Comparing Face Recognition Algorithms to Humans on Challenging Tasks

ALICE J. O'TOOLE, XIAOBO AN, JOSEPH DUNLOP, and VAIDEHI NATU, The University of Texas at Dallas
P. JONATHON PHILLIPS, National Institute of Standards and Technology

We compared face identification by humans and machines using images taken under a variety of uncontrolled illumination conditions in both indoor and outdoor settings. Natural variations in a person's day-to-day appearance (e.g., hair style, facial expression, hats, glasses, etc.) contributed to the difficulty of the task. Both humans and machines matched the identity of people (same or different) in pairs of frontal view face images. The degree of difficulty introduced by photometric and appearance-based variability was estimated using a face recognition algorithm created by fusing three top-performing algorithms from a recent international competition. The algorithm computed similarity scores for a constant set of same-identity and different-identity pairings from multiple images. Image pairs were assigned to *good*, *moderate*, and *poor* accuracy groups by ranking the similarity scores for each identity pairing, and dividing these rankings into three strata. This procedure isolated the role of photometric variables from the effects of the distinctiveness of particular identities. Algorithm performance for these constant identity pairings varied dramatically across the groups. In a series of experiments, humans matched image pairs from the good, moderate, and poor conditions, rating the likelihood that the images were of the same person (1: sure same - 5: sure different). Algorithms were more accurate than humans in the good and moderate conditions, but were comparable to humans in the poor accuracy condition. To date, these are the most variable illumination- and appearance-based recognition conditions on which humans and machines have been compared. The finding that machines were never less accurate than humans on these challenging frontal images suggests that face recognition systems may be ready for applications with comparable difficulty. We speculate that the superiority of algorithms over humans in the less challenging conditions may be due to the algorithms' use of detailed, view-specific identity information. Humans may consider this information less important due to its limited potential for robust generalization in suboptimal viewing conditions.

Categories and Subject Descriptors: I.5.4 [**Pattern Recognition**]: Applications

General Terms: Algorithms, Human Factors, Verification, Experimentation

Additional Key Words and Phrases: Face recognition, human-machine comparisons

ACM Reference Format:

O'Toole, A. J., An, X., Dunlop, J., Natu, V., and Phillips, P. J. 2012. Comparing face recognition algorithms to humans on challenging tasks. ACM Trans. Appl. Percept. 9, 4, Article 16 (October 2012), 13 pages.
DOI = 10.1145/2355598.2355599 http://doi.acm.org/10.1145/2355598.2355599

This work was supported by funding from the Technical Support Working Group of the Department of Defense, USA. P. J. Phillips was supported in part by funding from the Federal Bureau of Investigation. The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

Authors' addresses: A. O'Toole, X. An, and J. Dunlop, School of Behavioral and Brain Sciences, GR4.1, The University of Texas at Dallas, Richardson, TX 75083-0688; P. J. Phillips, National Institute of Standards and Technology, 100 Bureau Dr., MS 8940, Gaithersburg, MD 20899; email: jonathon@nist.gov. Correspondence should be addressed to A. J. O'Toole; email: otoole@utdallas.edu.

DOI 10.1145/2355598.2355599 http://doi.acm.org/10.1145/2355598.2355599

1. INTRODUCTION

When humans and machines differ on a face recognition judgment, which "system" is more likely to be correct? It is widely but incorrectly believed that humans are more accurate than machines at most face recognition tasks. In fact, comparisons made between humans and algorithms in two recent international competitions, the Face Recognition Grand Challenge (FRGC) [Phillips et al. 2005], and the Face Recognition Vendor Test (FRVT 2006) [Phillips et al. 2010] showed that the best face recognition algorithms surpassed humans at the task of matching identity in frontal face images [O'Toole et al. 2007, 2008]. For those comparisons, one image in the pair was taken under controlled illumination (studio or "mugshot" quality) and the other image was taken under uncontrolled (indoor ambient) illumination. In O'Toole et al. [2007], the image pairs were prescreened by a pixel-based baseline algorithm into "easy" and "difficult" face pairs. Notably, on the easy face pairs, the performance advantage for algorithms over humans was substantially larger than on the difficult face pairs. For the top-performing algorithms, however, the performance advantage for machines was evident even for identifications prescreened to be "difficult" for the algorithms.

