ELSEVIER

Image and Vision Computing 16 (1998) 295–306

image AND vision COMPUTING

Promising Research

The FERET database and evaluation procedure for face-recognition algorithms

P. Jonathon Phillips[a,*], Harry Wechsler[b], Jeffery Huang[b], Patrick J. Rauss[a]

[a]*U.S. Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783, USA*
[b]*George Mason University, Computer Science Department, Fairfax, VA 22030, USA*

Received 28 April 1997; revised 1 September 1997; accepted 31 October 1997

Abstract

The Face Recognition Technology (FERET) program database is a large database of facial images, divided into development and sequestered portions. The development portion is made available to researchers, and the sequestered portion is reserved for testing face-recognition algorithms. The FERET evaluation procedure is an independently administered test of face-recognition algorithms. The test was designed to: (1) allow a direct comparison between different algorithms, (2) identify the most promising approaches, (3) assess the state of the art in face recognition, (4) identify future directions of research, and (5) advance the state of the art in face recognition. © 1998 Published by Elsevier Science B.V. All rights reserved.

Keywords: Face recognition; Algorithm evaluation; Image databases

1. Introduction

In the last few years, face recognition has become an active area of computer vision. The interest has been fueled by potential applications and by the simultaneous development of algorithmic techniques and inexpensive computers with the computational power to run these algorithms. These developments yielded a large number of papers on face recognition, with a majority reporting outstanding recognition results (usually $> 95\%$ correct recognition) on limited-size databases (usually < 50 individuals). Few of these algorithms reported results on images from a common database; fewer met the desirable goal of being evaluated against a standard testing protocol that includes separate training and testing sets. As a consequence, there was no way to make a quantitative assessment of the algorithms' relative strengths and weaknesses. Unfortunately, this is not an isolated case, but an endemic problem in computer vision research.

In the computer vision literature, various articles and discussions have argued for methods that would allow comparative assessments of algorithms [1–5]. The benefits of such methods would include: (1) placing computer vision on a solid experimental and scientific ground, (2) assisting in developing engineering solutions to practical problems, (3) allowing accurate assessment of the state of the art, and (4) providing convincing evidence to potential users that computer vision research has found a solution to their problems.

Despite these well-founded arguments, the computer vision community for the most part has not heeded the call. There are a few exceptions: in handwritten character recognition, standard databases are available from the National Institute of Standards and Technology (NIST) (NIST Special Database 3) and from the UNIPEN consortium of companies and universities [6]. The recent work of Heath et al. [7] and Hoover et al. [8] established methods of comparing edge detectors in range and intensity images. In face recognition, Robertson and Craw [9] propose a method for evaluating algorithms and discuss how a database of facial images should be collected. However, they do not compare algorithms and limit experiments to a database of five subjects.

In this paper we present a comprehensive method of evaluating face-recognition algorithms, developed as part of the Face Recognition Technology (FERET) program [10,11]. The FERET evaluation methodology consists of an integrated data collection effort and testing program. These two parts are integrated through the FERET database of facial images; the database is divided into a development portion, which is provided to researchers, and a sequestered portion for testing. The partition of the database enables algorithms to be trained and tested on different, but related, sets of images. The FERET evaluation procedure is a set of

* Corresponding author. Current e-mail: jonathon@nist.gov

PII S0262-8856(97)00070-X

standard testing protocols: the FERET tests are independently administered and each test is completed within three days. The use of a standard testing protocol allows for the direct comparison of algorithms developed by different groups, as well as for measuring improvements made by any single group over time.

2. Test design principles

The FERET tests are administered using a testing protocol, which states the mechanics of the tests and the manner in which the test will be scored. In face recognition, for example, the protocol states the number of images of each person in the test, how the output from the algorithm is recorded, and how the performance results are reported.

There is a direct connection among the problem being evaluated, the images in the database, and the testing protocol. The testing protocol and the images define the problem to be evaluated. The characteristics and quality of the images are major factors in determining the difficulty of the problem. For example, if the faces are in a predetermined position in the images, the problem is different from that for images in which the faces can be located anywhere in the image. In the FERET database, variability was introduced by the inclusion of images taken at different dates and locations (see Section 4.2). This resulted in changes in lighting, scale, and background.

The goals for the FERET evaluation process were to assess the state of the art, advance the state of the art, and point to future directions of research. Accomplishing all three goals was a delicate process, and the keys to success were the database and the tests. If algorithms existed that could easily solve the problem, then the evaluation process would be reduced to 'tuning' existing algorithms. On the other hand, if the images defined a problem that was beyond current algorithmic techniques, then the results would have been poor and would have not allowed an accurate assessment of current algorithmic capabilities. The key was to find the right balance, so that if the problem formulated could not be solved satisfactorily by existing methods, it would be possible to develop algorithms that could solve it.

The collection of the FERET database was initiated in September 1993, and the first FERET test was administered in August 1994. A standard database of facial images was established in the first year and made available to researchers; this database provided the images for the Aug94 FERET test. (Throughout this article, date-based names such as Aug94 are used to refer to the same FERET test, even when the tests were administered on other dates.) The Aug94 FERET test established a performance baseline for fully automatic face-recognition algorithms. A fully automatic algorithm does not require the location of the face in the image as input: the algorithm locates and identifies the face in the image.

The Aug94 FERET test was designed to measure the performance of algorithms that could automatically locate, normalize, and identify faces from a database. The test consisted of three subsets: the large gallery, false alarm, and rotation tests. The first tested the ability of algorithms to recognize faces from a set of 317 known individuals (referred to as the *gallery*; an image of an unknown face presented to the algorithm is a *probe*, and the collection of probes is called the *probe set*). The second subtest, the false-alarm test, measured how well an algorithm rejects faces not in the gallery. The third baselined the effects of pose changes (rotations) on performance. On the basis of the Aug94 FERET test, it was concluded that algorithms needed to be evaluated on (1) larger galleries and (2) a substantially increased number of duplicate probes. (A *duplicate* is defined as an image of a person whose corresponding gallery image was taken on a different date.)

A second FERET test, first administered in March 1995 (and referred to as the Mar95 FERET test), was designed based on the conclusions from the Aug94 FERET test. The Mar95 FERET test evaluated algorithms on larger galleries and probe sets with a greater number of duplicates. This required that additional images be collected, with an emphasis on images of the same people taken months or years apart.

