Face Recognition Algorithms Surpass Humans

Alice J. O'TOOLE, P. Jonathon PHILLIPS, Fang JIANG, Janet AYYAD, Nils PENARD,

and Hervé ABDI*

***Abstract*—There has been significant progress in improving the performance of computer-based face recognition algorithms over the last decade. Although algorithms have been tested and compared extensively with each other, there has been virtually no work comparing the accuracy of computer-based face recognition systems with humans. We compared seven state-of-the-art face recognition algorithms with humans on a face-matching task. Humans and algorithms determined whether pairs of face images, taken under different illumination conditions, were pictures of the same person or of different people. Three algorithms surpassed human performance matching face pairs prescreened to be "difficult" and six algorithms surpassed humans on "easy" face pairs. Although illumination variation continues to challenge face recognition algorithms, current algorithms compete favorably with humans. The superior performance of the best algorithms over humans, in light of the absolute performance levels of the algorithms, underscores the need to compare algorithms with the best current control—humans.**

***Index Terms*—face and gesture recognition, performance evaluation of algorithms and systems, human information processing**

I. INTRODUCTION

An increase in security concerns worldwide has focused public attention on the accuracy of computer-based face recognition systems for security applications. How accurate must a face recognition algorithm be to contribute to these applications? Over the last decade, academic computer vision researchers and commercial product developers have improved the performance of automated face recognition algorithms on a variety of

Manuscript received March 14, 2006. This work was supported by a contract to A. J. O'Toole and H. Abdi from TSWG. P. J. Phillips was supported in part by funding from the National Institute of Justice.
A. J. O'Toole is at the School of Behavioral and Brain Sciences, GR4.1 The University of Texas at Dallas Richardson, TX 75083-0688, USA, (e-mail: otoole@utdallas.edu).
P.J. Phillips is with the National Institute of Standards and Technology, 100 Bureau Dr., MS 8940 Gaithersburg MD 20899 jonathon@nist.gov
Fang Jinag is at the School of Behavioral and Brain Sciences, GR4.1 The University of Texas at Dallas Richardson, TX 75083-0688, USA, (e-mail: fxj018100@utdallas.edu).
Janet Ayyad is at the School of Behavioral and Brain Sciences, GR4.1 The University of Texas at Dallas Richardson, TX 75083-0688, USA.
Nils Pénard is at the School of Behavioral and Brain Sciences, GR4.1 The University of Texas at Dallas Richardson, TX 75083-0688, USA, (e-mail: npenard@utdallas.edu).
Hervé Abdi is at the School of Behavioral and Brain Sciences, GR4.1 The University of Texas at Dallas Richardson, TX 75083-0688, USA, (e-mail: herve@utdallas.edu).

challenging face recognition tasks. Information about the performance of automated face recognition systems is available to the public in the form of scholarly journal articles (for a review, see [1]), commercially disseminated product information, and United States Government sponsored evaluations (e.g., [2]-[5]). Virtually no information is available, however, about the accuracy of face recognition algorithms relative to humans. Because humans currently perform face recognition tasks in most real world security situations, it is unclear whether the use of algorithms improves security or puts it at greater risk.

Notwithstanding the issue of human performance, it is difficult even to make direct comparisons among algorithms in a way that gives an accurate idea of the current state of the art [6]. This is because most journal articles and product pamphlets consider only one or two systems at a time and report results that are based on facial image sets that vary widely in number and quality. United States Government-funded competitions allow for direct comparisons among multiple recognition algorithms. Performance is measured in these competitions using standardized evaluation procedures applied to large sets of facial images (e.g., FERET, [4]; Face Recognition Vendor Test 2000, [2]; Face Recognition Vendor Test 2002, [5]; Face Recognition Grand Challenge, FRGC, [3]).

The present work builds on the most recent of these large-scale tests of face recognition algorithms—the FRGC. The evaluation procedures used in this test provide a unique opportunity to benchmark human performance against current face recognition algorithms. This competition, initiated in 2004 and currently ongoing, is open to academic, industrial, and research lab competitors. The developers of computer-based face recognition systems volunteer their algorithms for evaluation on one or more face matching tasks varying in difficulty.

