This is NISTIR 6348, National Institute of Standards and Technology, June 1999.

http://www.nist.gov/itl/div894/894.03/pubs.html#face

Assessing Algorithms as Computational Models for Human Face Recognition

P. Jonathon Phillips
National Institute of Standards and Technology
Gaithersburg, MD 20899-9840
jonathon@nist.gov

Alice J. O'Toole
The University of Texas at Dallas
Richardson TX 75083-0688
otoole@utdallas.edu

Yi Cheng
The University of Texas at Dallas
Richardson TX 75083-0688
yicheng@utdallas.edu

Brendan Ross
The University of Texas at Dallas
Richardson TX 75083-0688

Heather A. Wild
The University of Texas at Dallas
Richardson TX 75083-0688

Abstract

We assessed the qualitative accord between several automatic face recognition algorithms and human perceivers. By comparing model- and human-generated measures of the similarity between pairs of faces, we were able to evaluate the suitability of the algorithms as models of human face recognition. Multidimensional scaling (MDS) was used to create a spatial representation of the subject response patterns. Next, the model response patterns were projected into the MDS space. The results revealed a bimodal structure for both the subjects and for most of the models, indicating that the qualitative performance of the subjects and models was related. The bimodal subject structure reflected strategy differences in making similarity decisions. For the models, the bimodal structure was related to combined aspects of the representations and the distance metrics used in the implementations.

1 Introduction

Understanding the principles and processes involved in recognizing faces has been an active area of research in the domains of neuroscience, psychology, and computer science. One goal of this research has been to develop computational algorithms capable of recognizing human faces. An excellent resource for comparing the performance of computational face recognition algorithms has been made available recently under the "FERET" program. In the FERET program, the US Government evaluated 18 state-of-the-art automatic face recognition algorithms between August 1994 and March 1997. The primary interest of the

US Government in the FERET project was to explore the potential of these algorithms for automating visual monitoring activities. Several comprehensive tests of the accuracy of the algorithms have been reported in detail elsewhere [7].

Although the FERET program provides a wealth of information about the relative strengths and weaknesses of individual algorithms for the problem of automatic face recognition, these algorithms have not been compared to the most flexible face recognition system currently available: the human perceiver. A thorough comparison between human and model performance is of practical value for anticipating qualitative changes in the errors made when replacing or supplementing human monitors with automated face recognition systems. Additionally, evaluating the accord between human and machine performance can give insight into the nature of human representations and recall processes for faces. Specifically, each computational algorithm instantiates a representational system for faces and implements a "distance" metric for comparing target faces with "stored" faces. The similarity between human and model performance can be taken as an indication of the suitability of the algorithms as models of human face recognition.

In the present study, we compared the performance of 13 of the 18 FERET-evaluated algorithms[1] with that of human subjects on a similarity-rating task. The perceptual similarity of a face to other faces in a reference category (e.g., young adult Caucasian males) is a measure of the face's "typicality". Psychological experiments have shown that face typicality is a robust predictor of face recognition accuracy[3]. This may indicate that human perception is structured by statistical properties of the faces encountered most commonly in one's everyday environment (e.g., the "other-race effect"[6]). Most currently available neural network/statistical algorithms of face recognition are likewise sensitive to the statistical structure of the system's "experience" with faces. For example, face recognition systems based on principal component analysis (PCA) represent faces with respect to feature axes derived from the statistical structure of the set of known faces. The implementation of an algorithm, however, varies with the nature of the facial encoding (i.e., representation) and with the distance metrics used to compare "query" or "target" faces with stored faces[5].

In this study, we measured the similarity between all possible pairs of a subset of faces in the FERET data base. We did this for 13 algorithms and 22 human subjects. We then derived measures of the typicality of each face for the algorithms and humans to determine the suitability of the algorithms as models of human face processing. The novelty of this work is that we (a) focus on *qualitative* rather than quantitative aspects of the accord between human and machine performance, and that we (b) concentrate on a set of highly similar or "confusable" faces. There are two reasons for concentrating on the qualitative aspects of performance in relating the human and model performance. First, for the FERET-evaluated algorithms, performance levels were very close to ceiling for the image set sizes used in our experiments (20 similar young adult Caucasian male faces).[2] Thus, the absolute level of performance is not useful for determining the suitability of the algorithms as models of human processes. Second, human performance levels can be varied arbitrarily by varying the parameters of a psychological experiment, e.g., shortening exposure time. These parameters are not relevant for assessing the performance of the algorithms. By contrast, qualitative measures, such as the pattern of similarities among faces, are computed at the level of individual faces and provide a very much under-exploited set of observations for evaluating the suitability of computational algorithms as models of human face recognition. In short, we are asking, "do the algorithms and human subjects find the same faces similar?"

