To appear in the proceedings of the First International Conference on Audio and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, 12-14 March 1997.

The FERET September 1996 Database and Evaluation Procedure

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

[1] U.S. Army Research Laboratory, AMSRL-SE-SE,
2800 Powder Mill Rd., Adelphi, MD 20783-1197
[2] Department of Electrical and Computer Engineering
State University of New York at Buffalo, Amherst, NY 14260
[3] Department of Applied Sciences
College of Staten Island of City University of New York, Staten Island, NY 10314

Abstract. Two of the most critical requirements in support of producing reliable face-recognition systems are a large database of facial images and a testing procedure to evaluate systems. The *Face Recognition Technology (FERET)* program has addressed both issues through the FERET database of facial images and the establishment of the FERET tests. In this paper, we report on the FERET database and the September 1996 FERET test. This test is the third in a series of supervised face-recognition test administered under the FERET program.

1 Introduction

In September 1996, the Face Recognition Technology (FERET) program administered third in a series of tests of the performance of face recognition algorithms. The results of this test, which was based on the images in the FERET database, are reported here.

The FERET database and evaluation procedure are designed to advance the state of the art in face recognition, with the images collected directly supporting both the development and the evaluation of algorithms. The database is divided into a development set, which is provided to researchers, and a set of sequestered images for testing. The FERET database has made it possible for researchers to develop algorithms on a common database and to report results in the literature using this database.

A comprehensive method for evaluating face-recognition algorithms was also developed as part of the FERET program.[1] The independently administered FERET evaluation methodology allows for a scientific assessment of the relative strengths and weaknesses of different approaches.

[*] Please direct correspondence to P. Jonathon Phillips; email jonathon@arl.mil
[†] This work was done while H. Moon and S. Rizvi were visiting researchers at ARL.
[1] The Face Recognition Technology (FERET) program is sponsored by the U.S. Department of Defense's Counterdrug Technology Transfer Program with the U.S. Army Research Laboratory (ARL) serving as the technical agent.

The FERET tests clarify the current state of the art in face recognition and point out general directions for future research. Through this type of assessment, the community learns in an unbiased and open manner of the important technical problems to be addressed and how the community is progressing toward solving these problems.

2 FERET Testing History

The first FERET tests took place in August 1994 and March 1995 (for details see Phillips *et al.* [8]); FERET database collection had begun in September 1993 along with the FERET program [8, 10].

The August 1994 test established a performance baseline for face-recognition algorithms. This test was designed to measure performance on algorithms that could automatically locate, normalize, and identify faces from a database. The test uses both *gallery* and *probe* images. The *gallery* is the set of known individuals. An image of an unknown face presented to the algorithm is called a *probe*, and the collection of probes is called the *probe set*. The August 1994 test indicated three main directions of future research: (1) increasing the size of the database, (2) recognizing people from images collected months or years apart, and (3) correcting for variations in pose.

The second FERET test, which took place in March 1995, measured progress toward the above objectives. One emphasis of the test was on duplicate images in the probe set. A *duplicate* is defined as an image of a person whose gallery image was taken on a different date.

Building on the results of the March 1995 test, the FERET program established the main goals of the September 1996 test: to measure progress in face recognition since the March 1995 test, to assess current approaches, to identify the strengths and weaknesses of algorithms, and to point out future directions of research.

3 The FERET database

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 8-bit gray scale images. 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 of the image was taken, and special variations. The identity number was keyed to the person photographed, so that any future images collected on that person would have the same ID number associated with the images.

To maintain a degree of consistency throughout the database, the photographer used the same physical setup in each photography session. However, because the equipment must be reassembled for each session, there was variation from session to session (fig. 1).

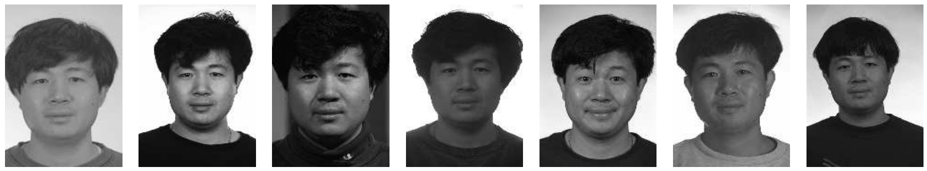

Fig. 1. Duplicate images (examples of variations among collections).

Images of an individual were acquired in sets of 5 to 11 images, collected under relatively unconstrained conditions. Two frontal views were taken (**fa** and **fb**); a different facial expression was requested for the second frontal image. For 200 sets of images, a third frontal image was taken with a different camera and different lighting (this is referred to as the **fc** image). The remaining images were collected at six head aspects: right and left profile, right and left quarter profile, and right and left half profile.

