Assessment of Time Dependency in Face Recognition: An Initial Study

Patrick J. Flynn[1], Kevin W. Bowyer[1], and P. Jonathon Phillips[2]

[1] Dept of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556 USA
{flynn,kwb}@nd.edu
[2] National Institute of Standards and Technology
100 Bureau Dr., Stop 8940, Gaithersburg, MD 20899 USA
jonathon@nist.gov

Abstract. As face recognition research matures and products are deployed, the performance of such systems is being scrutinized by many constituencies. Performance factors of strong practical interest include the elapsed time between a subject's enrollment and subsequent acquisition of an unidentified face image, and the number of images of each subject available. In this paper, a long-term image acquisition project currently underway is described and data from the pilot study is examined. Experimental results suggest that (a) recognition performance is substantially poorer when unknown images are acquired on a different day from the enrolled images, (b) degradation in performance does *not* follow a simple predictable pattern with time between known and unknown image acquisition, and (c) performance figures quoted in the literature based on known and unknown image sets acquired on the same day may have little practical value.

1 Introduction

Although automatic face recognition has a rich history [4], it has only recently emerged as a potentially viable component of authentication and access control systems. The research community has responded to this increased interest in various ways. Firstly, the variety of approaches to face recognition continues to broaden (*e.g.*, new modalities [2][5][6]). Secondly, standardized approaches for assessment (and more importantly, comparison) have emerged [1] and been used in meaningful evaluations of vendor systems [7][8]. Such efforts require databases that are large enough and controlled well enough to admit meaningful statistical analyses, but also are representative of applications. The emergence of viable systems is shifting emphasis from development of new algorithmic techniques to gaining an understanding of the basic properties of face recognition systems. One of the least well-understood phenomena is variation of face appearance over short (weeks or months), medium (years), and long (decades) periods of time. This paper presents an

J. Kittler and M.S. Nixon (Eds.): AVBPA 2003, LNCS 2688, pp. 44-51, 2003.

ongoing collection effort to support understanding the short and medium term changes in face appearance. Initial analyses are measuring the effects of temporal variation on algorithm performance and estimating the distribution of matches of images of the same person. The database will be released to the research community to support development of algorithms that are robust to temporal variations. The control of location, camera, and lighting allows variations in face appearance due to elapsed time to be investigated without masking from environmental changes. It has been observed [7,8] that face recognition systems are challenged in uncontrolled lighting (*e.g.*, outdoors or uneven indoor ambient illumination). The database described here contains images with uncontrolled lighting that can be used as challenge data.

The *de facto* standard in the area of performance evaluation of face identification algorithms is the FERET methodology [1]. This methodology and most subsequent work employ the concept of a *training* image set used to develop the identification technique, a *gallery* image set that embodies the set of persons enrolled in the system, and a *probe* image set containing images to be identified. Identification of a probe image yields a ranked set of matches, with rank 1 being the best match. Results are presented as cumulative match characteristics (CMC), where the x-axis denotes a rank threshold and the y-axis is the fraction of experiments that yield a correct match at ranks equal to or lower than the threshold. General aspects of the FERET methodology include precise specification of training, gallery, and probe image sets drawn from a large database of face images, defined methods for computing performance metrics, and sequestration of test set images until after the test is performed. In FERET tests in March of 1997, two algorithms were able to achieve 95% or greater correct identification of the rank-one match, based on a gallery of 1196 neutral-expression face images and a probe set of 1195 alternative-expression face images of the same subjects taken on the same day (Figure 3 of [1]). These algorithms operated in partially automatic mode, meaning that the eye coordinate locations were manually identified and supplied to the algorithms. When a somewhat smaller probe set was used, containing normal-expression images taken in a different image acquisition session, all algorithms scored less than 60% correct on a rank-one match (Figure 4 of [1]). This dramatic performance difference clearly points to the importance of studying how face identification performance changes as probe images are acquired at varying lengths of time separation from the gallery images. The complete FERET database was assembled in 1996 and has 14,126 images from 1,199 subjects [1]. Face images taken for one subject at one image acquisition session typically include multiple standard lighting conditions and multiple facial expressions. Some subjects participated in multiple image acquisition sessions separated by as much as two years in time, but following subjects over time was not a specific focus of the FERET effort. Only a handful of subjects participated in as many as ten different sessions.

There are a number of face database efforts described in the literature. The XM2VTS database assembled at Surrey [10] contains 295 subjects with images taken at one-month intervals and has been used for face authentication (verification) research. The PIE database at Carnegie Mellon University [11] contains 41,368 images of 68 people collected over a three-month period, reflecting a large variety of

poses and lighting conditions. Other databases assembled recently include the AR [12] and Oulu databases [13].