[1] We restrict our attention to algorithms evaluated with the FERET Sep96 partially-automated test [7]. For unrelated technical reasons, the later of the UMD algorithms was not included in our study.

[2] With only 20 faces, subjects must make 400 similarity comparisons, which they found time-consuming and tedious.

This paper is organized as follows. First, we describe the psychological methods. Next, we give a brief description and classification of the subset of FERET-evaluated models considered in the present paper. Finally, we present a set of combined analyses aimed at determing the accord between human subjects and the algorithms.

2 Methods

2.1 Human Face Similarity Judgments

The purpose of the human experiments was to generate a measure of the perceived similarity between all possible pairs of faces for a selected subset of faces in the FERET data base. The stimuli consisted of a smiling and neutral expression picture of twenty Caucasian male faces in their twenties, chosen to be difficult for both the human subjects and the computational algorithms. The pictures were centered in a frame and clothing was edited out digitally to prevent matching the faces by clothing cues. Twenty-two students from The University of Texas at Dallas volunteered to participate in the experiment. On each trial, subjects viewed a pair of facial images consisting of a neutral and smiling expression face for 1 second. A computer prompt then asked the subject to rate the face pair as consisting of: (0) identical persons, (1) similar persons, or (2) dissimilar persons. Subjects viewed all possible pairs of faces, presented in random order, for a total of 400 trials.[3]

Subject Accuracy. Errors can be divided into (a) "misses", defined as "similar person" or "dissimilar person" responses to identical face pairs, i.e., a smiling and neutral expression version of the same person, and (b) "false alarms", defined as "identical person" responses to face pairs comprised of different individuals. Across all subjects, the proportion of miss errors was .09, and the proportion of false alarm errors was .04, indicating relatively high accuracy on the task.

Individual Subject Similarity Data. Next, we created a 20-by-20 perceptual similarity matrix for each subject. Each cell of this matrix, $S_i(j,k)$, contained the similarity rating (0, 1, or 2) given by the i^{th} subject to the j^{th} (neutral expression) and k^{th} (smiling expression) face pair. The result is a sort of "distance" matrix. Each value on the diagonal contains a zero if no "miss" errors were made for the face (the distance between identical faces is 0). The off-diagonal elements contained relatively higher numbers (perceptual distances) for relatively more dissimilar face pairs. This matrix differs from a standard distance matrix, however, because it is not symmetric, i.e., $S_{j,k} \neq S_{k,j}$, i.e., human subjects could judge the similarity between the neutral version of the j^{th} face and the smiling version of the k^{th} face to be different than the similarity between the neutral version of the k^{th} face and the smiling version of the j^{th} face. We will refer to these matrices as the *individual subject similarity matrices*.

Individual Subject Typicality Data. The perceived typicality of a face is a well-known and robust predictor of human recognition accuracy for the face [3]. Typicality can be measured as the overall perceived similarity of a face with respect to all other faces in a category. We calculated this by averaging the columns of the subject similarity data, giving a typicality vector, t_i, containing the average of the ratings each face received in combination with all other faces in the set. Thus, the j^{th} element of t_i contained the average of the 20 similarity ratings given by the i^{th} subject to the j^{th} face. Faces rated as similar to most other faces were considered "typical", whereas faces rated dissimilar to most other faces were considered "distinctive". We will refer to these vectors as the *individual subject typicality vectors*.

[3] All experimental events were controlled by a computer programmed with PsyScope[1].

Table 1: Computational Algorithms

ALGORITHM	REPRESENTATION	DISTANCE METRIC
Excalibur Co. (EX)	Unknown	Unknown
MIT95	PCA-based	L2
MIT96	PCA-difference space	MAP Bayesian Statistic
Michigan St. U. (MSU)	Fischer discriminant	L2
Rutgers U.	Greyscale projection	Weighted L1
U. of So. CA (USC)	Dynamic Link Architecture	Elastic matching
U. of MD (UMD97)	Fischer discriminant	L2
NIST (L1)	PCA	$\sum_{i=1}^{k} \lvert x_i - y_i \rvert$
NIST (L2)	PCA	$\sum_{i=1}^{k} (x_i - y_i)^2$
NIST (MD)	PCA	$-\sum_{i=1}^{k} x_i y_i z_i$
NIST (AN)	PCA	$-\frac{\sum_{i=1}^{k} x_i y_i z_i}{\sqrt{\sum_{i=1}^{k} x_i^2 \sum_{i=1}^{k} y_i^2}}$
NIST (ML1)	PCA	$\sum_{i=1}^{k} \lvert x_i - y_i \rvert z_i$
NIST (ML2)	PCA	$\sum_{i=1}^{k} (x_i - y_i)^2 z_i$

2.2 Computational Algorithms

The 13 computational algorithms we considered can be divided into two groups. The first group consisted of seven algorithms developed by researchers not involved in designing the FERET evaluation method. For this group, the FERET evaluation was an independent assessment of performance. These algorithms were developed at: Excalibur Corp. (EX), Michigan State University (MSU) [9], the Media Lab at the Massachutsetts Institute of Technology (MIT95, MIT96) [4], University of Maryland (UMD97) [9], University of Southern California (USC) [2] and Rutgers University (RUT) [8].