3. Algorithms evaluated

We report results of FERET evaluations of face-recognition algorithms developed at three institutions: the MIT Media Laboratory, the Computational Neuroscience Laboratory of the Rockefeller University, and the Computational and Biological Vision Laboratory of the University of Southern California (USC).

The MIT Media Laboratory algorithm is based on principal component analysis. In principal component analysis, a set of reference faces is used to compute a set of 'eigenfaces' [12–15], which are the eigenvectors produced by the analysis. A face in an image is represented as its projection onto the eigenvectors. The algorithm identifies faces by comparing their projections in eigenspace. The eigenvectors are computed from a subset of the images in the database and are not modified as the database changes.

The Rockefeller algorithm is based on factorial learning and local feature analysis [16–18]. Local feature analysis is a sparsely distributed coding of decorrelated local features. This encoding is a local low-dimensional compact representation of the face. Local feature analysis for representing faces is presented in Penev and Atick [17], but details for face recognition are not discussed.

The USC algorithm uses the dynamic-link-architecture paradigm, which projects an image onto a set of Gabor jets [19–21]. A Gabor jet is a set of Gabor wavelets with different scales and orientations, all centered at the same pixel. The location of each jet is a vertex of a planar graph, where the planar graph is a geometric model of the face. A face is represented as the coefficients derived from

projecting the image onto the Gabor jets, and the distances between vertices in the graph. The similarities between the two faces are determined by comparison of the Gabor jet coefficients and the graphs; this process is referred to as elastic graph matching.

4. FERET database

4.1. Purpose

The FERET database was established to support both algorithm development and evaluation. Two guiding principles were followed. First, the evaluation of algorithms requires a common database of images for both development and testing. In the FERET evaluation, the images in the test are from both the development and sequestered portions of the FERET database. Second, the variety and difficulty of the problems defined by the images in the database must increase incrementally.

The need to test algorithms against a database is obvious, but the development function of the database is equally important (if less obvious). For the evaluation procedure to produce meaningful results, the images in the developmental portion of the database must resemble those on which algorithms are to be tested. The development and testing data sets must be similar in both quality and quantity. For example, if the test will consist of a gallery of 1000 individuals, it is not appropriate for the development database to consist of 50 individuals. The algorithms tested will be only as good as the database from which they are developed. The FERET evaluation procedure followed this principle by partitioning the FERET database into the developmental and sequestered portions, where the developmental portion was representative of the sequestered portion (details are provided in Section 4.2).

The other principle is that the problem defined by the images in the database must mesh with the current level of algorithm development, and the difficulty of the database must grow as the sophistication of the algorithms increases. As explained in Section 2, if the database defines a problem that is too easy, testing the algorithm becomes a mere tuning exercise. At the other extreme, if the problem is too far beyond the state of the art, the test will not produce any meaningful results. To prevent the FERET database from becoming stale, we continuously expanded and adjusted the database to the state of the art in face recognition.

At the beginning of the database collection effort (September 1993), most algorithms worked on small databases (usually < 50 individuals). In these databases, a face was usually rigidly aligned in the image so that the algorithms did not have to locate or normalize it. The database at the MIT Media Laboratory [12] and the database used in Lades et al. [19] were two notable exceptions. The Media Laboratory database of 7562 images was collected with the eyes registered and the faces against a black background. In Lades et al., the database consisted of 88 images taken without rigid alignment of the face.

Based on the state of the art in September 1993, the FERET database collected images to support the development of algorithms that could locate a face in an image and recognize that face from a gallery of approximately 300 individuals. The images were collected without rigid alignment of the faces. This allowed variations in scale, position of the face in the image, illumination, and pose. Following the Aug94 FERET test, additional images were collected to support evaluation of algorithms on (1) images of people taken at different dates and (2) galleries larger than those in the Aug94 FERET test.

4.2. Description

The images in the FERET database were initially acquired with a 35-mm camera. The film used, color Kodak Ultra, was processed and the images placed onto a CD-ROM via Kodak's multiresolution technique. These color images were retrieved from the CD-ROM and converted into eight-bit gray-scale images. 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. Each image was assigned a unique file name that encodes the image ground truth. This includes the subject's identity number, the nominal pose of the image, the date the image was taken, and special variations. The identity number was keyed to the person photographed, so that any future images collected of that person would have the same ID number.

The facial images were collected in 11 sessions from August 1993 to December 1994. Conducted at George Mason University and at US Army Research Laboratory facilities, the session lasted one or two days, and the location and set-up did not change during a session. In an effort to maintain a degree of consistency throughout the database, the same physical set-up was used in each photography session. However, because the equipment was reassembled for each session, there was variation from session to session (Fig. 1). Each image consisted primarily of an individual's head and neck, and sometimes the upper part of the shoulders.

Images of an individual were acquired in sets of 5–11 images, collected under relatively unconstrained conditions (see Fig. 2). Two frontal views were taken (**fa** and **fb**); a different facial expression was requested for the second frontal image (Fig. 3 shows variations between **fa** and **fb** images). Images were also collected at the following head aspects: right and left profile (labeled **pr** and **pl**), right and left quarter profile (**qr** and **ql**), and right and left half profile (**hr** and **hl**). Additionally, five extra locations (**ra**, **rb**, **rc**, **rd**, and **re**) irregularly spaced among the basic images, were collected if time permitted. To add simple variations to the database, the photographers sometimes took a second set of images, for which the subjects were asked to put on their glasses and/or pull their hair back. Sometimes a second

Fig. 1. Duplicate images (examples of variations between sessions).

set of images of a person was taken on a later date; such a set of images is referred to as a duplicate set. Such duplicate sets result in variations in scale, pose, expression, and illumination of the face (Fig. 4). For some people, there was nearly a year between their first and last sittings, with some subjects being photographed multiple times (Fig. 1).

By August 1994, 673 sets of images had been collected, for 4143 images. By March 1995, 1109 sets of images were in the database, for 8525 total images. The database contained 884 individuals and 225 duplicate sets of images. The development portion of the database for the Aug94 and Mar95 tests consisted of 304 individuals (327 sets, 1843 images). This was the largest database available to most researchers in August 1994 and March 1995.