2 Data Collection

To assess the performance of face recognition systems under large elapsed time, a suitable database of such imagery must be obtained. Such a collection will be assembled during 2002-2004. The acquisition plan has the following elements:

a. Weekly acquisitions of each subject.
b. At a minimum, four high-resolution color images (two facial expressions, two controlled studio lighting configurations) and two images with "unstructured" lighting will be taken of each subject.
c. Subjects will participate in the study as long as possible over the two-year period.

Fig. 1. Ten FERET-style images of one subject taken over a period of eleven weeks

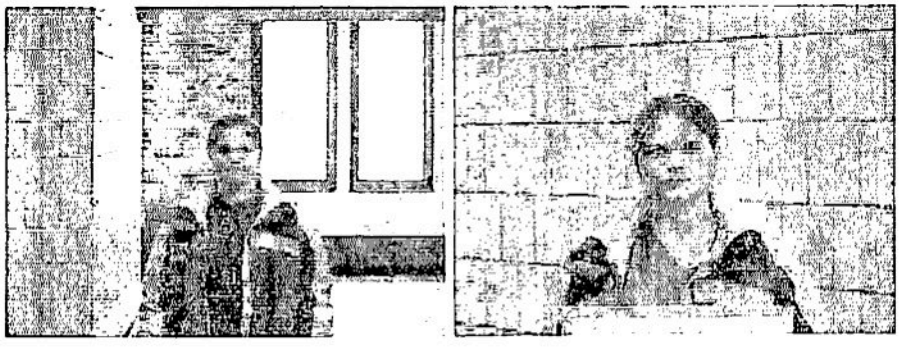

Fig. 2. Two representative "unstructured" images from the Spring 2002 pilot collection

In Spring 2002, a pilot acquisition study was undertaken to obtain experience with a large-scale study as well as to prompt the development of necessary hardware and software support. The Spring 2002 study involved ten acquisition sessions conducted over an eleven-week period (spring break occurred in the middle of the project). Color images were acquired with Sony MVC-95 cameras, which provided 1600x1200 images in JPEG format with minimal compression artifacts visible. Ten views of a single subject (one from each week of the pilot study) appear in Fig. 1. Three Smith-Victor A120 lights with Sylvania Photo-ECA bulbs provided studio lighting. The lights were located approximately eight feet in front of the subject; one was approximately four feet to the left, one was centrally located, and one was located four feet to the right. All three lights were trained on the subject face. One lighting configuration had the central light turned off and the others on. This will be referred to as "FERET style lighting" or "LF". The other configuration has all three lights on; this will be called "mugshot lighting" or "LM". In nine of the ten weeks of the Spring 2002 study, two additional images were obtained for each subject, under less well-controlled lighting and camera configuration. Generally, these images were taken in a hallway outside the laboratory, with a different camera and subject position each week. Fig. 2 shows two representative images of this "unstructured" type; the subject is the same as that in Fig. 1.

The Spring 2002 study yielded 3378 color images. Fig. 3 depicts the number of subjects participating in each week of the project. Eye coordinates in each image were selected manually and used for image registration in the PCA system described below.

3 Experimental Designs and Results

In this section, we describe a series of experiments designed to investigate the baseline performance of a face recognition system employing the data described in Section 2, and to investigate performance variations due to elapsed time. The software suite used in these experiments was developed at Colorado State University [3]. In keeping with the evaluation methodology introduced in the FERET study [1], each experiment is characterized by three image sets, all disjoint.

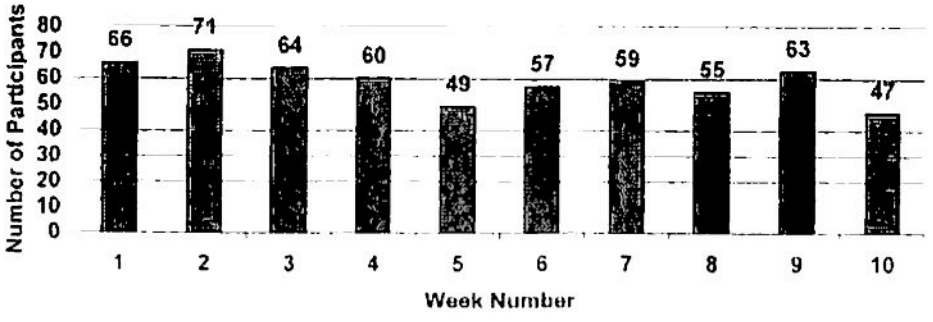

Fig. 3. Participation counts per week

a. The *training set* is used to form the projection matrix used in the PCA technique; in the experiments reported here, 1115 images drawn from the FERET data set (neutral facial expression) formed the training set. We experimented with other sets of training data and did not discover significant variations in performance.
b. The *gallery* set contains the set of "enrolled" images of subjects to recognize. All galleries used in this paper were drawn from the Spring 2002 image acquisitions described above.
c. The *probe* set is a set of images to be recognized via matching against the gallery. We employ a closed universe assumption (every probe will have a corresponding match in the gallery). Probes were drawn from the Spring 2002 database discussed above.