By July 1996, 1564 sets of images were in the database, for 14,126 total images. The database contains 1199 individuals and 365 duplicate sets of images. For some people, over two years elapsed between their first and most recent sittings, with some subjects being photographed multiple times (fig. 1). Approximately 500 sets of images were released to researchers as a developmental data set, and the remaining images were sequestered by the government.

For the evaluation procedure to produce meaningful results, the FERET database was established as a common database of images for development and testing. The FERET database has fulfilled the data requirements for both development and testing, becoming the de facto standard for face recognition from still images. (Results on the FERET database have been reported by Gordon [3], Mauer and von der Malsburg [4], Moghaddam*et al.* [5], Pentland *et al.* [7], Phillips and Vardi [9], and Wiskott *et al.* [13].)

4 Performance Evaluation Procedure

The traditional method of testing a face-recognition algorithm is to provide the algorithm with two sets of images, the gallery and the probe set, which do not intersect [8]. Unfortunately, this method severely limits one's ability to analyze the data. To overcome this deficiency, we modified the testing protocol to allow for a more detailed analysis of face-recognition algorithm performance. We designed the testing protocol so that algorithm performance can be computed for a variety of different galleries and probe sets.

In the new protocol, an algorithm is given two sets of images: the *target set* and the *query set*. We introduce this terminology to distinguish these sets from the gallery and probe sets that are used in computing performance statistics. The target set is given to the algorithm as the set of known facial images. The images in the query set are the unknown facial images to be identified. For each image q_i in the query set $\mathcal{Q}$, an algorithm reports the similarity $s_i(k)$ between q_i and each image t_k in the target set $\mathcal{T}$.

We compute algorithm performance from virtual galleries and probe sets. A virtual gallery $\mathcal{G}$ is a proper subset of the target set $\mathcal{T}$. Similarly, a virtual probe set $\mathcal{P}$ is a subset of the query set $\mathcal{Q}$. For a given virtual gallery $\mathcal{G}$ and virtual probe set $\mathcal{P}$, the performance scores are computed by examination of the similarity measures $s_i(k)$ such that $q_i \in \mathcal{P}$ and $t_k \in \mathcal{G}$.

The virtual gallery and probe set technique allows us to characterize algorithm performance by different categories of images. The different categories include (1) rotated images, (2) duplicates taken within a week of the gallery image, (3) duplicates where the time between the images is at least one year, (4) galleries containing one image per person, and (5) galleries containing more than one image per person. We can create a gallery of 100 people and estimate the algorithm's performance at recognizing people in this gallery. Using this as a starting point, we can then create virtual galleries of $200, 300, \ldots, 1000$ people and determine how performance changes as the size of the gallery increases. Another avenue of investigation is to create n different galleries of size 100 and calculate the variation in algorithm performance with the different galleries of size 100.

In the September 1996 FERET test, the target set contained 3323 images and the query set 3816 images. All the images in the target set were frontal images. The query set consisted of all the images in the target set plus rotated images and digitally modified images. We designed the digitally modified images to test the effects of illumination and scale [8]. For each query image q_i, an algorithm outputs the similarity measure $s_i(k)$ for all images t_k in the target set. The output for each query image q_i is sorted by the similarity scores $s_i(\cdot)$. Except for the rotated and digitally modified images, the target and query sets are the same. Thus, the test output contains every target image matched with itself.

In the previous FERET test, the algorithms had to be fully automatic: i.e, the algorithms had to automatically locate and recognize the face. Because of this requirement, it was not possible to separate the performance of the face-locating portion of the algorithm from the face-identification portion. In addition, this requirement restricted the algorithms that could be tested.

To be able to characterize the performance of the feature location and identification modules, we devised two versions of the September 1996 test. The target and query sets are the same for each version. The first version requires that the algorithms be fully automatic. In the second version of the test, the coordinates of the eyes are provided. By comparing the performance between the two versions, one can estimate the performance of the face-locating and identifying portions of an algorithm.