To aid in the evaluation of algorithms, we augmented the sequestered database with a set of digitally altered images. The digital modifications changed scale, illumination, or color of clothes. To test sensitivity to scale changes, we reduced 40 images by 10, 20, and 30% along each axis. To study the effects of illumination, we changed illumination levels: 40 images were non-linearly changed by 40 and 60%. We reversed the color of the clothes to see if the algorithms were using cues from the clothing for recognition.

5. Decision theory and performance evaluation

The basic models for evaluating algorithm performance are the closed and open universes. In a closed universe, every probe is in the gallery; in an open universe, some probes are not in the gallery. Both models reflect different and important aspects of face-recognition algorithms and report different performance statistics.

The closed-universe model allows us to ask how good an algorithm is at identifying a probe image; the question is not always 'is the top match correct?' but 'is the correct answer in the top n matches?'. This lets us know how many images have to be examined to obtain the desired level of performance. The performance statistics are reported on a cumulative match score (see Fig. 6 below). The horizontal axis gives the rank, and the vertical axis is the percentage correct. For example, for the MIT curve in Fig. 6 below, the correct answer was rank 1 for 80% of the probes scored, and the correct answer was rank 10 or less for 87% of the probes scored (in other words, the correct answer was in the top 10 87% of the time). The computation of a cumulative match score is not restricted to the entire probe set. In fact, if the complete probe set contains probes that are not in the gallery, the cumulative match score cannot be computed. The cumulative match score can be calculated for any subset of the probe set. We do this to evaluate an algorithm's performance on different categories of probes, i.e. rotated probes or scaled probes. The computation of the score is quite simple. Let $\mathcal{P}$ be the number of probes to be scored, and R_k the number of these probes in the subset that are in the top k. The fraction reported correctly is $R_k/\mathcal{P}$.

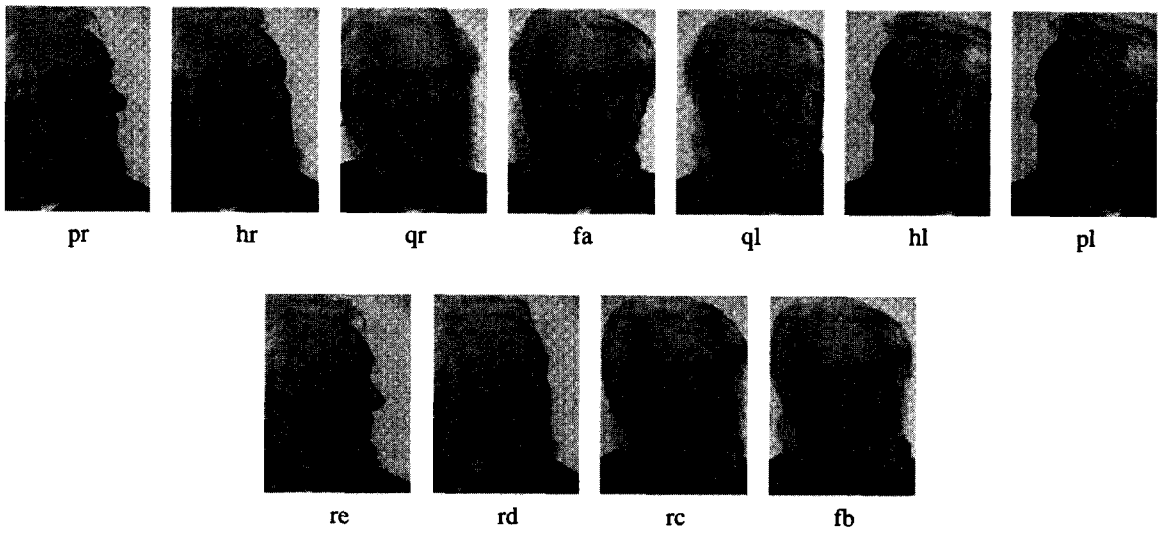

Fig. 2. Image set.

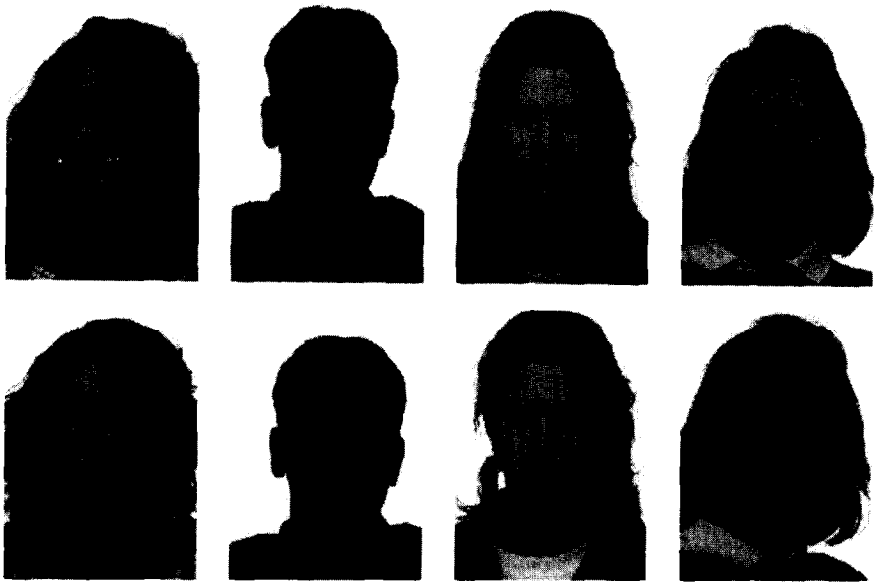

Fig. 3. Variation between **fa** (top row) and **fb** (bottom row) images.

In an open-universe test, the results are reported on a receiver operating characteristic (ROC). The ROC plots the trade-off between the probability of false alarm and the probability of correct identification. In an open-universe test there are two classes of probes. The first class is probes that are 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 $P_F = \hat{F}/F^*$, where F^* is the number of probes not in the gallery and $\hat{F}$ is the number of probes reported as false alarms. The second class is correctly identified probes, whose performance is characterized by P_I. The probability $P_I = \hat{I}/I^*$, where I^* is the size of a set of probes and $\hat{I}$ is the number of these probes that are correctly identified.

Fig. 4. Differences between duplicate images of the same person. In each column are two images of the same person taken on different dates. The images show variations is pose, scale, illumination, background, and overall effects of time between images.

There is a trade-off between P_F and P_I. If no probes are declared to be in the gallery, then $P_F = 0$ and $P_I = 0$. At the other extreme, if every probe is reported to be in the gallery, 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. 12 below; the horizontal axis is P_F and the vertical axis is P_I).

