can be reasonably expected for all approaches tested. Currently, there is no definite set of performance statistics, because upper limits on the ability of algorithms to recognize faces have not been established.

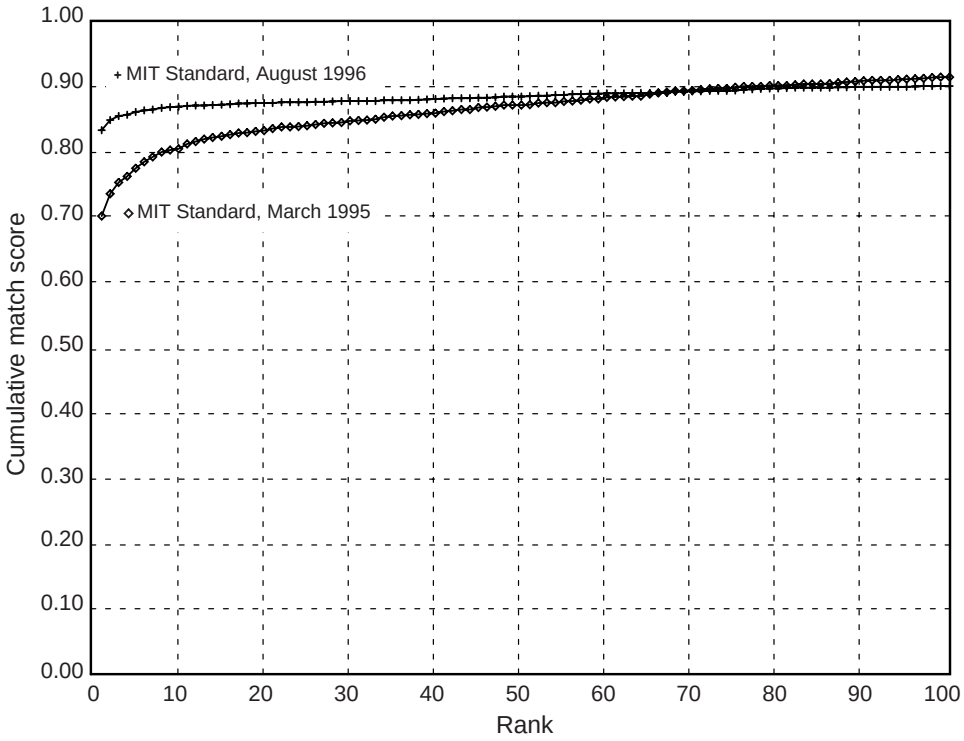


Figure A-1. Comparison of overall scores for March 1995 and August 1996 algorithms.