We carried out a direct comparison between humans and seven face recognition algorithms participating in the most difficult experiment of the FRGC: matching two-dimensional face images taken under different illumination conditions (Fig. 1). The problem of face recognition over changes in illumination is widely recognized to be difficult for humans [7]-[11] and for algorithms [12], [13]. Illumination changes can vary the overall magnitude of light intensity reflected back from an object, as well as the pattern of shading and shadows visible in an image [14]. Indeed, varying the illumination can result in larger image differences than varying either the identity [15] or the viewpoint [14] of a face. These illumination changes have well-documented, detrimental effects on human accuracy recognizing faces (for a review, see [16]).

For face recognition algorithms, the difficulty of the illumination problem is evident in comparing the FRGC experiments for matching controlled versus uncontrolled-illumination image pairs. In the controlled illumination experiment of the FRGC, algorithms match the identity of faces in two pictures taken under similar, controlled illumination conditions. The 17 algorithms participating in the controlled illumination experiment, to date, have achieved a median verification rate of .91 at the .001 false acceptance rate—where verification rate is the proportion of matched pairs correctly judged to be of the same person and false acceptance rate is the proportion of mismatched pairs judged incorrectly to be of the same person. By contrast, there are only seven algorithms competing in the uncontrolled illumination experiment. These algorithms have achieved a median verification rate of .42 at a false acceptance rate of .001. The difficulties posed by variable illumination conditions, therefore, remain a significant challenge for automatic face recognition systems.

In the FRGC, algorithms must determine if two faces in a pair of images are of the same person or of different people. Algorithms in the FRGC uncontrolled illumination experiment are required to match identities in approximately 128 million pairs of faces. Specifically, the test data for this experiment consist of all possible pairs of 16028 "target" faces and 8014 "probe" faces [13]. The *target* images were taken under controlled illumination conditions typical of those used for passport photographs and the *probe* images were taken under uncontrolled illumination conditions (e.g., in a corridor) (Fig. 1).

Participating algorithms produce a 16028 x 8014 matrix of *similarity scores* for all possible face pairs. A similarity score represents the end result of an algorithm's computations to establish the likelihood that the faces in the image pair are the "same" person. A higher similarity score indicates a greater likelihood that the two faces are the same person. The similarity score matrix is used to compute the algorithms' "same person" versus "different person" judgments. Face pairs with similarity scores greater than or equal to a set criterion *c* are judged to be of the "same" person; pairs with similarity less than or equal to *c* are judged to be of "different" people. By varying the criterion *c* over the full range of similarity scores, a complete receiver operating characteristic (ROC) curve is generated. An ROC curve indicates performance level for all possible combinations of correct verification and false acceptance rate. The use of an ROC curve for characterizing the performance of algorithms makes it relatively straightforward to compare human face matching data to the algorithm data.

Testing humans on 128 million pairs of facial images is impossible. We focused, therefore, on a sample of the "easiest" and "most difficult" face pairs with a sampling

procedure defined by a baseline algorithm. In three experiments, we tested human performance on "easy" and "difficult" face pairs, varying the exposure time of the face pairs from unlimited time to 500ms. Humans rated the likelihood that the two images were of the same person. Next, we extracted similarity scores from the seven algorithms on the same set of face pairs matched by the human participants. Finally, ROC curves were generated for the algorithms and humans. These form the primary basis of our comparisons between humans and machines on the face-matching task.