From the ROC, it is possible to compare algorithms. However, it is not possible to compare two algorithms from a single performance point from each. Say we are given algorithm **A** and algorithm **B**, along with a performance point (P_I^A, P_F^A) and (P_I^B, P_F^B) (false-alarm and identification probabilities) from each. Algorithms **A** and **B** cannot be compared from (P_I^A, P_F^A) and (P_I^B, P_F^B), for two primary reasons: 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. If one is given 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 returns the gallery image that is most similar to the probe. In addition, a confidence score of this match is reported, with a high score implying greater likelihood similarity. One generates the ROC by varying a threshold on the confidence score. For a given threshold, a probe is estimated to be in the gallery if the confidence score is above the threshold; otherwise, it is estimated to be not in the gallery. A false alarm occurs when the confidence score of a probe that is not in the gallery is above the threshold. A probe is correctly identified if the

algorithm estimates the correct identity, and the probe's confidence score is greater than the threshold.

In generating the ROC, P_F and P_F are recomputed for each threshold. Initially, the threshold is set higher than the highest match score. This will generate the point $P_F = 0$ and $P_I = 0$. As the threshold is incrementally lowered, P_F and P_I monotonically increase, and the ROC is swept out.

6. Testing on the FERET database

As described in Section 2, the Aug94 and Mar95 FERET tests and evaluation procedures were developed for different purposes. The Aug94 FERET test was designed to baseline face-recognition algorithms, and the Mar95 FERET test measured progress in the abilities of algorithms to handle large galleries and duplicate probes. The Aug94 FERET test procedure consisted of a suite of three subtests that evaluated different aspects of face-recognition algorithms: the large-gallery, false-alarm and rotation tests. The Mar95 FERET test consisted of only a large-gallery test.

In this paper we present a summary of the major results for both tests; we show improvements in algorithm performance from August 1994 to August 1996. A full report on all aspects of both tests can be found in Phillips et al. [10].

We only discuss results for those groups that took the Mar95 FERET test (MIT, Rockefeller, and USC). We present the following: Aug94 FERET large-gallery test results for MIT and USC; Aug94 FERET false-alarm results of MIT, Rockefeller, and USC; and Mar95 FERET test results for MIT, Rockefeller, and USC. The tests were administered at different dates to the groups. Table 1 summarizes the dates that the tests were administered.

The test was designed so the algorithms had to be fully automatic. (Thus, algorithms that required the face to be in a specified position were precluded from taking the Aug94 and Mar95 tests.) The processing of the gallery and the probe images was done without human intervention. The input to the algorithms for both the gallery and the probe was a list of image names, along with the nominal pose of the face in the image. The faces in the images were not placed in a predetermined position or normalized. If such steps were necessary, then the prepositioning or normalization had to be performed by the face-recognition system. For both tests, a group had three days to complete the test on less than 10 UNIX workstations (this limit was not reached). We did not record the time or number of workstations because execution time can vary according to machines used, machine and network configuration, and the amount of time the developers spent optimizing their code (we wanted to encourage algorithm development, not code optimization). (We imposed the time limit to encourage the development of algorithms that could be incorporated into fieldable systems.)

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

At the start of the test, the testee was given two lists of images: the gallery and the probe set. Fig. 5 presents a schematic diagram of the testing procedure. To ensure that matching was not done by file name, we gave the images random names. The nominal pose of each face was provided to the testee.

6.1. Large gallery tests

One of the major questions in face recognition is the performance of algorithms against large galleries. Before the FERET database was assembled, this was a hard question to answer because most databases consisted of less than 100 images, and the few large databases in existence were not widely available. To address this issue, we designed the large-gallery tests, which consisted of as large a gallery as the FERET database could provide at the time of the test.

Two large-gallery tests were designed: one for the Aug94 FERET test and one for the Mar95 FERET test. In both tests, the gallery consisted of one image per person, with the Aug94 FERET test having 317 individuals and the Mar95 FERET test having 831. (The selection of one image per person was a design decision that we felt reflected law enforcement scenarios. The number of images per person in the gallery varies by application, and tests for a specific problem would reflect this). The images in the probe set were divided into a number of categories, where each category tests different aspects of an algorithm. Table 2 gives a breakdown of the gallery and probe set by category for the Aug94 FERET large-gallery and false-alarm tests, and for

Table 1
Dates tests were administered

Type of test	Algorithm tested	Date test administered
Aug94 FERET (large-gallery)	MIT, USC	August 1994
Aug94 FERET (false-alarm)	MIT, USC	August 1994
	Rockefeller	November 1995
Mar95 FERET (large-gallery)	MIT, USC	March 1995
	Rockefeller	November 1995
	MIT	August 1996

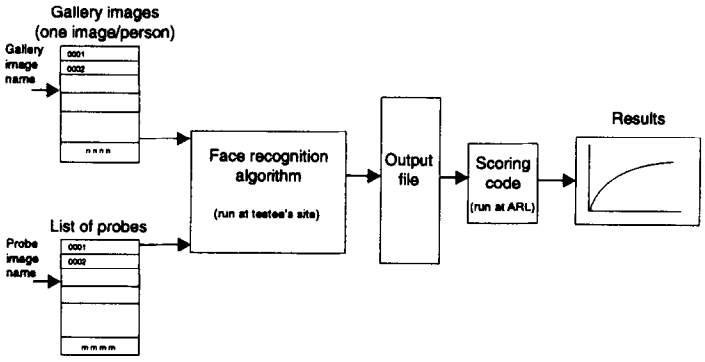

Fig. 5. Schematic diagram of the FERET testing procedure.

the Mar95 FERET test. Performance is broken out by probe category and is reported as cumulative match versus rank.

Each set of facial images includes two frontal images, **fa** and **fb** (see Fig. 2). One of these images is allocated to the gallery and is referred to as the **FA** image. The other frontal image (not in the gallery) is called the **FB** image and is allocated to the probe set. In the Aug94 FERET test, all the **fa** images (see Section 4.2) were selected to be the **FA** images. In the Mar95 FERET test, the allocation was random, with a 50–50 chance of either image being selected as the **FA** image. In the Aug94 FERET test, 316 **FB** images were in the probe set, and these images made up the largest category in the probe set. Tests on **FB** images allowed the comparison of algorithm performance on a large number of images and a preliminary assessment of the potential of face recognition in general. If algorithms could not perform well on the **FB** images, then there was little hope of acceptable performance on duplicate images.