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

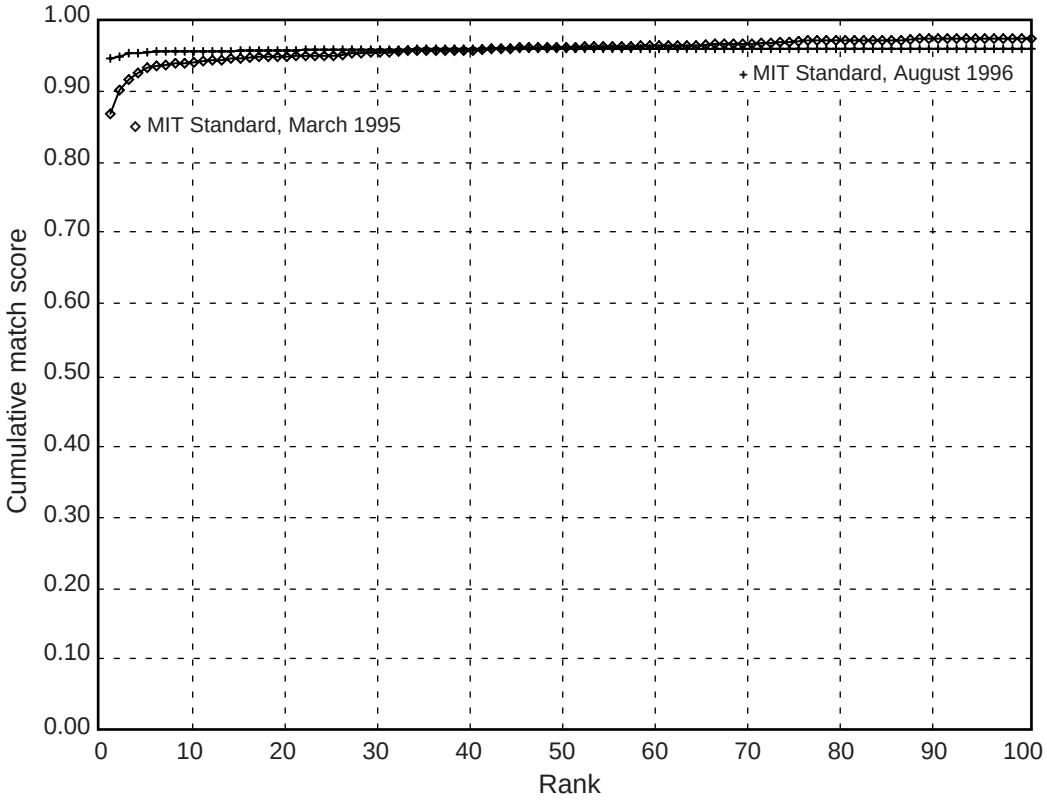


Figure A-2. Comparison of FA versus FB (alternative frontal images) scores for March 1995 and August 1996 algorithms.

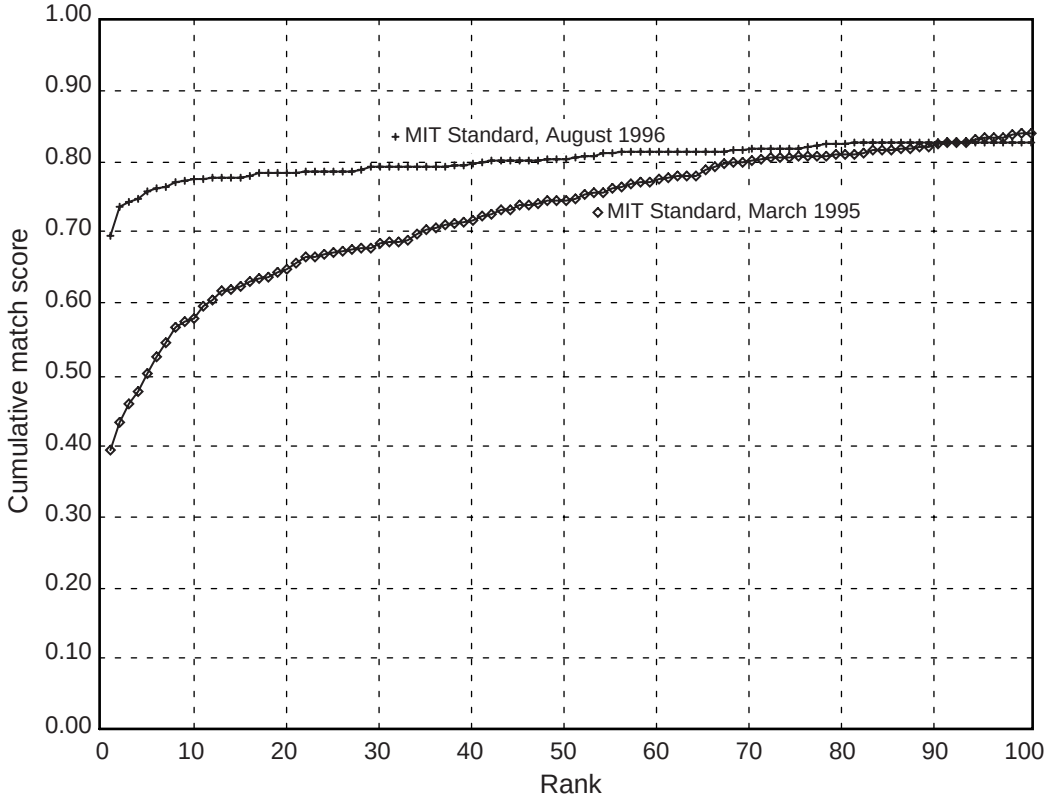


Figure A-3. Comparison of duplicate image scores for March 1995 and August 1996 algorithms.