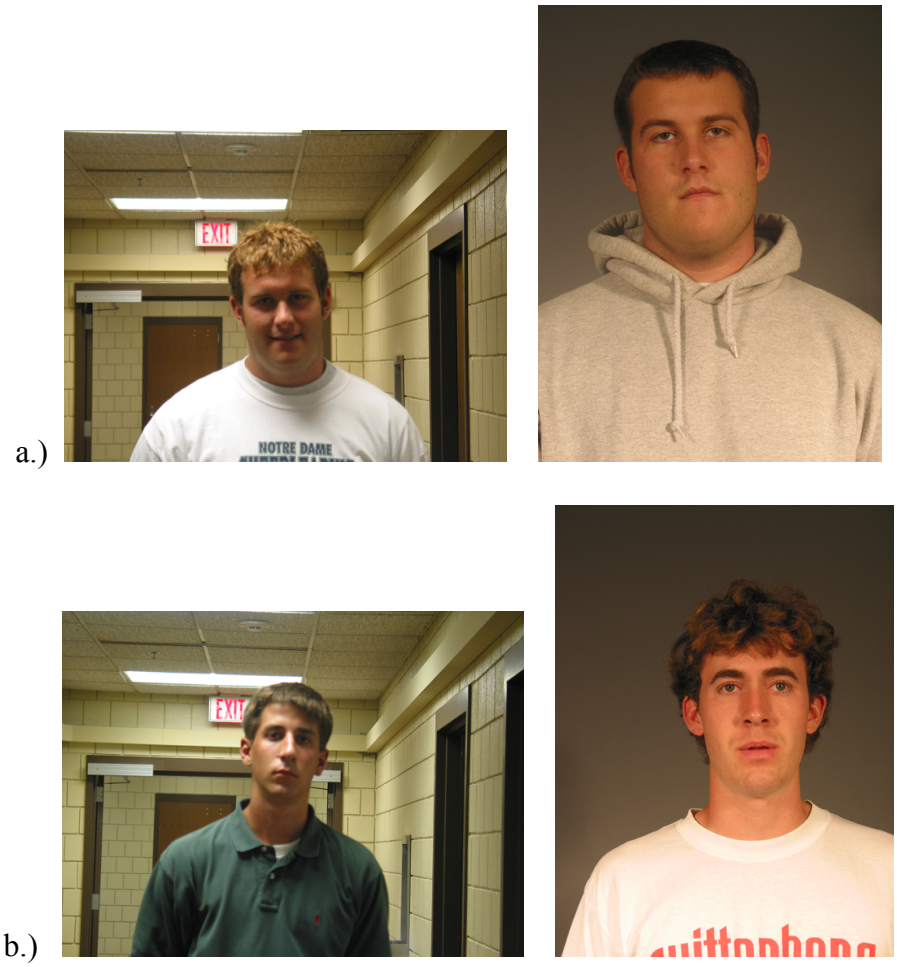

Fig. 1 A sample pair of face images from a "match" trial (a) and a " no match" trial (b). Participants responded by rating the likelihood that the pictures were of the same person using a five-point scale ranging from "1.) sure they are the same person" to "5.) sure they are not the same people."

II. EXPERIMENTS

A. Methods

1) Stimuli

Face stimuli were chosen from a large database developed for the FRCG study [13]. As noted, the uncontrolled illumination experiment from the FRCG used 8014 probe faces and 16028 target faces. The uncontrolled illumination probe faces had a resolution of 2272 x 1704 pixels. The controlled illumination target faces had a resolution of 1704 x 2272 pixels. For these experiments, we sampled face pairs from 128,448,392 pairs available. Of these, 407,352 (0.32%) were of the same people (match pairs) and 128,041,040 (99.68%) were of different people (non-match pairs, see Fig. 1 for an example pair).

To make the task as challenging as possible, we narrowed the available pairs to include only Caucasian males and females. Restriction to this relatively homogeneous set of faces eliminates the possibility that algorithms or humans can base identity comparisons on surface facial characteristics associated with race or age. For the same reason, the face pairs presented to participants were matched by sex.

Next we divided the face pairs into "easy" and "difficult." To estimate pair difficulty, we used a control algorithm based on principal components analysis (PCA) of the aligned and scaled images. PCA algorithms are an appropriate baseline because they have been available and widely tested since the early 1990's [17], [18]. The FRGC version of this baseline was designed to optimize performance [19]. We show that this algorithm reliably predicts "easy" and "difficult" sets of face pairs for both humans and algorithms. The algorithm itself, however, is not considered "state-of-the-art".