The impressive performance of algorithms relative to humans in these tests might be interpreted to suggest that the problem of automatic face recognition from frontal images is "solved," even when illumination varies. A more recent assessment of algorithm performance with a more challenging dataset of frontal face images indicates that this conclusion is premature [Phillips et al. 2011]. The test was conducted by National Institute of Standards and Technology, and called *The Good, The Bad and The Ugly (GBU) Challenge* [Phillips et al. 2011]. Its purpose was to understand the characteristics of face image pairs that are identified by current algorithms with high (the "good"), moderate (the "bad"), and poor ("the ugly") accuracy. In particular, the effects of natural variations in a person's day-to-day appearance (hair, facial expression, etc.) and variations in illumination across both indoor and outdoor settings were considered. An important control imposed on the stimulus set in this test was that all three data sets were made up of identity pairs of the *same* individuals. Thus, only the images, not the individual identities, changed across the three groups. This provides an assurance that the accuracy differences for algorithms were due to factors other than the particular set of face identities tested.

To arrive at the performance-based data conditions in the GBU evaluation, three top-performing face recognition algorithms from the FRVT 2006 algorithm test [Phillips et al. 2010] were computationally "fused" to produce a single algorithm. This involved combining estimates generated by the three algorithms of the similarity between each pair of face images (cf., Methods section for details of the fusion procedure). This "similarity score" forms the basis for an identification judgment, with higher similarity scores indicating a higher likelihood that two images are of the same individual. For each same-identity pairing of an individual, multiple similarity scores were available from different image pairings. These scores were ranked and partitioned into three groups. The face pairs with the highest third of the scores comprised the "high accuracy condition," the face pairs with scores in the middle third comprised the "moderate accuracy condition," and face pairs in the lowest third of the scores comprised the "low accuracy condition." Figure 1 shows three pairs of images of the same person, sampled from the good (left), moderate (middle), and poor (right) performance conditions. This figure illustrates the wide variation in the appearance of a person across frontal images. It also highlights the difficulties that may occur in matching identity in pairs of images that are taken in different settings and which include variations in expression and appearance-based features such as hairstyle. These factors become even more salient in combination (cf., Figure 1, right).

More concretely, Figure 2 contains the distribution of algorithm-generated similarity scores for matched (images of the same person) and nonmatched (images of different people) identity pairs found