The duplicate images tested the algorithms' performance on probe images taken on different dates from those of the corresponding images in the gallery. Duplicate images most closely resemble the type of images expected in real-world applications. The number of duplicates in the Aug94 FERET test was small because, in that test, the priority was placed on collecting a medium-size database for baselining algorithm performance. For diagnostic purposes, in both tests, we placed copies of **FA** images in the probe set; these should produce exact matches with their originals in the gallery.

The remaining categories of images allowed us to examine algorithm performance under different conditions. The number of images in each of these categories was small. The categories were designed to give an indication of how algorithms would respond to these variations. *Quarter* and *half rotations* refer to the head rotation in those images. The remaining categories consisted of the digitally altered frontal images discussed in Section 4.2.

Fig. 6 reports overall performance for the Aug94 test, 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 digitally altered images. Fig. 7 shows performance on the **FB** and duplicate categories of probes, revealing that duplicate categories of probes are harder to identify than **FB** probes.

Table 2
Number and type of images in FERET tests

Type of image	Number of images		
	Large-gallery test (Aug94 FERET test)	Large-gallery test (Mar95 FERET test)	False-alarm test (Aug FERET test)
Gallery:			
FA	317	831	25
Probes:			
FA	48	71	–
FB	316	780	25
Probes not in gallery	50	45	204
Duplicate frontal images	60	463	–
Quarter rotations	26	33	–
Half rotations	48	48	–
40% change in illumination	40	40	10
60% change in illumination	40	40	9
10% reduction in scale	40	40	19
20% reduction in scale	40	40	19
30% reduction in scale	40	40	–
Contrast reversed clothes	22	40	19
Size of probe set	770	1680	306

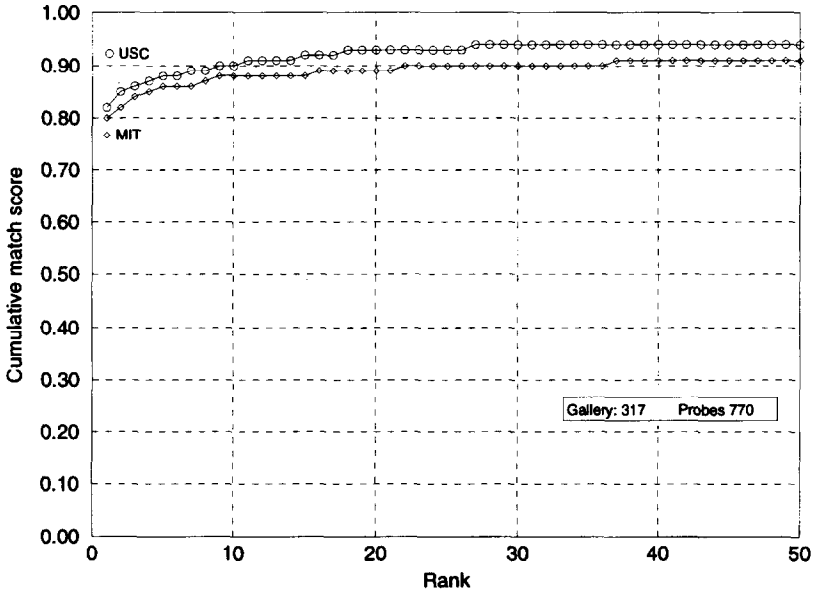

Fig. 6. Overall performance for the Aug94 FERET large-gallery test.

The Mar95 FERET test was an enlarged version of the Aug94 large-gallery test. The main differences were the size of the gallery (831 individuals) and the number of duplicate images in the probe set (463). Fig. 8 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 digitally altered images. Fig. 9 shows the performance on the **FB** and duplicate categories of probes. Fig. 10 shows performance on quarter rotated images.

After making substantial improvements in their algorithm, MIT requested to retake the Mar95 FERET test in August 1996. The result from this retest cannot be directly compared to previous tests, because additional images were released to researchers after they took the Mar95 FERET test. Thus, a large development set was available for the second test. Nevertheless, it is still noteworthy that MIT algorithm's performance improved between March 1995 and August 1996 (a comparison of Figs 9, and 11 shows a substantial increase in performance).

6.2. False-alarm test

The false-alarm test evaluates the ability of an algorithm to reject faces not in the gallery, while simultaneously correctly identifying those faces that are in the gallery. To

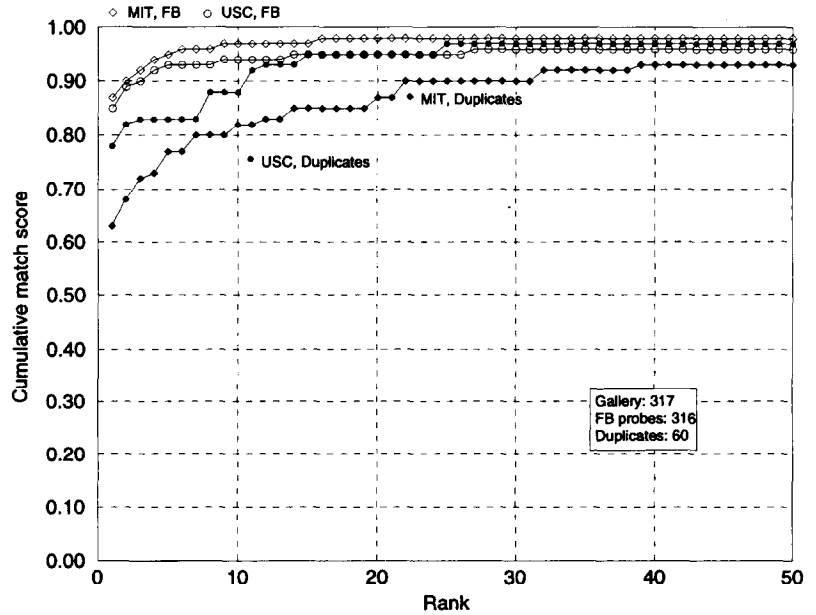

Fig. 7. **FB** and duplicate performance for the Aug94 FERET large-gallery test.