The baseline algorithm generated a 16028 x 8014 similarity matrix. We defined *difficult match pairs* to be image pairs of the same person with similarity scores less than two standard deviations below the average similarity score for the same-person match pairs (i.e., highly dissimilar images of the same person). *Easy match pairs* were image pairs of the same person with similarity scores greater than two standard deviations above the mean similarity score for the same-person pairs (i.e., highly similar images of the same person). *Difficult non-match pairs* and *easy non-match pairs* were chosen analogously (e.g., image pairs of different faces with similarity scores greater/less than two standard deviations above/below the average similarity score for the different-person pairs). The face pairs used in the experiments were selected randomly from face pairs meeting the above criteria.

2) Participants

Students from the School of Behavioral and Brain Sciences at University of Texas at Dallas volunteered to participate in these experiments in exchange for a research credit in a psychology course. A total of 91 students participated in the experiments (Exp. 1, n = 22, 10 females and 12 males; Exp. 2, n = 49, 25 females and 24 males; Exp. 3, n = 20, 10 females and 10 males).

3) Procedure

For all experiments, the task was to determine whether two face images, which appeared side by side on a computer screen, were pictures of the same person or of different people. Probe images were displayed on the left and target images were displayed on the right. The participants were asked to rate the image pairs using a 1 to 5

scale, “1.) sure they are the same; 2.) think they are the same; 3.) don’t know; 4.) think they are not the same; and 5.) sure they are not the same.”

In Experiment 1, 120 pairs of male faces served as stimuli. Half of the pairs were prescreened to be “easy” and half were prescreened to be “difficult”. Participants had unlimited time to enter a response for each pair, with images remaining on the screen until a response was entered.

Experiment 2 was similar to Experiment 1, but with exposure limited to two seconds. In this experiment, we included an equal number of male (120) and female (120) face pairs and balanced the inclusion of male and female human participants. Again, half of the pairs were prescreened to be “easy” and half were prescreened to be “difficult”. The exposure time was reduced to two seconds with the purpose of increasing error rates. This exposure time reduction did not diminish human performance significantly, and so Experiment 3 was conducted.

Experiment 3 was identical to Experiment 2, but with exposure time set to 500 milliseconds. This reduced performance considerably. A pilot study, not reported here, showed that the error rates did not increase substantially from the unlimited time condition when the exposure time was reduced to 1 second.

The algorithm-human comparisons we report in this study are based on data from Experiment 2. We chose this experiment for comparisons because we think that two seconds is a “realistic” exposure time for many security applications, and because the experiment includes a balanced number of male and female participants and male and female faces.

B. Results

1) Behavioral Data.

For the purposes of conducting inferential statistical analyses on the behavioral data, participant responses were transformed into "same" or "different" judgments for individual pairs of faces, so that *d'* could be computed. Responses 1 and 2 were deemed "same" judgments and responses 3, 4, and 5 were deemed "different" judgments. The correct verification rate (i.e., hit rate) was computed as the proportion of matched pairs correctly judged to be of the same person. The false acceptance rate (i.e., false alarm rate) was calculated as the proportion of non-match pairs judged incorrectly to be of the same person. A *d*' was then computed from the hit and false alarm rates as $Z_{\text{hit rate}} - Z_{\text{false alarm rate}}$. ROC curves were computed from the full range of response data that assigned certainty values to each match/no-match judgment. These curves provide analogous data to the ROCs computed from the algorithms, across the range of verification and false acceptance rates. The *d'* values served as the dependent variable in the analysis of variance (ANOVA) results reported for each experiment.

The difficulty level of face pairs (i.e., as defined by baseline control algorithm) was evaluated as a within-subjects variable in all three experiments. The effects of sex of the participant (between-subjects variable) and sex of face pair (within-subjects variable) were examined in Experiment 2, where both factors were closely balanced. No significant differences or interactions were found for these gender variables, and so we do not consider them further.