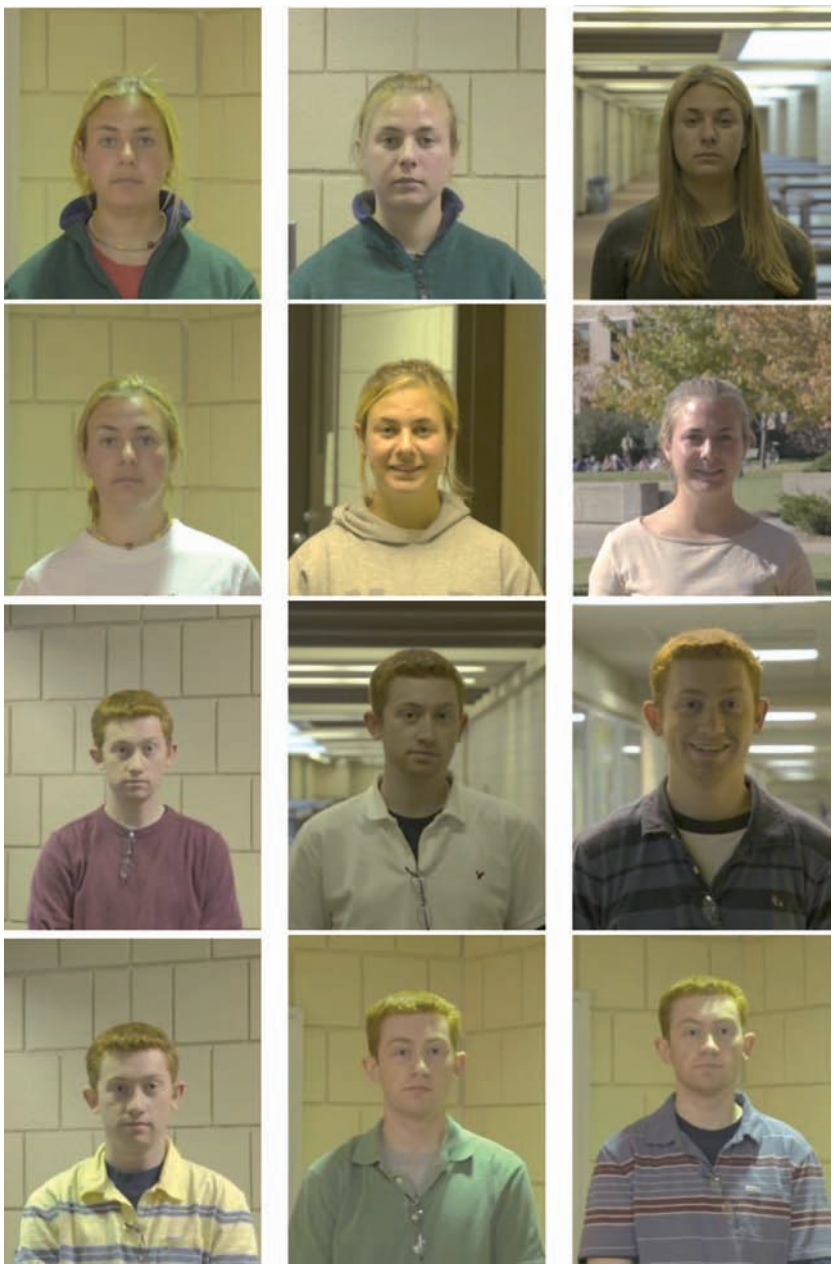

Fig. 1. Two example stimuli from the good, moderate, and poor distributions to illustrate the variation in the challenge level associated with matching identity in the three performance partitions. The top two rows show three pairs of images of the same person, sampled from the good (left), moderate (middle), and poor (right) performance conditions. The second two rows show the same type of sample for a second person. These images are from the Face and Ocular Challenge Series dataset, reprinted here with permission of Prof. Patrick Flynn.

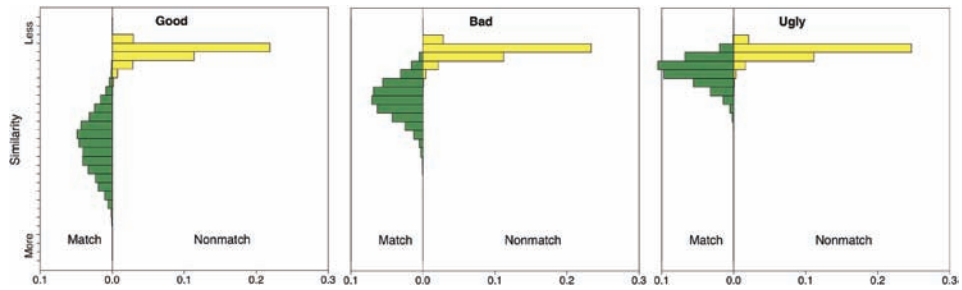

Fig. 2. This figure shows the distribution of algorithm-generated similarity scores for matched (images of the same person) identity pairs and nonmatched (images of different people) identity pairs for each GBU performance partition.

in the GBU challenge. As is clear, the good condition (the top third of cases) yields minimal overlap between similarity scores for the matched and nonmatched identity pairs, whereas the poor condition (the bottom third of cases) shows strong overlap between the matched and nonmatched distributions. The performance of the algorithm is highly accurate for the best cases and quite poor for the worst cases. Note that the three distributions in Figure 2 contain the same pairs of face identities, isolating the impact of photometric and appearance-based variation on recognition.

The purpose of the present study was to compare humans to face recognition algorithms with stimuli that portray faces with natural variations in appearance and illumination. Specifically, we compared machines and humans as a function of the level of challenge. Illumination variation is well-known to affect both machine and human face recognition performance [Gross et al. 2005; Adini et al. 1997; Braje 2003; Braje et al. 1999; Hill and Bruce 1996; Johnston et al. 1992]. Although these natural changes in appearance (e.g., hair style, expression) are common in everyday life, they have not been studied extensively either by psychologists or by computer vision researchers. One exception is a recent study by Jenkins and colleagues that provides a striking demonstration of the variability of a person's appearance across multiple images [Jenkins et al. 2011]. In that study, participants were asked to decide how many different individuals were pictured among 40 "Web-gathered" images, comprised of 20 images each of two individuals. On average, participants "found" between seven and eight unique identities in the 40 images.