Appendix B. Availability of Data for Outside Research

To advance the state of the art in face recognition, the Army Research Laboratory (ARL) will make the Face Recognition Technology (FERET) database available to researchers in face recognition on a case by case basis. All requests for the FERET database must be submitted in writing to the FERET technical agent at ARL. Inquiries for further information may be made to the Program Manager at

U.S. Army Research Laboratory
Dr. P. Jonathon Phillips
AMSRL-SE-RT
2800 Powder Mill Rd
Adelphi, MD 20783-1197

Phone: 301-394-5000
e-mail: jonathon@arl.mil

Appendix C. Research Release Form

George Mason University is conducting research on automated means for face recognition. The subjects are expected to allow their pictures to be taken in five poses: frontal, 3/4 view, and/or profile. Participation in this research is voluntary. Full confidentiality will be maintained regarding the identity of the subject, and coding for person-identifiable data will be done with alphanumeric tags. This project has been reviewed according to George Mason University procedures governing your participation in this research. You may also contact the George Mason University Office for Research at 703-993-2295 if you have any questions or comments regarding your rights as a participant in this research.

I understand that these pictures may be published in reports documenting the results of this research.

I have read this form and agree to participate in the study.

Date: _____

Subject signature: _____

Witness: _____

Appendix D. Algorithm Approaches

D-1. MIT Approach

The Massachusetts Institute of Technology (MIT) Media Laboratory Face Processing system consists of a two-stage object detection and alignment stage, a contrast normalization stage, and a feature extraction stage whose output is used both for the recognition stage and for coding the gallery. Object detection begins by locating regions in the image that have a high likelihood of containing a face. It assumes that there is a 3:1 ratio of possible face scales (e.g., that people are between x and $3x$ distance from the camera). Currently four independent and parallel processors are used, one designed for each of the four standard poses (frontal, quarter, half, and full profile). This head localization is performed by multiscale saliency computation. In addition to the saliency computation based on likelihood, the current version incorporates likelihoods based on the first two moments of the grayscale histogram (mean and variance), as well as spatial location. Each of these factors is incorporated independently through the Mahalanobis distances based on previously computed means and covariances from training data. After the best head location and scale are determined, the original image is linearly scaled and translated so that the head is centered in the frame at a fixed scale.

Once the head-centered image is obtained, parallel searches for the four facial features (the left eye, right eye, nose, and mouth) are conducted in essentially the same manner as that for locating the head. The saliency computation is restricted to certain regions (windows) in the head-centered frame and is also modulated by a prior probability distribution for the location of the features in these windows. The top N candidate locations for each feature are verified and pruned of false alarms based on the geometrical constraints of a face (the relative location of the individual features). An exhaustive combinatorial search of all possible pairings of the top N candidates for the four features is performed. For each possible combination (which forms a candidate four-node spatial graph), a likelihood score is generated based on a Mahalanobis distance, in terms of a 12-dimensional feature vector, which consists of the length and orientation of the six links of this graph. The individual scores (likelihoods) of each candidate location are also taken into consideration. The final score is the product of these four individual likelihoods and the likelihood score from their geometry.

The final feature locations are then used to warp the head-centered image so as to align the detected feature locations with those of a canonical model. A rigid transform is used based on the locations of the two eyes in the image with those in the canonical model. After scaling and alignment, the warped image is masked so that the background is removed. It is then normalized by linear remapping of the grayscale to a specified mean and standard deviation.

Finally, the geometrically aligned and normalized image is projected onto a custom set of eigenfaces, producing a feature vector that is then used for recognition, as well as facial image coding of the gallery images.

D-2. Rutgers University Approach

The Rutgers University Center for Computer Aids for Industrial Productivity (CAIP) face-recognition system possesses three attributes that distinguish it from other approaches. The first of these is the use of grayscale projections, wherein a two-dimensional image of a face is compacted into a small number of one-dimensional signatures. These signatures are obtained by the addition of the grayscale values of pixels across the image in a direction perpendicular to the angle of the signature; e.g., horizontal projections are obtained by the addition of pixels across rows, and vertical projections are obtained by the addition of pixels down the columns. This initial stage of data reduction greatly reduces the complexity of the subsequent processing without sacrificing significant amounts of information necessary for recognition. Because robustness to rotation of the head about the vertical axis was important, three signatures are used as a source of features for recognition: the horizontal projection on the original image, the horizontal projection of the image electronically rotated 7° left of the center of the face, and the horizontal projection rotated 7° to the right.

The second attribute is transform coding of the grayscale projections. Transform coding of the sampled projections decorrelates the data, allows for additional data reduction (elimination of high spatial frequencies and the dc term), and distributes the local errors (e.g., due to a smile or frown) over all the output samples in the transform domain. For this effort, the discrete cosine transform (DCT) was used. It provides results closely approaching those of the Karhunen-Loeve transform (the eigenface approach in two-dimensional (2D) systems), but can be computed with a fast algorithm.