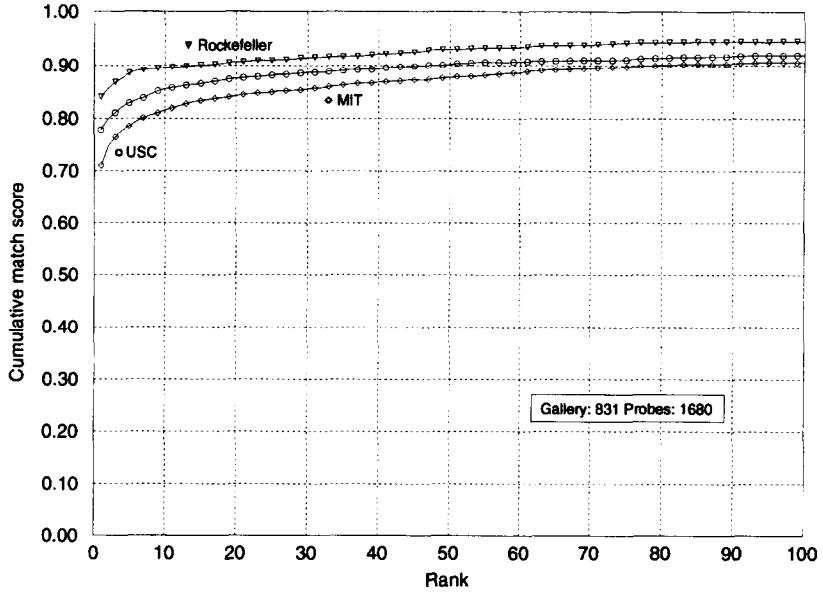

Fig. 8. Overall performance for the Mar95 FERET test.

accomplish this, we designed the false-alarm test with a small gallery and a large probe set, with the majority of the probes not being in the gallery. The gallery consisted of 25 people with one image per person, and the probe set had 305 images (Table 2). All images for this test were frontal images. The results are reported on the ROC in Fig. 12. For computing the results in Fig. 12, the number of probes not in the gallery is 204, and the number of probes in the gallery is 101 (the remaining probe categories).

7. Conclusions and lessons learned

The FERET database and evaluation effort has collected a database of facial images, which has been used for the development and evaluation of face-recognition algorithms. The evaluation procedure allowed us to assess the state of the art in face recognition, make comparisons between different approaches to face recognition, and identifying directions for future research. The FERET database is the largest generally available facial-image database, and the FERET evaluation is the only independent evaluation procedure. As a result of these accomplishments, the FERET database and evaluation procedures have become de facto standards.

The FERET database activities have succeeded in making a large database available to researchers. Because the time and expense of collecting a database are out of the reach of most researchers, the availability of the database has made research in face recognition possible for a much larger group. Results using the FERET database have been

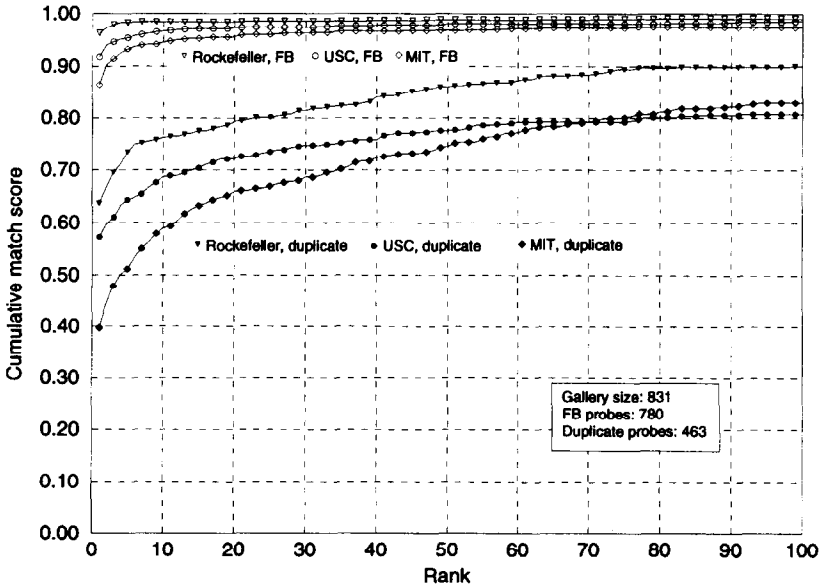

Fig. 9. **FB** and duplicate performance for the Mar95 FERET test.

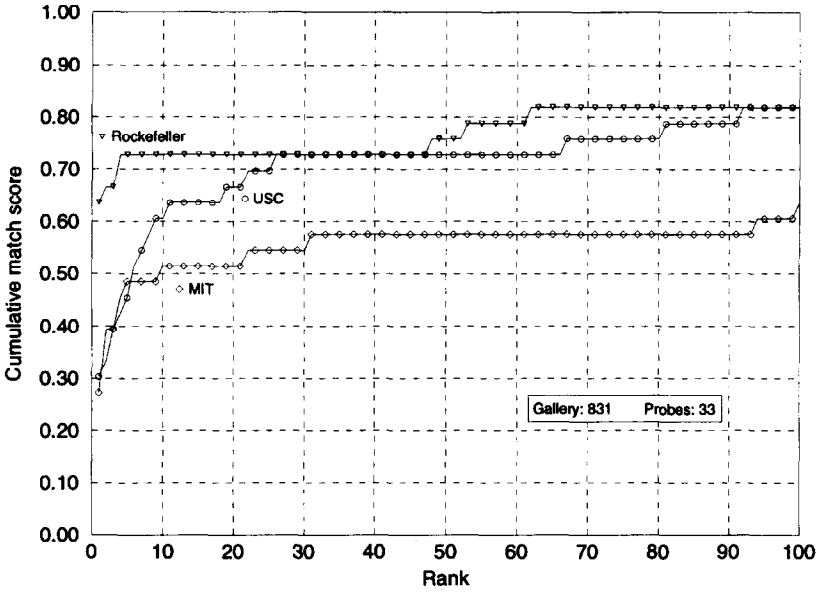

Fig. 10. Quarter rotation performance for the Mar95 FERET test.

reported in Cox et al. [22], Gordon [23], Gutta et al. [24], Mauer and von der Malsburg [20], Moghaddam and Pentland [12], Moghaddam et al. [25], Penev and Atick [17], Pentland et al. [14], Phillips and Vardi [26], Phillips [27], Swets and Weng [28], Wilder et al. [29], and Wiskott et al. [21]. These results enable a reader to get a feel for the strengths and maturity of the algorithms, although they do not allow for a direct comparison. (A direct comparison is not possible because each algorithm solves a slightly different problem, uses different training and testing sets, and computes the results differently).