The data from the GBU challenge provided a unique opportunity to look at human performance over natural variations in viewing conditions that are problematic for machine-based recognition systems. In the first set of experiments, we measured human performance matching the identity of face pairs sampled from the match and nonmatch distributions generated by the GBU evaluation (Figure 2). These represent different degrees of difficulty for algorithms. How do humans compare to algorithms as the difficulty for the algorithms increases? In the first set of experiments, we sampled the distributions near their respective means and compared short exposure times with unlimited viewing times. In a second set of experiments, we sampled the moderate and difficult distributions more broadly and tested two separate groups of participants with a larger number of image pairs. The comparison across the first two experiments offers insight into the extent to which humans maintain a stable criterion for determining the level of similarity needed to affirm or reject an identification match. In combination, the results of the study provide a basis for understanding how human face recognition abilities compare to those of current algorithms as the level of difficulty increases. This is especially interesting in cases, such as the present one, where neither the human nor machine performs flawlessly.

2. EXPERIMENTS

2.1 Stimulus Set

The data set for the GBU challenge was constructed from face images in the multi-biometric database collected at Notre Dame University for the FRVT 2006 [Phillips et al. 2010, 2012]. These images were selected from a larger set collected either outside or with ambient indoor illumination in a corridor. All images were acquired with a 6 Mega-pixel Nikon D70 camera, and were taken between Aug. 2004 and May 2005. As noted, because multiple images were available for all of the subjects in the database, it was possible to assess algorithm performance for matching the identity of the same person with many image pairings. This was also true for pairings of different identities. The creation of three performance-based stimulus partitions was carried out by ranking the similarity scores generated by the fused algorithm (see below) and dividing them into three strata to produce the good, moderate, and poor performance partitions. Stimuli from these algorithm-generated partitions formed three conditions for the human participants in these experiments. We refer to these conditions henceforth as the good, moderate, and poor conditions.

Each condition was constructed using two sets of images. Set 1 included 1085 images of 457 individuals and Set 2 included 1085 different images of the same individuals. To balance subject counts across the three conditions, the number of images per person was the same in Sets 1 and 2. To assure that identity matches could not be based on trivial appearance cues (clothing, hair), the images in all matched identity pairs were taken on different days.

The task of the algorithm was to assign a similarity score to all possible pairs of images from Sets 1, and 2. Thus, for each condition, the algorithm produced a 1085×1085 similarity matrix of $1,177,225$ similarity scores, where element $s_{i,j}$ of the matrix contained the similarity between the i^{th} image in Set 1 and the j^{th} image in Set 2. For each condition, the matrix contained similarity scores for 3297 matched identity pairs and $1,173,928$ nonmatched identity pairs. The matched and nonmatched identity pairings across these three conditions were exactly the same.

2.2 Algorithm Source

The FRVT 2006, a U.S. Government-sponsored international competition for face recognition algorithms, was the source of the algorithm used in this study. Complete information and results for this competition can be found elsewhere [Phillips et al. 2010]. For present purposes, the algorithm we tested was a fusion of three top-performing algorithms from the FRVT 2006. All three of these algorithms were submitted by commercial enterprises. We note at the outset that the FRVT 2006, and similar US Government-sponsored algorithm competitions, are conducted without access to the source code of the algorithms entered. Instead, executable versions of the programs, installed at NIST, are used for the test. The use of executables protects the proprietary nature of the code and makes it more likely that the very best commercial algorithms will participate in the evaluation. The downside, however, is that it is not possible to know precisely how individual algorithms operate. For this reason, it makes sense to use a combination of several good algorithms rather than any individual algorithm.

This fusion algorithm was created by combining the similarity scores generated by the three algorithms for all possible pairs of the Set 1 and Set 2 images. The fusion was computed in a two-step process. In the first step, for each algorithm, the median and the median absolute deviation (MAD) were estimated from 1 in 1023 similarity scores ($median_k$ and MAD_k are the median and MAD for algorithm k). This sampling method was used to avoid "over tuning" the estimates to the data. The similarity scores were selected to evenly sample the images in the experiment. The fused similarity

scores were computed as follows: If s_k is a similarity score for algorithm k and s_f is a fusion similarity score, then $s_f = \sum_k (s_k - \text{median}_k)/\text{MAD}_k$.