The third attribute is training and classifying via the CAIP-developed Neural Tree Network (NTN). The NTN is a hierarchical classifier that effectively combines neural networks and decision trees. It can be implemented cost-effectively on extremely simple hardware, i.e., a single reprogrammable neuron.

The CAIP system was designed to find and identify people standing in front of a uniform, consistently illuminated background. The first step in the process is to segment the person from the background by the computation of an edge picture (the maximum of 0°, +45°, –45°, +90° gradients followed by thresholding and morphological growing to fill in gaps). The edge image was used to set all background pixels in the gray level image to zero. The edge picture was also used to locate the top, left, and right edges of the head. These boundaries established the limits for horizontal and vertical projections. These projections are used to locate the eyes, nose, and mouth. The locations of the eyes and mouth are then used to scale the face to a standard size. The final side of the box around the face is gener-

ated by heuristic techniques for finding the top of the forehead and the chin. Projections are then computed within the region from the forehead to the chin. DCT's are computed on the projections, and low-pass spatial frequency components selected as features for training and recognition.

The NTN classifier performs at least as well as any direct distance-based classifier in finding the most likely candidate for recognition. However, testing results were required to include a rank order of the 50 most likely candidates for each face presented. It was more efficient during the tests to compute a function of the L_1 norm of the distance between the test vector and the training vectors for each member of the database (D_{ui}). The ranking metric is $1 - D_{ui}/D_{umax}$, where D_{umax} was the L_1 norm of the distance from the test vector to the most distant training vector. If the ranking metric is below a given threshold (0.6), the test vector is rejected (as not belonging to the database) if its distance to the mean vector of the database is greater than 1.5 times the distance of the outermost member of the database to the mean of the database. The metric is computed for the feature vectors derived from each of the three projections (0,±7°), stored for each member of the database; the largest is selected as representing the distance to that member. Then, these maximum values are rank-ordered across the database.

D-3. TASC Approach

The major emphasis of the effort by The Analytic Science Company (TASC) is the use of information about the 3D shape of the face to both detect and compensate for viewing angle variation. Most approaches to face recognition rely on low-level image-pattern comparisons to compute similarity between two face images. If the pose of the head is not roughly the same in both images, these types of comparison methods will produce incorrect results. As the number of subjects in the database increases, this source of error will become more and more important.

The computation of 3D structure or position information requires the use of multiple views of the subject. Since the 3D pose of the head cannot be computed from one image, it is not possible even to detect this source of error if only one image of the subject is available. Under this effort, two uncalibrated views, frontal and profile, were considered. The profile view provides information about the relief of the face that cannot be computed from the frontal view. This information can be used to better distinguish two subjects whose frontal views might be incorrectly compared because of differences in view angle (e.g., tilt of the chin). This scenario is one of the simplest multiview conditions available, and also describes a real-world application: matching against traditional mugshots. Hence this problem is valuable both in the short term and in the long term as a baseline for future work, in which 3D models will be constructed from more complex multiview scenarios (e.g., video sequences).

The TASC system processes both the frontal and profile views in a similar fashion. Feature extraction is used first to identify two fiducials in the im-

age that are used to perform geometric normalization, including adjustments of image plane rotation and scale. Since the images are uncalibrated, these normalization factors are specific to each view. Template regions are extracted from the normalized images and stored in the database along with the location of fiducials from the original images. A total of five template regions are extracted. At the lowest level, two subjects are compared on the basis of general pattern-matching techniques with only the extracted normalized templates. This comparison method performed quite well on the database provided, with the largest source of error being the location of the fiducial points used for geometric normalization.

The system can be run in two modes. Comparison can be made on the basis of only the frontal view, or on both views.

D-4. USC Approach

The general approach to face recognition used by the University of Southern California (USC) Computational Vision Lab is based on the dynamic link architecture (DLA) theory of brain function. The program, known as SCFacerec, is an algorithmic abstraction of DLA called elastic graph matching, which is better suited for processing on conventional digital computers than is DLA.

Broadly speaking, elastic graph matching finds a mapping between the image and model domains and compares features sampled at corresponding points in the mapping. Two stages of elastic graph matching are used by SCFacerec: a spatially coarser stage, in which the face is found and normalized with respect to scale and position in the image, and a finer stage, in which features of the face are located for comparison with a gallery of mug shots. The same basic graph matching scheme is used for both coarse and fine stages; indeed, many of the same functions are called in both steps.