The Aug94 FERET test provided a baseline for algorithm performance and a method of comparing different approaches to face recognition. The results of this test indicated that three main directions for future research were: (1) to increase the size of the database, (2) to recognize people in duplicate probes taken months or years apart from the corresponding gallery image, and (3) to handle variations in pose. Subsequent image collections were made to increase the size of the database and to support research on the second item.

The Mar95 FERET test was designed to measure progress in recognizing faces from duplicate images and in performance as the size of the gallery increases. Although it is not possible to directly compare the results of the two tests, examining the **FB** performance allows an indirect comparison. Fig. 13 shows that performance did not decrease even though the difficulty of the test increased (the size of

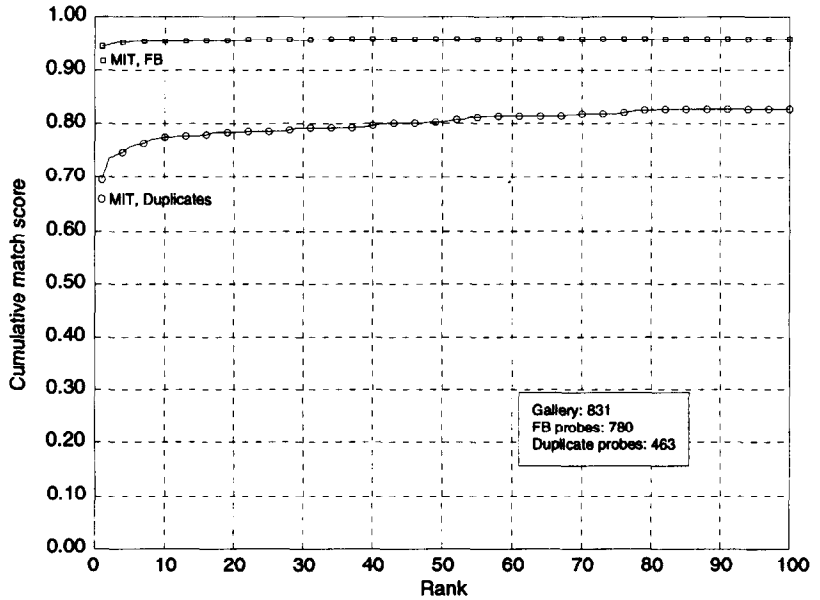

Fig. 11. MIT's performance on **FB** and duplicate probes for the test administered in August 1996.

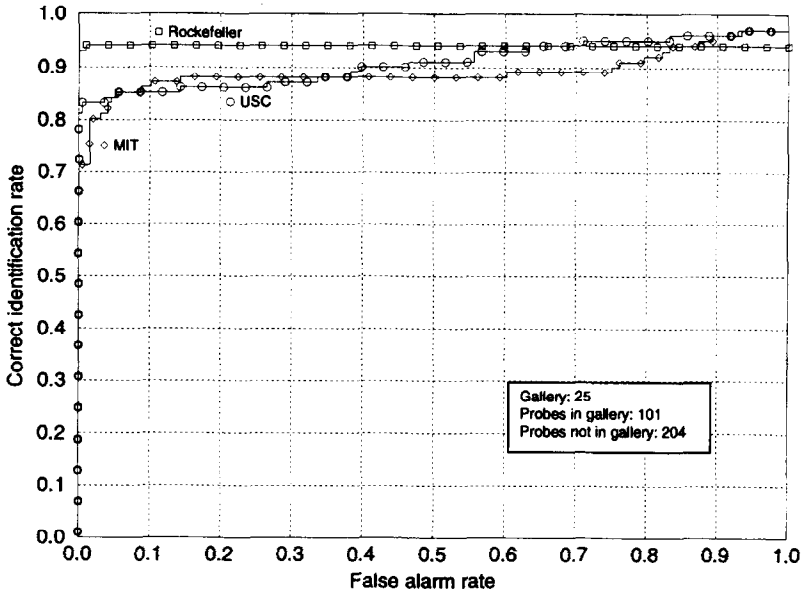

Fig. 12. Performance for the Aug94 FERET false-alarm test.

the gallery increased from 317 to 831). Performance on the duplicate images decreased (Figs 7, and 9). However, this result followed a two-fold increase in the difficulty of the problem: first, increased the number of duplicates from 60 to 463; second, increased the illumination variation because the number of sessions increased from 5 to 11 (the variations are due to moving and reassembling equipment). (This, more than time between duplicates, was the primary factor causing the increase in difficulty).

The FERET test is not the only possible test, but it is a standard test. It does not address many practical issues associated with fielding a system for a particular application; rather, it measures overall progress in the field. The suitability of an algorithm for an application must be determined by evaluation of its performance on images representative of that application.

The future direction of the FERET database effort is to increase the size of the database in terms of the number of both individuals and duplicates. The effort will be expanded to include infrared images [29] and video sequences. The most recent FERET test (which took place in September 1996) concentrated on large galleries and a greater number of duplicates in the probe set. This test is substantially different in design from the tests reported here (preliminary results are in Phillips et al. [30]). Beyond this test, the effort will develop protocols capable of testing algorithms that detect and identify faces from video streams.