Six additional algorithms, implemented by researchers involved in designing the FERET evaluations, were included: (a) to provide a performance baseline control model using a standard PCA algorithm, and (b) to gain a better understanding of the impact of varying the "retrieval" stage of the model via variations of distance metric implemented in the nearest neighbor classifier. These NIST control algorithms are as follows: L1 distance (L1), L2 distance (L2), Mahalanobis distance (MD), city block distance metric (ML1), Euclidean distance (ML2), angular distance and cosine (AN). The overall accuracy of these algorithms is detailed elsewhere[5] and indicates that variations in the distance metric impact the performance of the PCA substantially.

The algorithms and their basic characteristics are listed in Table 1.[4] For present purposes, the computational performance measures for all algorithms were calculated using the FERET September 1996 evaluation method [7]. This method supplies a similarity measure between all possible pairs of images that is analogous in form to the human measures.

[4] $\{x_i\}$ and $\{y_i\}$ are points in a PCA-face space and $z_i \approx \frac{1}{\sqrt{\lambda_i}}$ and λ_i are the eigenvalues of the PCA.

2.3 Combined Analysis

We assessed the degree of accord between the model and human estimates of the similarity among the set of faces by creating a spatial representation of human subject response patterns to the 400 face pairs. We next projected analogous data from the computational algorithms into this subject similarity space. Algorithms producing patterns of similarity scores comparable to the human subjects should cluster close to the subjects in the space.

More formally, we analyzed the individual subject typicality vectors from the 22 human subjects using metric MDS[5]. This produces a low-dimensional spatial representation of the 22 subjects on the task. The first 2 axes of this analysis, which explain .73 and .08 of the variance, respectively, appear in Figure 1. Each individual subject is marked with an "*". The figure reveals two clusters of human subject response patterns captured by the first axis of the analysis. This indicates simply that subjects' response patterns with respect to this axis are bimodally distributed.

We next generated analogous typicality vectors for the algorithms and projected them into the space derived from the human subject data. These appear superimposed on the subject data in Figure 1. Each algorithm is marked with a ".", and is labeled with its abbreviated name. A few descriptive points are worth noting. First, the bimodal distribution of subject response patterns is mirrored in the model data. In one cluster of subjects, we find the algorithms from UMD97, USC, MSU and EX with the PCA control models employing the ML1, ML2 and L2 distance metrics. In the second cluster, we see the MIT96 algorithms and the PCA control models employing the AN, L1 and MD metrics. The MIT95 and RUT algorithms are not embedded in subject clusters. Second, all of the algorithms fall reasonably close to subject data, indicating at least some relationship between the human and model processing of the faces.

3 Interpretation and Discussion

The bimodal subject distribution indicates that subjects can be divided into two groups, possibly based on differences in the similarity criteria they applied to the task (e.g., some subjects might have considered hairstyle in their similarity judgments, whereas others may have used only the internal facial features). Although there is no way to assess this directly, one can apply an MDS to the "faces" (rather than subjects) separately for the two groups of subjects and then "lay out" pictures of the faces in the MDS space to attempt to interpret the axes. In fact, as part of our pilot work, we analyzed the original face similarity data for all subjects in this way and were able to interpret the first two axes of this analysis reasonably confidently. The first axis was "age/maturity", which contrasted "college student" faces at one end and with more mature "twenty-something" faces at the other end. The second axis, simply put, contrasted ruggedness/atheletes to studiousness/nerds. To interpret the bimodal split in the face typicality data, we performed this analysis separately for subjects on the left and right sides of Figure 1. This yielded a simple interpretation. For the left cluster of subjects, the age dimension dominated and for right cluster, the athletic dimension dominated.

Before linking these criteria to the model distribution, which mirrors the bimodal split seen for the subjects, we would like first to rule out a few obvious factors that might have contributed to the split. First, the extensive FERET tests assures us that performance accuracy differences are not responsible for the clustering. The three most accurate algorithms, MIT96, UMD97, and USC, are divided between the clusters. Second, somewhat surprising, the algorithms do not separate *exclusively* based on their underlying representation

[5] Metric multidimensional scaling is based on a linear distance metric and hence reduces to a standard PCA.