The PCA algorithm's ability to predict accuracy for humans was verified in all three face-matching experiments. Humans were significantly more accurate on face pairs

estimated by the PCA to be easy than they were on the face pairs estimated to be difficult (Fig. 2). This was true when people had unlimited time to match the pairs (Exp. 1, $F(1,20)=19.78$, $p < .0002$); when the pairs were presented for two-seconds (Exp. 2, $F(1,48)=96.53$, $p < .0001$); and when the pairs were presented for 500 milliseconds (Exp. 3, $F(1,18)=62.65$, $p < .0001$). Performance on the face-matching task was good, but not perfect (see Fig. 2, for the Exp. 2 results)**.** Remarkably, human participants performed no better with unlimited time to examine each pair than with a 2-second time limit ($F(1,176) = 1.65$, *ns.*). Match accuracy declined, however, when exposure time was limited to 500 milliseconds ($F(1,176) = 22.37$, $p < .0001$).

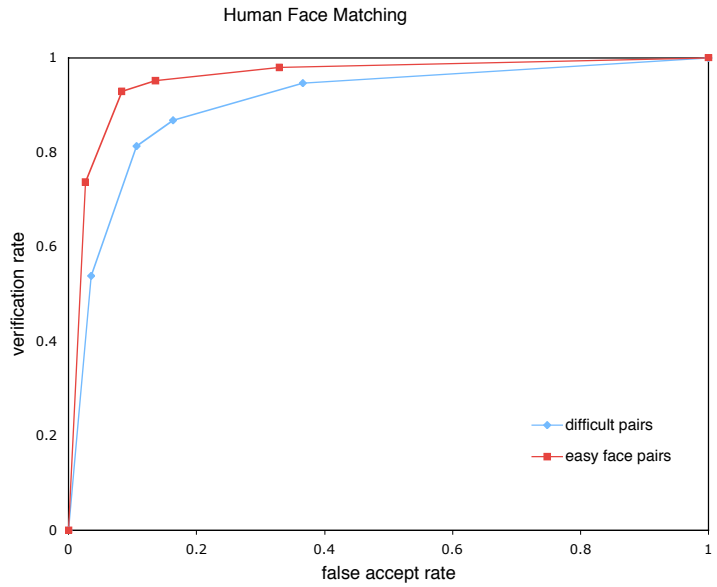

Fig. 2. The accuracy of humans matching face identities for "easy" versus "difficult" face pairs, with easy/difficult estimated by the baseline PCA algorithm. Face pairs found to be more difficult for the PCA were likewise more difficult for humans.

Combining across the three experiments, two findings are worth noting. First, human face matching accuracy was roughly constant for exposure times varying from two seconds to "unlimited" time. Only in Experiment 3, where exposure time was reduced to

500ms, was there a substantial decline in performance. This suggests that more time, or the use of a more analytic and time-consuming strategy by humans, would not have pushed performance levels to perfection. Whatever strategy humans employ to reach these "good" to "excellent" levels of performance seems to operate quickly and efficiently. The relatively stable performance of humans across the exposure times supports the use of these data as meaningful benchmark for algorithm performance.

A second result is that PCA can serve as an effective tool for pre-screening easy versus difficult face pairs for humans.

2) Human-Algorithm Comparison

We compared human performance in Experiment 2 to the performance of seven algorithms participating in the illumination experiment of the FRGC. Three of these algorithms were developed at academic institutions and four were developed by commercial enterprises.

As noted, each participating algorithm produced a 16028 x 8014 matrix of similarity scores for all the face pairs. The similarity scores of the 240 (120 male and 120 female) face pairs presented to participants in Experiment 2 were extracted from each similarity matrix and analyzed as follows. A full ROC curve was generated for each algorithm by sweeping a match criterion, c, across the distributions of "match" and "no-match" similarity scores and assigning same or different responses for individual face pairs, (e.g., similarity score greater than c yielded "same" and similarity score less than or equal to c yielded "different"). The verification and the false acceptance rates were used to compute ROC curves for the algorithms.

Figure 3a shows the performance of the algorithms and humans on the difficult face pairs. Three algorithms were more accurate than humans at this task and four algorithms were less accurate than humans. Two of the three algorithms that surpassed human performance were developed at academic institutions [20],[21]. The third algorithm is from industry and is described partially in a recent paper [22].