5 September 1996 Test Results

For the September 1996 test, we report the results for seven algorithms. Results are restricted to the semi-automatic case with eye-coordinates given. (MIT Media

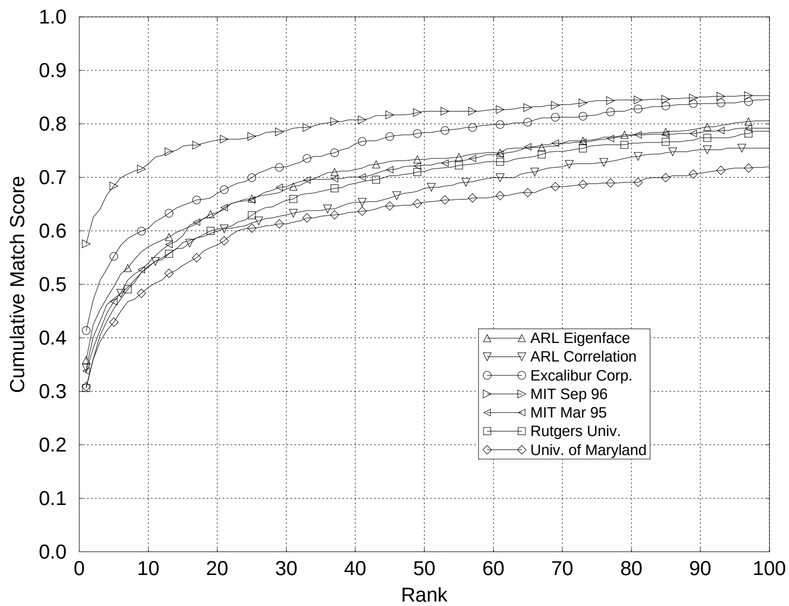

Fig. 2. Performance of individual algorithms on duplicate probes. (Number of images in gallery = 1196; Number of duplicate images in probe set = 722.)

Lab was the only group to take fully automatic test.) Two of these were developed at the MIT Media Laboratory. The first was the same algorithm that took the March 1995 test [6, 7]. This algorithm was retested so that any improvement since March 1995 could be determined. The second algorithm was based on more recent work [5]. Algorithms were also tested from Excalibur Corp., Rutgers University [12], and the University of Maryland [2]. The final two algorithms were our implementation of Turk and Pentland's eigenfaces [11] and normalized correlation. These two algorithms provide a baseline for algorithm performance.

In our implementation of eigenfaces, we measured the similarity between two faces by the angle between feature vectors in eigenspace. We found that the performance for this similarity measure is better than the Euclidean distance. Our correlation algorithm compared faces by correlating the whole face, where each face had been normalized and masked. This method yielded better performance than the feature-based correlation of Brunelli and Poggio [1].

We present results for algorithm performance on different categories of probe sets and the effects of different galleries. The results presented are a small subset of the actual ongoing analysis to be completed. We report on two sets of experiments. The first set summarizes algorithm performance on large galleries

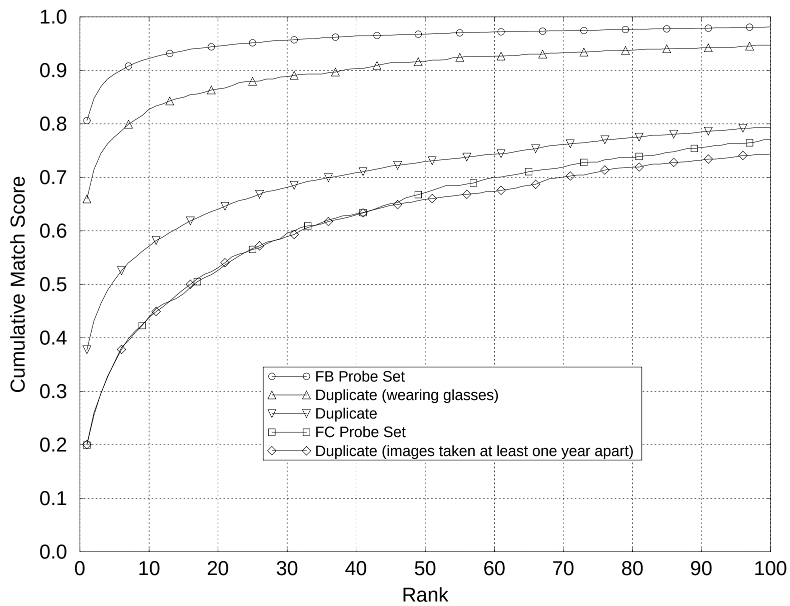

Fig. 3. Comparison of performance against different categories of probes. (Gallery consisted of 1196 frontal image and probe set contains 722 frontal images.)

for a number of different categories of probe sets. The second set shows how performance varies with different galleries.

For a probe, the algorithms outputs the entire gallery ordered by the similarity score. Because we are using a closed universe, the correct identity of the probe will appear in the list and has a rank k. For a probe set $\mathcal{P}$, we calculate the percentage of probes that have rank n or less, for $n = 1, ..., 100$. Performance statistics are reported as a cumulative match score graph [8], where the horizontal axis is the rank. The vertical axis is the percentage of probes in a probe set that have rank n or less.