From the authors' experience with the FERET project, we

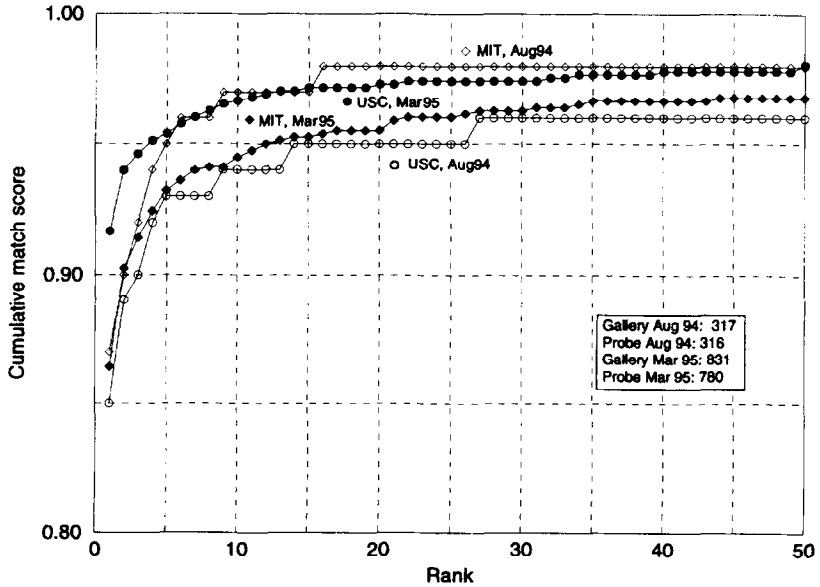

Fig. 13. Comparison of **FB** performance between the Aug94 FERET large-gallery test and the FERET Mar95 test.

have drawn two important though subjective conclusions. First, the FERET database and evaluation procedure have pushed the development of face recognition. By designing the protocols for image collection and testing, the FERET effort has advanced the state of the art. This is because the tests were sufficiently beyond the state of the art that researchers had to develop new algorithmic techniques.

Second, independent testing of face-recognition technology increases its credibility with potential users. When they are approached by researchers who all claim recognition rates of 99%, such users are skeptical. However, when presented with results from an independent test, they give automatic face recognition more serious consideration. Even if the test does not mirror their particular task, it can demonstrate that the algorithms are mature enough to be adapted to a given application.

Acknowledgements

The FERET program is sponsored by the U.S. Department of Defense's Counterdrug Technology Development Program Office. The U.S. Army Research Laboratory (ARL) is the technical agent for the FERET program. ARL designed, administered, and scored the FERET tests. George Mason University collected, processed, and maintained the FERET database. Inquiries regarding the FERET database or test should be directed to P. Jonathon Phillips.

References

[1] O. Firschein, M. Fischler, T. Kanade, Creating benchmarking problems in machine vision: scientific challenge problems, DARPA Image Understanding Workshop, 1993, pp. 177–182.

[2] R. Haralick, Computer vision theory: the lack of thereof, Computer Vision, Graphics, and Image Processing 36 (1986) 372–386.

[3] R. Jain, T. Binford, Ignorance, myopia, and naivete in computer vision systems, CVGIP: Image Understanding 53 (1) (1991) 112–117.

[4] T. Pavlidis, Why progress in machine vision is so low, Pattern Recognition Letters 13 (1992) 221–225.

[5] K. Price, Anything you can do, I can do better (no you can't), Computer Vision, Graphics, and Image Processing 36 (1986) 387–391.

[6] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, UNIPEN project of on-line data exchange and recognizer benchmarks, 12th Int. Conf. on Pattern Recognition, 1994, pp. B:29–31.

[7] M. Heath, S. Sarkar, T. Sanocki, K. Bowyer, Comparison of edge detectors: A methodology and initial study, Proceedings Computer Vision and Pattern Recognition 96 (1996) 143–148.

[8] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbob, R. Fisher, An experimental comparison of range image segementation algorithms, IEEE Trans. PAMI 18 (7) (1996) 673–689.

[9] G. Robertson, I. Craw, Testing face recognition systems, Image and Vision Computing Journal 12 (9) (1994) 609–614.

[10] P.J. Phillips, P. Rauss, S. Der, FERET (face recognition technology) recognition algorithm development and test report, Technical Report ARL-TR-995, U.S. Army Research Laboratory, 1996.

[11] P. Rauss, P.J. Phillips, A.T. DePersia, M. Hamilton, Face recognition technology program overview and results. 25th AIPR Workshop: Emerging Applications of Computer Vision, SPIE 2962 (1996) 253–263.

[12] B. Moghaddam, A. Pentland, Face recognition using view-based and modular eigenspaces. Proc. SPIE Conference on Automatic Systems for the Identification and Inspection of Humans, SPIE 2277 (1994) 12–21.

[13] B. Moghaddam, A. Pentland, Maximum likelihood detection of faces and hands, in: M. Bichsel (Ed.), International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 122–128.

[14] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, Proceedings Computer Vision and Pattern Recognition 94 (1994) 84–91.

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

[16] J. Atick, A.N. Redlich, Convergent algorithm for sensory receptive field development, Neural Computation 5 (1993) 45–60.

[17] P. Penev, J. Atick, Local feature analysis: a general statistical theory for object representation, Network: Computation in Neural Systems 7 (3) (1996) 477–500.

[18] A.N. Redlich, Redundancy reduction as a strategy for unsupervised learning, Neural Computation 5 (1993) 289–304.

[19] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, W. Konen, Distortion invariant object recognition in the dynamic link architecture, IEEE Trans. on Computers 42 (1993) 300–311.

[20] T. Maurer, C. von der Malsburg, Single-view based recognition of faces rotated in depth, in: M. Bichsel (Ed.), International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 248–253.

[21] L. Wiskott, J.-M. Fellous, N. Kruger, C. von der Malsburg, Face recognition and gender determination, in: M. Bichsel (Ed.), International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 92–97.

[22] I. Cox, J. Ghosen, P. Yianilos, Feature-based face recognition using mixture-distance, Proceedings Computer Vision and Pattern Recognition 96 (1996) 209–216.

[23] G.G. Gordon, Face recognition from frontal and profile views, in: M. Bichsel (Ed.), International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 47–52.

[24] S. Gutta, J. Huang, D. Singh, I. Shah, B. Takacs, H. Wechsler, Benchmark studies on face recognition, in: M. Bichsel (Ed.), International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 227–231.

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

[26] P.J. Phillips, Y. Vardi, Data driven methods in face recognition, in:; M. Bichsel (Ed.), International Workshop on Automatic Face and Gesture Recognition, 1995, pp. 65–70.

[27] P.J. Phillips, Representation and Registration in Face Recognition and Medical Imaging, Ph.D. thesis, RUTCOR, Rutgers University, 1996.

[28] D. Swets, J. Weng, Discriminant analysis and eigenspace partition tree for face and object recognition from views, 2nd International Conference on Automatic Face and Gesture Recognition, 1996, pp. 192–197.

[29] J. Wilder, P.J. Phillips, C. Jiang, S. Wiener, Comparison of visible and infrared imagery for face recognition, 2nd International Conference on Automatic Face and Gesture Recognition, 1996, pp. 182–187.

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