SCFacerec may be broken down into the following components, each of which is described in more detail below: (1) a fiducial graph, (2) lists of features or “jets,” (3) a similarity function for comparing jets, (4) heuristic moves for registering the graph with a facial image, and (5) a prior knowledge about faces for use in graph matching (also known as general face knowledge or “GFK”).

The fiducial graph consists of a graph of nodes corresponding to anatomically identifiable points on the face. Choice of a reproducible set of nodes for the graph allows comparison of the same facial points across different poses and between individuals. Fiducial graphs are also necessary for the use of differential weighting of graph nodes in recognition and to introduce jet transformations to account for the effects of rotation in depth.

The system uses a bank of multiple-scale and multiple-orientation Gabor wavelet filters for feature extraction. This representation is based on a simple model of the receptive fields found experimentally in the neurons of the mammalian primary visual cortex. Use of these features gives the

system insensitivity to changes in absolute illumination and, with the similarity measure described below, to overall changes in contrast of an image. Use of the absolute power (i.e., modulus of the wavelet transform) of the Gabor features leads to some insensitivity to the exact positioning of the graph nodes. The responses to the eight orientations and five spatial frequencies of Gabor wavelet filters used by SCFacerec are coded as a 40-dimensional vector or jet.

Jets are extracted and compared at each node of the graph both in the graph-matching phase of the algorithm and in comparing faces in probe and gallery image lists. The generalized direction cosine between two jets is used for the comparison. The normalization of jet length in the calculation of the direction cosine leads to an insensitivity to changes in the level of contrast in the image. In positioning graph nodes (locations to extract jets), a similarity measure is used that also takes into account the phase of the Gabor transform. In comparing graphs for identity recognition, only the magnitude of the transform is used.

The algorithm samples the image in a hierarchical fashion to determine the position and scale of the face. This is effectively a three-parameter search. Parameter changes or graph moves are accepted if the match with the GFK (explained below) is improved. Finally, each node is allowed to "diffuse" or move independently of the rest of the graph to improve the fit with the individual probe face. Graphs are automatically positioned on both probe and gallery faces by this method. Jets may then be extracted and compared at corresponding points in probe and gallery graphs for recognition.

The general face knowledge (GFK) consists of a stack of example faces on which fiducial graphs have been positioned manually. A GFK stack usually contains between 10 and 70 examples, depending on the requirements of the matching problem. Once constructed, a GFK stack may be reused for different probe and gallery stacks: reliable matching is fairly insensitive to the exact details of the examples used to construct the GFK. For each trial position of a graph node in the matching process, the GFK is searched for the most similar jet at that node. This information is used to compute an overall similarity of the probe graph with the GFK stack and evaluate whether a graph move improves or worsens the fit of the graph to the probe face.

The components described above are integrated into a system with a convenient graphical user interface. The system may be run in batch modes, for testing recognition performance, or in demo mode, where individual images are processed for recognition.

Distribution

Admnstr
Defns Techl Info Ctr
Attn DTIC-OCP
8725 John J Kingman Rd Ste 0944
FT Belvoir VA 22060-6218

US Customs Service Rsrch & Dev Div
Attn B Armstrong
1301 Constitution Ave NW
Washington DC 20229

US Dept of Energy
Ofc of Safeguard & Security
Attn NN-513.4 C Pocratsky
Washington DC 20585

NIJ/OST
Attn A T DePersia
Attn R Downs
633 Indiana Ave NW
Washington DC 20531

Army Rsrch Lab Physics Div
Attn AMSRL-PS W Gelnovatch
FT Monmouth NJ 07703

DARPA/ISO
Attn T Stratt
3701 N Fairfax Dr
Arlington VA 22203-1714

Hdqtrs Dept of the Army
Attn DAMO-FDQ MAJ M McGonagle
400 Army Pentagon
Washington DC 20310-0460

Nav Surface Warfare Ctr
Attn Code B07 D Grenier
Attn Code B07 J Pennella
17320 Dahlgren Rd Bldg 1655
Dahlgren VA 22448-5100

Ofc of Nav Rsrch
Attn Code 342CN T McKenna
300 N Quincy Stret
Arlington VA 22217-5000

Fed Bureau of Investigation
Advanced Technologies Unit
Attn M Gilchrist Rm 11401
10th Stret & Pennsylvania Ave NW
Washington DC 20537