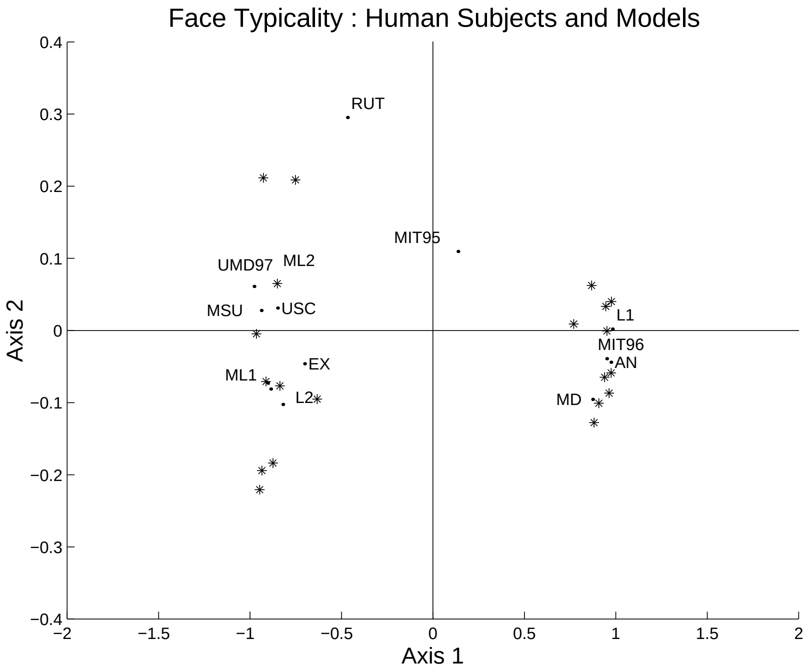

Figure 1: Two-dimensional MDS solution of the pattern of face similarities produced by human subjects. Subject are marked by "*", and algorithms' by labeled "·"'s.

bases. From Figure 1 it is clear that algorithms with rather different representations can cluster together (and vice versa). Although some clustering relates to representation (e.g., the UMD97 and MSU algorithms, both based on linear Fischer discriminant analysis, are very close), representation is apparently not the only factor.

Third, the distance metrics seemed to have a potent effect on the qualitative performance of the algorithms.[6] This is clear in the distribution of the NIST control implementations, which are scattered across the space. Given that only the distance metric varies in these models, distance by itself cannot explain the cluster structure. We are left then with the conclusion that the bimodal model structure is determined by complex trade-offs between the representations and distance metrics.

Returning to the subject similarity criteria, although such global criteria are "abstract", they are nonetheless likely to relate to physical properties of the shapes and textures of faces, e.g., athletic faces might be more muscular/masculine than faces we label as "studious". In fact, this kind of "global feature dimension" (e.g., masculinity) has been observed previously in alternative versions of some of these algorithms, applied to the task of classifying faces by sex. It is not surprising, therefore, that such dimensions relate to model predictions

[6] See [5] for a quantitative performance assessment of the distance metrics, which is consistent with the qualitative findings reported here.

of face similarity.

In summary, most of the automatic face recognition algorithms we tested performed in ways that are qualitatively similar to humans. Human performance is best characterized by the similarity criteria on which the subjects focus. Model behavior is best characterized by the complex interactions between the representations and distance metrics implemented.

Acknowledgments

This work was supported by grants from the National Institute of Justice to P. J. Phillips and A. J. O'Toole through NIST's Office of Law Enforcement Standards.

References

[1] J. D. Cohen, B. McWhinney, M. Flatt, and J. Provost. Psyscope : A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments and Computers*, 25:257–271, 1993.

[2] K. Okada et. al. The Bochum/USC face recognition system. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.

[3] L. Light, F. Kayra-Stuart, and S. Hollander. Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, 5:212–228, 1979.

[4] B. Moghaddam and A. Pentland. Beyond linear eigenspaces: Bayesian matching for face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.

[5] H. Moon and P. J. Phillips. Analysis of PCA-based face recognition algorithms. In K. W. Bowyer and P. J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision*. IEEE Computer Society Press, Los Alamitos, CA, 1998.

[6] A. J. O'Toole, K. A. Deffenbacher, D. Valentin, and H. Abdi. Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, 22:208–224, 1994.

[7] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.

[8] J. Wilder. Face recognition using transform coding of gray scale projections and the neural tree network. In R. J. Mammone, editor, *Artifical Neural Networks with Applications in Speech and Vision*, pages 520–536. Chapman Hall, 1994.

[9] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.