The performance of algorithms relative to humans for the easy face pairs was even better. Six of the seven algorithms performed more accurately than humans (Fig. 3b). The seventh algorithm exceeded human performance only at low false acceptance rates.

In all cases, algorithm performance was better for the "easy" face pairs than for the "difficult" face pairs, echoing the pattern for humans.

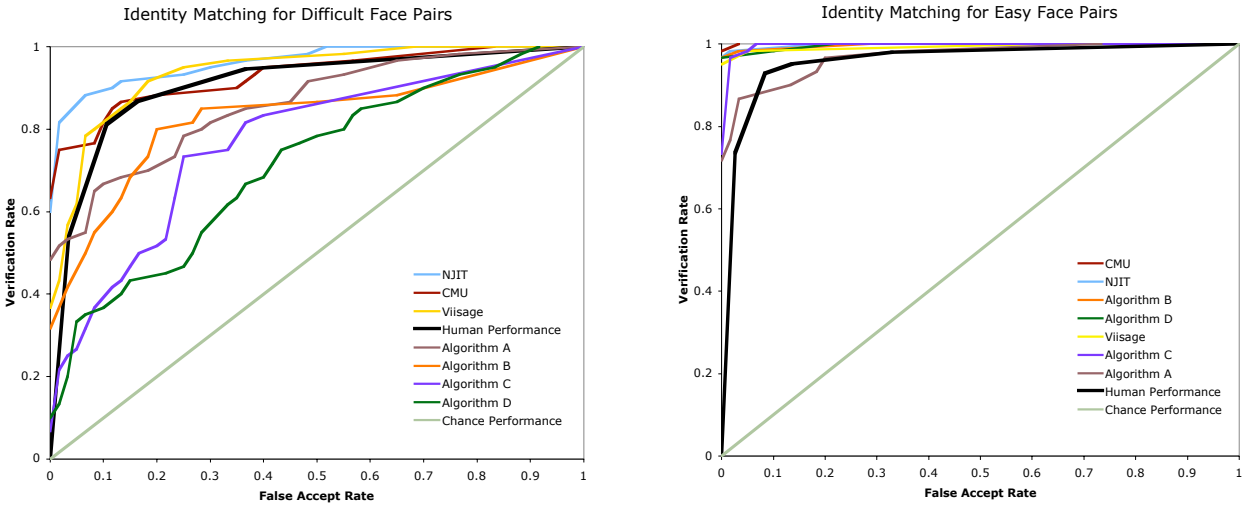

Fig. 3. Performance of humans and seven algorithms on the difficult face pairs (Fig. 3a) and easy face pairs (Fig. 3b) shown with ROC curves. Three algorithms outperform humans on the difficult face pairs at most or all combinations of verification rate and false accept rate (cf., [20] NJIT, [21] CMU for details on two of the three algorithms). Humans out-perform the other four algorithms on the difficult face pairs. All but one algorithm performs more accurately than humans on the easy face pairs.

Did human performance suffer from fatigue or waning attention over the course of the experiment? To examine this possibility, we assessed human performance over the

sequence of face pairs presented. Accuracy did not vary with trial number for either the verification rate (r = .07, *ns.*) or for the false acceptance rate (r = -.04, *ns.*). It is unlikely, therefore, that the comparatively lower performance of humans versus algorithms on the easy face pairs was due to difficulties in maintaining attention for the duration of the experiment.

III. GENERAL DISCUSSION

There is an implicit assumption among computer vision researchers, psychologists, and indeed much of the general public that human abilities recognizing and matching faces are currently beyond reach for machines. The present experiments challenge this assumption by showing that some current face recognition algorithms can compete with humans on a challenging task—matching face identity between photographs that are taken under different illumination conditions. Although algorithm performance may not seem impressive in absolute terms, it nonetheless compares favorably with humans.

The comparisons we report may lead us to wonder if human abilities with faces are overrated. Before concluding that they are, we note that human participants in these experiments were asked to match the faces of people previously unknown to them. This is an appropriate task for testing the abilities of a human security guard and/or algorithm, but may not show human face recognition skills at their best. The robust recognition abilities that characterize human expertise for faces may be limited to the faces of people we know well. Indeed, by contrast to performance with unfamiliar faces, human face recognition abilities for familiar faces are relatively robust to changes in viewing parameters such as illumination and pose [23], [24].