In the large gallery experiments, the gallery consisted of 1196 frontal images with one image per person. None of the people in the gallery were wearing glasses. We computed performance of the large gallery against different categories of probes. In the first category, the probes set contained 722 duplicate images. Figure 2 shows the performance on this category for the seven algorithms that we tested.

In figure 3 we compare the difficulty of different categories of probe sets. Whereas figure 2 reports the performance for each algorithm, figure 3 shows a single curve that is an average of the performance of all the algorithms. For

Algorithm	gallery 1	gallery 2	gallery 3	gallery 4	gallery 5
ARL Eigenface	6	6	3	2	5
ARL Correlation	7	4	4	4	6
Excalibur Corp.	2	3	2	3	1
MIT Sep 96	1	1	1	1	1
MIT Mar 95	4	2	5	6	7
Rutgers Univ.	3	4	7	5	4
Univ. of Maryland	4	6	6	7	1
Average	0.220	0.587	0.626	0.512	0.653
Number of Probes Scored	143	64	194	277	44

Table 1. Variations in performance over 5 different galleries of fixed size (200) on duplicate probes. Algorithms are ordered by performance (1 to 7). The order is by percentage of probes correctly identified (rank 1). Also included in the table is average rank 1 performance for all algorithms and number of probes scored.

example, we compute the first rank score for the duplicate probe set by averaging the first rank score for all algorithms in figure 2. Figure 3 reports performance according to five different categories of images. The first category consists of the second frontal images (**fb**) from the same set as the gallery image (section 3). The second category consists of images of people wearing glasses. The third category is the set of all duplicate images (figure 2). The fourth category consists of the frontal **fc** images (images taken on the same day, but with a different camera and lighting). The final category consists of images that were taken at least a year and a half apart. (The gallery in this category is a subset of the gallery in the other categories.)

In table 1, we show how algorithm performance varies if the people in the galleries change. For this experiment, we constructed five galleries of approximately 200 individuals, in which each individual was in only one gallery. Table 1 orders each algorithm by performance for each of the five galleries.

6 Discussion and Conclusion

We draw three major conclusions from the September 1996 FERET test. The first is that substantial progress has been made in face recognition. This is directly supported by the improvement in performance of the algorithms developed at the MIT Media Lab and is indirectly supported by the number of groups that took this test.

The second is that the performance of six of the seven algorithms can be grouped together. Except for new MIT algorithm, no algorithm is clearly better than another algorithm. (In table 1, the University of Maryland's performance ranked first for some galleries and seventh for others.)

The final conclusion is that algorithm performance is dependent on the gallery and probe sets. Figure 3 shows categories of probe sets characterized by difficulty. Images from the same set (**fa** versus **fb**) are the least difficult to recognize, and images taken a year or more apart are the most difficult. Performance on the **fc** images shows the effects of changes in illumination. The effect of changing the gallery is shown in table 1.

This test shows that definite progress is being made in face recognition and that the upper bound in performance has not been reached. The results show that work needs to be done in recognizing faces from duplicate images and handling variations due to illumination. Work also needs to be done on understanding how changing the gallery affects algorithm performance.

References

1. R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Trans. PAMI*, 15(10), 1993.
2. K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. In *ICASSP '96*, pages 2148–2151, 1996.
3. G. G. Gordon. Face recognition from frontal and profile views. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 47–52, 1995.
4. T. Maurer and C. von der Malsburg. Single-view based recognition of faces rotated in depth. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 248–253, 1995.
5. B. Moghaddam, C. Nastar, and A. Pentland. Bayesian face recognition using deformable intemsity surfaces. In *Proceedings Computer Vision and Pattern Recognition 96*, pages 638–645, 1996.
6. B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *Proceedings of the Inter. Conf. on Computer Vision*, pages 786–793, 1995.
7. A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eignspaces for face recognition. In *Proceedings Computer Vision and Pattern Recognition 94*, 1994.
8. P. J. Phillips, P. Rauss, and S. Der. FERET (face recognition technology) recognition algorithm development and test report. Technical Report TR # 995, U.S. Army Research Laboratory, 1996.
9. P. J. Phillips and Y. Vardi. Data driven methods in face recognition. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 65–70, 1995.
10. P. Rauss, P. J. Phillips, A. T. DePersia, and M. Hamilton. Face recognition technology program overview and results. In *AIPR-96*, October 1996.
11. M. Turk and A. Pentland. Eigenfaces for recognition. *J. Neuroscience*, 3(1), 1991.
12. J. Wilder. face recognition using transform coding of gray scale projection projections and the neural tree network. In *Artifical neural networks with applications in speech and vision*. Chapman Hall, 1994.
13. L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition and gender determination. In M. Bichsel, editor, *International Workshop on Automatic Face and Gesture Recognition*, pages 92–97, 1995.