Gallery and probe selections were made to allow straightforward experimental comparisons akin to those typically reported in the literature. For the sake of brevity, we report four such experiments in this paper. However, the availability of several weeks of data on the same subjects allows for other studies explicitly addressing time dependence.

3.1 Experiment 1

The scenario for this experiment is a typical enroll-once identification setup. The 62 gallery images were neutral-expression, LF images of all subjects photographed in session 1 of the Spring 2002 study. The 438 probe images were all neutral-expression, LF images of subjects in sessions 2 through 10 of the Spring 2002 study. Hence, this experiment controls for same lighting and type of expression. For each subject, there is one enrolled gallery image and up to nine probe images, each acquired in a distinct later session. Figure 4 shows a cumulative match characteristic (CMC) plot for this recognition study. Since the Colorado State PCA software supports several metrics for subspace matching, we experimented with several choices. This plot illustrates a striking difference in performance between the straightforward Euclidean metric and the Mahalanobis metric that attempts to factor out scaling and correlation effects. The approximately 95% first-rank recognition result using the Mahalanobis angle metric is an encouraging result. The quality of the Mahalanobis angle metric has been noted by other researchers [3,9].

3.2 Experiment 2

This experiment controls for different expressions in the gallery and probe sets. The gallery used was identical to that of Experiment 1. The 438 probe images were *alternate-expression* FERET-lit subjects. CMC curves are depicted in Figure 5. As expected, performance in this experiment degrades significantly in comparison to Experiment 1, when the same expression is used in gallery and probe. Note that the performance degradation of the Euclidean metric is more than twice that of the Mahalanobis angle metric. This suggests that the Mahalanobis angle metric is effectively normalizing out some of the variation due to the change in expression. However, the performance degradation of more than 10% for the Mahalanobis angle metric indicates that more needs to be done to handle variation in expression.

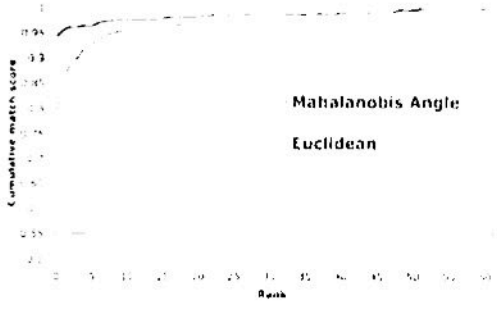

Fig. 4. CMC plot for experiment 1

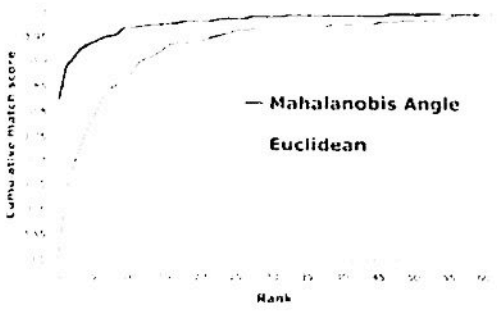

Fig. 5. CMC plot for Experiment 2

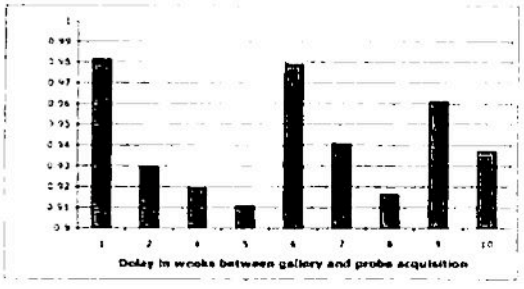

Fig. 6. Rank-1 correct match percentage for ten different delays between gallery and probe acquisition

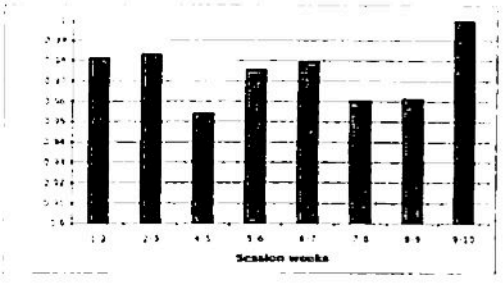

Fig. 7. Rank-1 correct match percentage for eight experiments with gallery and probe separated by one week

3.3 Experiment 3

Experiment 3 was designed to reveal any obvious effect of elapsed time (between gallery and probe acquisition) on performance. The experiment consisted of nine sub-experiments. The gallery set is the same as that used in Experiments 1 and 2. Each of the probes was a set of neutral-expression, FERET-lit images taken within a single session after session 1 (*i.e.*, subexperiment 1 used session 2 images in its probes, subexperiment 2 used session 3, and so forth). Figure 6 plots, for each week, the percentage of top-ranked matches that were correct in the nine sub-experiments (the graph depicts performance between 90% and 100%). The graph reveals differences in performance from week to week, but there is no clearly discernable trend in the results. Week 5 has the worst results of the ten weeks and week 6 is essentially the same as week 1 in performance.