3. EXPERIMENTS 1A AND 1B: HUMAN PERFORMANCE - SAMPLING FROM NEAR THE DISTRIBUTION AVERAGES

In this first set of experiments, humans matched the identity of face pairs sampled from the good, moderate, and poor distributions generated by the fused face recognition algorithm (see Figure 2). In Experiment 1a, participants saw each pair of faces for 2 seconds before entering a judgment. In Experiment 1b, participants were allowed to view the faces for an unlimited amount of time. For all three conditions, the stimulus pairs presented to participants were selected from close to the mean of their respective matched and nonmatched distributions. Because of the large number of available pairs of images in each distribution, and the smaller number of pairs we can reasonably present to human participants, it was possible to select a set of matched and nonmatched pairs with virtually no overlap in similarity scores for all three conditions. In other words, on these image pairs, the algorithm performed perfectly for the the good, moderate, and poor conditions (see Figure 2). Despite the perfect performance of the algorithm with samples chosen in this way, as we will see, human performance can be linked to these algorithm-generated similarity scores, independent of the local distribution of scores in the experimental conditions.

3.1 Methods

3.1.1 *Participants.* Undergraduate students from the School of Behavioral and Brain Sciences at The University of Texas at Dallas volunteered to participate in these experiments in exchange for a research credit in a psychology course. A total of 21 students (14 females and 7 males) participated in Experiment 1a and 22 students (15 females and 7 males) participated in Experiment 1b.

3.1.2 *Stimuli.* For each condition, image pairs were chosen by beginning at the mean of each distribution and sampling pairs with similarity scores directly around these means. Any image pairs with identities duplicated in the condition were eliminated. We selected 40 pairs of matched identity images and 40 pairs of non-matched identities for each condition. The images were 752 pixels wide and 500 pixels in height. For consistency, the images were presented on the same high quality 24-inch Apple monitor for all experiments.[1]

3.1.3 *Procedure.* Participants were instructed about the purpose and procedure of the experiment. On each trial, they viewed a pair of images, presented side by side on the computer screen. In Experiment 1a, the images remained visible for 2s, and then disappeared, after which a text prompt appeared asking the participant to choose one of the following responses: (1) Sure they are the same person; (2) Think they are the same person; (3) Don't know; (4) Think they are different people; (5) Sure they are different people. In Experiment 1b, the face pair appeared on the screen along with the prompt, and remained visible until the participant entered a response.

In both cases, the participant's rating was used in this study as a human-generated measure of the similarity of the faces in each pair. This is analogous to the similarity scores generated by the algorithm.

Both experiments consisted of 240 trials of face pairs (120 matched pairs and 120 non-matched pairs; 80 pairs from each condition).

[1]The monitor was not calibrated to control for gamma distortion for two reasons. First, given the variability of the illumination conditions in the images, any calibration would be likely to have inconsistent effects, possibly improving some images and making others worse. Second, in terms of ecological validity, humans who would perform this type of a task in a security application would likely do so on a standard uncalibrated computer monitor.

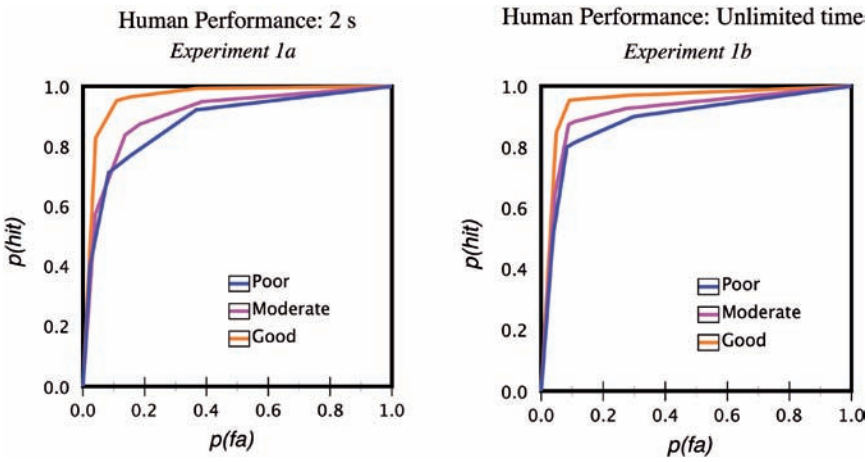

Fig. 3. This figure shows the performance of humans on face pairs from the good, moderate, and poor categories of algorithm performance with 2s exposures (left) and unlimited time (right). The figure also illustrates that the performance-based stratifications for algorithms are also seen with humans.