The consistent advantage of the algorithms over humans on the "easy" faces suggests that easy face pairs may simply be those in which the image-based features for the two faces are well matched and the illumination differences between images are minimal. Because image-based matching is a task that face recognition algorithms have done well for many years [1], it is perhaps not surprising that algorithms can compete with humans on this task. The algorithms might exploit useful, but subtle, image-based information that give them a slight, but consistent, advantage over humans.

An explanation for the better performance of algorithms on the difficult face pairs is less clear and may await more information and further tests of the algorithms that achieve this better performance. Although a discussion of these algorithms is beyond the scope of this paper, suffice to say that the Liu [20] and Xie et al. [21] algorithms represent a departure from past approaches. Both algorithms use kernel methods and both make efficient use of multiple "training" images of individuals to create a *face feature space*. The training images were made available to algorithm developers for "optional use," and were not included in the target or probe sets. The Liu [20] and Xie et al. [21] algorithms represent faces in the derived face feature space before matching by identity. It is possible that these algorithms work well because of information they can exploit from training images about the variability of individuals across changes in illumination. This may be similar to information humans acquire in developing face recognition skills and in becoming familiar with individuals.

The power of combining variable face images to improve face recognition has been explored also by Burton, Jenkins, Hancock, and White [25]. They used a technique based on averaging multiple aligned face images of the same person and found that it

outperformed comparable systems based on collections of instances. They demonstrated also that the averaging procedure produced faces that *humans* recognized more accurately than the original face images. Algorithms that make coordinated use of multiple images, therefore, can potentially offer insight into the question of how familiarity with multiple images of individual faces may help to buffer human and machine recognition against spurious changes in viewing parameters.

Finally, we tend to accept at "face" value, the pressing need to evaluate the performance of *any* algorithm that is placed in the field for a security application of importance. The tools and techniques for evaluating human face matching and recognition performance have been available for many years. These tools make it possible to test humans before assigning them to security tasks. This is an important, even critical, step for evaluating the level of security provided by human operators, by machine operators, and by the human-machine partnerships that will likely become commonplace in the near future for many security applications.

REFERENCES

[1] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, ″Face recognition: A literature survey, ″ *ACM Computing Surveys*, vol. 35, pp. 399-459, 2003.
[2] D. Blackburn, J.M. Bone, and P.J. Phillips, ″FRVT 2000 evaluation report. Technical report, ″ http://www.frvt.org, 2001.
[3] P.J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, W. Worek, ″Preliminary Face Recognition Grand Challenge Results, ″ *Proceedings of the Seventh International Conference on Automatic Face and Gesture Recognition*, in press.
[4] P.J. Phillips, H. Moon, P. Rizvi, and P. Rauss, ″The FERET evaluation method for face recognition algorithms, ″ *IEEE Transactions on Pattern Analysis and Machine Learning,* vol. 22, pp. 1090-1104, 2000.
[5] P.J. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J.M. Bone, ″FRVT 2002 evaluation report, ″ Tech. Rep. NISTIR 6965 http://www.frvt.org, 2003.
[6] P.J. Phillips, and E. Newton, ″Meta-analysis of face recognition algorithms, ″ *Proc. 5th Int. Conf. Auto. Face & Gest. Recog.,* pp. 235, 2002.
[7] W.J. Braje, ″Illumination encoding in face recognition: effect of position shift, ″ *Journal of Vision,* vol. 3, pp. 161-170, 2003.
[8] W.J. Braje, D. Kersten, M.J. Tarr, and N.F. Troje, ″Illumination effects in face recognition, ″ *Psychobiology*, vol. 26, pp. 371-380, 1999.
[9] W.J. Braje, G.E. Legge, and D. Kersten, ″Invariant recognition of natural objects in the presence of shadows, ″ *Perception,* vol. 29, pp. 383-398, 2000.
[10] H. Hill and V. Bruce, ″Effects of lighting on the perception of facial surface, ″ *Journal of Experimental Psychology: Human Perception & Performance,* vol. 22, no. 4, pp. 986-1004, 1996.
[11] N.F. Troje and H.H. Bülthoff, ″How is bilateral symmetry of human faces used for recognition of novel views? ″ *Vision Research,* vol. 38, pp. 79-89,1998.
[12] R. Gross, S. Baker, I. Matthews, and T. Kanade, ″Face recognition across pose and illumination, ″ in *Handbook of Face Recognition,* S. Z. Li and A.K. Jain, Eds. Springer, pp. 193-216, 2005
[13] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, ″Overview of the face recognition grand challenge, ″ *Proc. IEEE Computer Vision & Pattern Recognition*, vol. 1, pp. 947-954, 2005.