Fed Bureau of Investigation
Attn J K Kielman
Ninth Stret & Pennsylvania Ave NW
Washington DC 20535

United States Dept of Justice
Immgrtn & Naturalization Service
Attn V Fuentes Rm 4014
Attn S Schroffel
425 I St NW
Washington DC 20536

United States Dept of Justice
Drug Enforcement Administration
Attn A Antenucci
8199 Backlick Rd
Lorton VA 22079-1414

National Security Agency
Attn R22 Murley
9800 Savage Rd Ste 6516
FT Meade MD 20755-6516

Oak Ridge Natl Lab
CSM Div
Attn M B Shah
Bldg 6025
Oak Ridge TN 37831-6364

University of Wollongong
Dept of Computer Sci
Attn J Fulcher
Northfields Ave
Wollongong NSW 2522
Austrialia

Distribution

Dept of Mathematical Sci
Univ of Aberdeen
Attn I Craw
The Edward Wright Building Dunbar Stret
Aberdeen AB9 2TY
England

Dir Imaging Lab
Univ of IL at Urbana-Champaign
Attn T S Huang
Beckman Institute
405 N Matthews Ave
Urbana IL 61801

George Mason Univ
Dept of Computer Science
Attn H Wechsler
4400 University Dr
Fairfax VA 22030-4444

MA Instit of Techlgy
The Media Lab
Attn A Pentland Head Perceptual
Computing Sect
20 Ames St Rm E15-387
Cambridge MA 02130

MA Instit of Techlgy
Artificial Intelligence Lab
Attn T Poggio
545 Technology Square Rm NE 43-787
Cambridge MA 02139

Institut fuer Neuroinformatik, ND 03-36
Attn C von der Malsburg
Ruhr-Universitaet Bochum
44780 Bochum
Germany

Michigan State Univ
Attn J J Weng
A733 Wells Hall
East Lansing MI 48824-1027

Rockefeller Univ
Attn J J Atick
1230 York Ave
New York NY 10021

Rutgers Univ
Ctr for Computer Aids for Ind Prodctvty
Attn J Wilder
Frelinghuysen Rd Core Bldg
Piscataway NJ 08855-1390

Univ of Manchester
Dept of Medical Biophysics
Attn A Lanitis
Attn C Taylor
Oxford Rd Stopford Building
Manchester M13 9PT
United Kingdom

Univ of South Florida
Attn ENB 118 K W Bowyer
4202 East Flowler Ave
Tampa FL 33620-5399

Univ of Southern California
Dept of Psychology
Attn I Biederman
Hedco Neurosciences Bldg MC 2520
Los Angeles CA 90089-2520

Univ of Southern California
University Park HNB 007
Attn C von der Malsburg
Attn M Lyons
3614 Watt Way
Los Angeles CA 90089-2520

Adaptive & Learning Sys Grp
RACAL Rsrch Limited
Attn R Rickman
Worton Dr Worton Grange Industrial Estate
Reading Berkshire RG2 0SB
England

Distribution

Amherst Sys Inc
Attn C Bandera
30 Wilson Rd
Buffalo NY 14221

Head Dept 2
ATR Human Infor Processing Rsrch Lab
Attn S Akamatsu
2-2 Hikaridai
Seiko-cho Soraku-gun Kyoto 619-02
Japan

CISRO Div of Radiophysics
Leader, Sig & Imagning Technl Program
Attn G Poulton
PO Box 76
Epping New South Wales 2121
Australia

Computational Neurobiology Lab
The Salk Inst for Biological Studies
Attn L Wiskott
PO Box 85800
San Diego CA 92186-5800

David Sarnoff Rsrch Ctr
Dir Business Dev
Attn J L Frank
201 N Washington Rd
Princeton NJ 08543-8619

David Sarnoff Rsrch Ctr Inc
Attn CN5300 J Matey
Princeton NJ 08543-5300

Dept 2 ATR HIP Labs
Attn M Lyons
2-2 Hikari-dai Seika-cho
Soraku-gun Kyoto 619-02
Japan

EPFL DE/LTS
Attn J Bigun
Lausanne CH-1015
Switzerland

Excalibur Technologies
Attn M Willey
1959 Palomar Oaks Way
Carlsbad CA 92009

Fraunhofer Gesellschaft
Attn U Dieckmann
Institut fuer Integrierte Schaltungen
Am Weichselgarten 3 Erlangen D-91058
Germany