3.4 Experiment 4

Experiment 4 was designed to examine the performance of the face recognition system with a constant delay of one week between gallery and probe acquisitions. It consists of nine sub-experiments: the first used images from session 1 as a gallery and session 2 as probe, the second used session 2 as gallery and session 3 as probe, and so on. All images were neutral-expression subjects with FERET-style lighting. The top-

rank-correct percentages for this batch of experiments appear in Figure 7. We note an overall higher level of performance with one week of delay than with delays larger than one week (as plotted in Figure 6). However, there is no clear trend in performance with an increasing number of weeks between gallery and probe acquisition.

4 Conclusions and Future Work

In this paper, we have described a new data set for face recognition research containing images of several dozen subjects taken weekly over a ten-week interval. Another such collection effort with expanded scope is underway. We also describe the results of some baseline performance experiments using the collected data. We observed superior performance of the Mahalanobis angle metric over the Euclidean metric. Any delay between acquisition of gallery images and probes caused recognition system performance degradation. More than one week's delay yielded poorer performance than a single week's delay. However, there is no discernible trend (using the data in the pilot study) that relates the size of the delay to the performance decrease. This motivates the development of a larger database covering more subjects and a longer period of time. Such an acquisition is underway. Additional experiments to be performed include: identification of subjects who are consistently difficult to recognize correctly, determination of the effect of lighting change (from FERET lighting to another structured lighting scheme or to unstructured lighting) on performance, investigation of metrics for matching, and repetition of earlier experiments with new image data.

Acknowledgement and Disclaimer

This research was supported by the Defense Advanced Research Project Agency (DARPA) under AFOSR award F49620-00-1-0388 and ONR award N00014-02-1-0410. Commercial equipment is identified in this work in order to adequately specify or describe the subject matter. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment identified is necessarily the best available for this purpose.

References

[1] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Trans. on PAMI* 20, 10 (Oct 2000), 1090-1104.

[2] A.J. O'Toole, T.Vetter, H. Volz, and E.M. Slater. Three dimensional caricatures of human heads: distinctiveness and the perception of facial age, *Perception* 26, 719-732.

[3] W.S. Yambor, B.A. Draper and J.R.Beveridge, Analyzing PCA-based Face Recognition Algorithms: Eigenvector Selection and Distance Measures, *Proc. 2nd Workshop on Empirical Evaluation in Computer Vision, Dublin*, Ireland, July 1, 2000.

[4] R. Chellappa, C.L. Wilson, and S. Sirohey, Human and Machine Recognition of Faces: A Survey, *Proc. IEEE* 83(5), 705-740, May 1995.

[5] D.A. Socolinsky, L.B. Wolff, J.D. Neuheisel, and C.K. Eveland, Illumination Invariant Face Recognition Using Thermal Infrared Imagery, Proc. CVPR 2001, vol. I, 527-534, December 2001.

[6] V. Blanz, S. Romdhani and T. Vetter, Face identification across different poses and illuminations with a 3D morphable model, *Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition*, 202-207, 2002.

[7] D.M. Blackburn, J.M. Bone and P.J.Phillips, FRVT 2000 results. http://www.frvt.org/FRVT2000.

[8] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, J. M. Bone, "Face Recognition Vendor Test 2002: Evaluation Report, NISTIR 6965, 2003, http://www.frvt.org.

[9] H. Moon and P.J. Phillips, Computational and performance aspects of PCA-based face recognition algorithms, *Perception* 30:303-321, 2001.

[10] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, Comparison of face verification results on the XM2VTS database, *Proc. ICPR 2000*, Barcelona, v. 4, p. 4858-4863, Sept. 2000.

[11] T. Sim, S. Baker, and M. Bsat, The CMU Pose, Illumination, and Expression (PIE) Database, *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-51, May 2002.

[12] AR Face database. http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html.

[13] E. Marszalec, B. Martinkauppi, M. Soriano, M. Pietikäinen (2000), A physics-based face database for color research, *Journal of Electronic Imaging* Vol. 9 No. 1 pp. 32-38.