3.2 Results: Experiment 1a

The results were analyzed in two ways. First, we computed receiver operator characteristic (ROC) curves from the rating scale used by participants to indicate the likelihood that the people in the image pairs were the same [Macmillan and Creelman 1991]. These appear in Figure 3 (left) and show a stratification of performance for the good condition over the moderate and poor conditions. The moderate and poor conditions show roughly equivalent performance, with a small tendency for better performance in the moderate over the poor condition. Second, and more formally, we used the summary measure of *d'* to test for statistical differences between the conditions. The average accuracies for the conditions were as follows: good (*d' = 3.09, se = .12*), moderate (*d' = 2.32, se = .09*), and poor (*d' = 2.09, se = .09*). A one-factor within-subjects analysis of variance showed a main effect of condition, $F(2, 40) = 36.66; p < .0001$. As indicated by the standard errors, the difference between the good and the moderate conditions, $F(1, 40) = 39.74, p < .001$, and between the good and poor conditions, $F(1, 40) = 66.83, p < .001$, were highly significant. Human performance in the moderate and poor conditions did not differ significantly, $F(1, 40) = 3.48, ns$.

3.3 Results: Experiment 1b

The ROCs for the unlimited exposure time appear in Figure 3 (right) and show a rough stratification of performance for the good, moderate, and poor conditions. Again, the moderate and poor conditions show roughly equivalent performance, with a small tendency for better performance in the moderate over the poor condition. The average accuracies for the conditions were as follows: good (*d' = 3.19, se = .13*); moderate (*d' = 2.77, se = .14*); and poor (*d' = 2.42, se = .11*). A one-factor within-subjects analysis of variance showed a main effect of condition, $F(2, 42) = 25.54; p < .0001$. As indicated by the standard errors, the difference between the good and the moderate conditions, $F(1, 42) = 15.03, p < .001$, and between the good and poor conditions, $F(1, 42) = 51.10, p < .001$, were highly significant. Human performance in the moderate and poor conditions differed significantly in this case with unlimited exposures, $F(1, 42) = 10.63, p < .001$.

To determine if the unlimited viewing times given to the participants in Experiment 1b improved performance over the 2-second viewing times given in Experiment 1a, we computed an ANOVA across

the experiments with independent variables of exposure time (2 seconds and unlimited) and difficulty condition (good, moderate, and poor). This analysis revealed a small, but statistically significant, advantage for the unlimited viewing time condition, $F(1, 41) = 4.84$, $p < .03$).

3.4 Conclusion

The results of the experiment suggest that condition-based differences in the algorithm-generated similarity scores were reflected in the human performance data. Humans were most accurate for the face pairs from the algorithm-generated good partition. Humans showed roughly equal performance for the moderate and poor algorithm-generated partitions. The comparison between short exposures (2s) and unlimited viewing time indicated a small, but reliable, benefit for the unlimited time condition. As noted, because the face pairs in this experiment were sampled from the center of the appropriate distributions, algorithm performance for these stimulus pairs was perfect in all three conditions.

4. EXPERIMENTS 2A AND 2B: UNIFORM SAMPLING OF DISTRIBUTIONS

The goal of the second set of experiments was to compare human and machine performance across a range of similarity scores. To do this, we sampled face pairs broadly across each of the distributions. This yielded a meaningful level of overlap for the algorithm-generated match and nonmatch distributions for the moderate and poor conditions. For the human experiments, we focused here on the poor (Experiment 2a) and moderate (Experiment 2b) performance conditions, because there was virtually no overlap in the match and nonmatch distributions for the good condition. We also sampled a larger number of face pairs for the moderate and poor conditions. On average, we expected the summary statistics for human performance in the three conditions to replicate those found in Experiment 1. In addition, the sampling implemented here allowed us to directly examine the relationship between human- and algorithm-generated similarity scores. How well do machine-generated scores predict human identity judgments?