HNC Software Inc
Attn Y-T Zhou
5930 Cornerstone Ct W
San Diego CA 92121-3728

Info Technl Solutions Inc
Attn R Kukich
2 Eaton St Ste 908
Hampton VA 23669

Infotec Dev Inc
Attn EDS-ITP W Shen
1420 Spring Hill Rd Ste 205
McLean VA 22102

Interval Rsrch Corp
Attn G Gordon
1801 Page Mill Rd Bldg C
Palo Alto CA 94304

Miros Inc
Attn R Leibowitz
572 Washington Stret #18
Wellesley MA 02181

Nielson Media Rsrch
Attn ENT M Lee
375 Patricia Ave
Dunedin FL 34698

Ofc of Special Techlgy
Attn J David
10530 Riverview Rd
FT Washington MD 20744-5821

Distribution

SAIC
Attn F Shields
4001 N Fairfax Dr Ste 300
Arlington VA 22203

Samuels & Assoc
Attn A I Samuels
145 Green Acres Rd
Elizaville NY 112523

Sensci Corp
Attn L Acchione
1423 Powhatan Stret Ste 4
Alexandria VA 22314

Siemans Nixdorf
Advanced Technologies GmbH
Attn N Kunstmann
Scharfenberger Str 66 Dresden D-01139
Germany

Siemens AG
Attn B Wirtz
ZFE ST SN 5
Otto Hahn Ring 6 Munich 81739

SJB Services
Biometric Technology Today
Attn G Roethenbauch
London House Broad Stret
Somerton Somerset TA11 7NH
United Kingdom

Social Services Dev Ctr
Attn A T Simonelli
740 Notre Dame W 13th Floor
Montreal Quebec H3C 3X6
Canada

TASC
Attn G Gordon
55 Walkers Brook Rd
Reading MA 01867

TRW Integrated Engrg Div
Attn R Smith
One Federal Sys Park Dr
Fairfax VA 22033

Army Rsrch Lab
Attn AMSRL-WT I May
Aberdeen Proving Ground MD 21005-5000

US Army Rsrch Lab
Attn AMSRL-CI W H Mermagen Sr
Attn AMSRL-HR R L Keesee
Attn AMSRL-SL J Wade
Aberdeen Proving Ground MD 21005-5425

US Army Rsrch Lab
Attn AMSRL-VS W Elber
Hampton VA 23681-0001

US Army Rsrch Lab
Attn AMSRL-VP R Bill
21000 Brookpark Rd
Cleveland OH 44135-3191

US Army Rsrch Lab
Attn AMSRL-BE D R Veazey
White Sands Missile Range NM 88002-5513

US Army Rsrch Lab
Attn AMSRL-CS-AL-TA Mail & Records Mgmt
Attn AMSRL-CI-LL Tech Lib (3 copies)
Attn AMSRL-CS-AL-TP Tech Pub (5 copies)
Attn AMSRL-PP B Fonoroff
Attn AMSRL-SE J M Miller
Attn AMSRL-SE-IS V DeMonte
Attn AMSRL-SE-R E Burke
Attn AMSRL-SE-RT J Phillips (50 copies)
Attn AMSRL-SE-RT P Rauss
Attn AMSRL-SE-RT M Hamilton
Attn AMSRL-SE-RT S Der
Adelphi MD 20783-1197

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1996	Final, September 1993 to August 1996

4. TITLE AND SUBTITLE
FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results

5. FUNDING NUMBERS

6. AUTHOR(S)
P. Jonathon Phillips, Patrick J. Rauss, and Sandor Z. Der

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
Attn: AMSRL-SE-RT
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-995

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
DARPA
3701 N. Fairfax Drive
Arlington, VA 22203-1714

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
AMS code: 63889E00000
ARL PR: 51DJ10

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

As part of the Face Recognition Technology (FERET) program, the U.S. Army Research Laboratory (ARL) conducted supervised government tests and evaluations of automatic face recognition algorithms. The goal of the tests was to provide an independent method of evaluating algorithms and assessing the state of the art in automatic face recognition. This report describes the design and presents the results of the August 1994 and March 1995 FERET tests. Results for FERET tests administered by ARL between August 1994 and August 1996 are reported.

14. SUBJECT TERMS
Face recognition, face database collection, recognition algorithms

15. NUMBER OF PAGES
72

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

DEPARTMENT OF THE ARMY
U.S. Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783-1197