[14] M.J. Tarr and H.H. Bülthoff, ″Image-based object recognition in man, monkey and machine, ″ *Cognition*, vol. 67, pp. 1-20, 1998.
[15] Y. Adini, Y. Moses, and S. Ullman, ″Face recognition: the problem of compensating for changes in illumination direction, ″ *IEEE Transactions on Pattern Analysis and Machine Learning*, vol. 19, pp. 721-732, 1997.
[16] A.J. O'Toole, F. Jiang, D. Roark, and H. Abdi, ″Predicting human face recognition, ″ in *Face Processing: Advanced methods and models*, W-Y. Zhao and R. Chellappa, Eds. Elsevier, 2006.
[17] L. Sirovich and M. Kirby, ″Low-dimensional procedure for the characterization of human faces, ″ *Journal of the Optical Society of America,* vol. A, no. 4, pp. 519-524, 1987.
[18] M. Turk and A. Pentland, ″Eigenfaces for recognition, ″ *Journal of Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991.
[19] H.J. Moon and P.J. Phillips, ″Computational and performance aspects of PCA-based face recognition algorithms, ″ *Perception,* vol. 30, pp. 301-321, 2001.
[20] C. Liu, ″Capitalize on dimensionality increasing techniques from improving face recognition Grand Challenge performance, ″ *IEEE Transactions on Pattern Analysis and Machine Learning*, 2006, in press.
[21] C.M. Xie, M. Savvides, and V. Kumar, ″Kernel correlation filter based redundant class-dependence feature analysis (KCFA) on FRGC2.0 Data, ″ *IEEE International Workshop Analysis & Modeling Faces & Gestures,* pp. 32-43, 2005.
[22] M, Husken, B. Brauckmann, S. Gehlen, and C. von der Malsburg, ″Strategies and benefits of fusion of 2D and 3D face recognition, ″ *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 3, pp. 174, 2005
[23] P.J.B. Hancock, V. Bruce, and A.M. Burton, ″Recognition of unfamiliar faces, ″ *Trends in Cognitive. Sciences*, vol. 4, pp. 330-337, 2000
[24] A.J. O'Toole, D. Roark, and H. Abdi, ″Recognition of moving faces: a psychological and neural perspective, ″ *Trends in Cognitive Sciences*, vol. 6, pp. 261-266, 2002.
[25] A.M. Burton, R. Jenkins, P.J. Hancock, and D. White, ″Robust representations for face recognition: The power of averages, ″ *Cognitive Psychology,* vol. 51, pp. 256-284, 2005.

ACKKNOWLEGEMENTS

The primary goal of the FRGC is to encourage and facilitate the development of face recognition algorithms. To provide the face recognition research community with an unbiased assessment of state-of-the-art algorithms, research groups voluntarily submit similarity scores from prototyped experiments to the National Institute of Standards and Technology (NIST) for analysis. The results of the analysis by NIST are anonymous, unless otherwise agreed to by the participating algorithm developers. All participating groups were given the choice of remaining anonymous or being identified in this report. Performance results are from Jan. 2005 for all algorithms except Xie et al., 2005, where results are from Aug., 2005. This work was supported by a contract to A. O'Toole and H. Abdi from TSWG. P. J. Phillips was supported in part by funding from the National Institute of Justice. The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.