4.1 Methods

4.1.1 *Participants.* Undergraduate students from the School of Behavioral and Brain Sciences at The University of Texas at Dallas volunteered to participate in these experiments in exchange for a research credit in a psychology course. A total of 23 students (12 females and 11 males) participated in Experiment 2a and 30 (19 females and 11 males) students participated in Experiment 2b.

4.1.2 *Stimuli.* For each condition, we selected 120 pairs of matched identity images and 120 pairs of nonmatched identities for the moderate and poor performance conditions. Sampling was centered at the means of the appropriate distributions and was implemented to span the range of scores in the distributions.

4.1.3 *Human Procedure.* The human procedure was identical to that described for Experiment 1a with the exception that participants in Experiment 2a matched identity in 240 face pairs from the moderate condition, and participants in Experiment 2b matched 240 identity in face pairs from the poor condition.

4.1.4 *Algorithm Procedure.* To measure the algorithm performance in terms comparable to human performance, we computed the ROC curve for the algorithms in the good, moderate, and poor conditions. The ROC plots the trade-off between the hit rate and the false accept rate as a threshold is varied. Analogous to the human data analysis, a successful verification occurs when an algorithm correctly judges "same" in response to a pair of images of the same person. A false accept occurs when an algorithm incorrectly judges a nonmatched face pair as "same." It is worth noting that the nonmatch

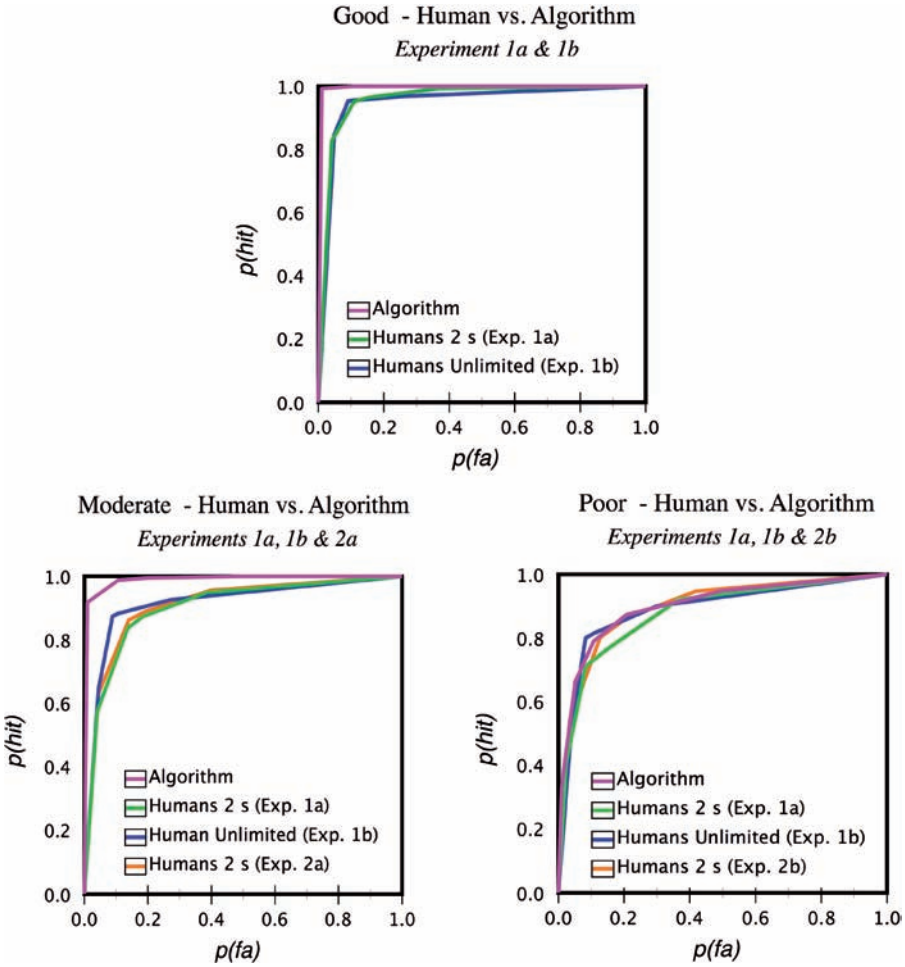

Fig. 4. This figure shows the performance of humans on face pairs from the good, moderate, and poor conditions for Experiments 1 and 2. The human performance is plotted with the algorithm performance for comparison. The superior performance of the algorithm is evident in the good and moderate conditions. The humans and algorithm perform comparably for the poor condition.

distributions for the algorithm ROC contain pairs of images that may or may not match in gender or race. This variability of the nonmatched pairs will tend to over-estimate algorithm performance by a small amount (cf., O'Toole et al. [2012] for additional quantitative detail). As we will see, the human-machine comparisons we present show quite large differences in performance that cannot be explained by the inclusion of cross-demographic nonmatch pairs.

4.2 Results

The ROC curves for these experiments appear in Figure 4, plotted with the comparable conditions from Experiment 1 and with data taken from the algorithm performance. For comparison, we also include the human versus machine performance for the good condition from Experiment 1 in the top panel of the figure. This top panel shows simply that the algorithm performs far better than humans on face pairs from the algorithm-generated good distribution.

For the moderate condition (middle), the algorithm also performs much better than humans. This was the case in all three of the experiments we conducted with this set of faces. This includes Experiments 1a (2s exposures) and 1b (unlimited time) with face pairs sampled from the centers of the distributions, as well as Experiment 2a (2s exposures) with more broadly sampled image pairs. It is also clear from this graph that Experiment 2a provides a close replication of the analogous Experiment 1a, $F(1, 42) < 1$.

In the most difficult cases, which we tested in the poor condition, human and algorithm performance was roughly comparable. The graph makes clear also that human performance was reasonably stable in this condition, again with a replication of accuracy level for Experiment 1b and the analogous Experiment 2b, $F(1, 49) < 1$.

4.2.1 *Human and Algorithm Similarity Scores.* Next, we considered whether human and machine responses were linked at the level of individual stimulus pairs. To do this we assessed the correlation between human and machine-generated similarity scores. First, we calculated an average human-generated similarity score for each face pair in Experiments 2a and 2b. This was defined as the average of the ratings human participants gave the pair on the 5-point scale (i.e., 1: sure the same, 2: think the same .... 5: sure different). Next, we computed a correlation coefficient for the human and machine scores assigned to the 240 pairs of images from the moderate and poor performance conditions. In both cases, the correlations were strong and statistically significant (moderate, $r(239) = -.8838$, $p < .001$, poor $r(239) = -.7112$, $p < .001$). (Note that for the algorithm, high scores indicate high similarity, whereas for the human ratings, the scale is reversed. Thus, we find a negative correlation between human and machine-generated similarity). Figure 5 shows the plot of human similarity scores against model similarity scores for the moderate (left) and poor (right) conditions. Though clearly related, there is sufficient scatter to suggest differences in the approaches applied by humans and machines.

It is perhaps worth noting that although it can be informative to look at the stimuli that make up the outliers in this graph, it is difficult to easily find pairs that completely characterize where the models err and the humans succeed (and vice versa). More often than not, human-machine pairs that produce strongly disparate similarity estimates contain combinations of mismatched poor-quality factors (slight blur, very small misalignments in view, illumination differences, and expression oddities). Thus, although it is often possible to look at these examples and guess what went wrong for the algorithm or human, we would hesitate to make general claims based on a few examples. A more systematic analysis of these examples, however, is worthy of further study.

5. CONCLUSION

The goal of this study was to compare human accuracy at identifying faces with accuracy of state-of-the-art face recognition algorithms. In particular, we were interested in how humans perform relative to machines as the level of difficultly increases. Our tests relied on images taken under a variety of uncontrolled illumination conditions in both indoor and outdoor settings and with natural variations in a person's day-to-day appearance. To date, these are the most variable illumination- and appearance-based recognition conditions on which humans and machines have been compared. The results of this study indicate that the best current face recognition algorithms perform better than humans in all but the most challenging cases of matching identity in frontal face images. Moreover, as the task difficulty increased, the gap between machines and humans narrowed, but never reversed. Thus, for frontal images, even with quite substantial changes in illumination and appearance, machines are now comparable to humans.

A second interesting finding was the close relationship between the similarity scores generated by humans and machines. This result suggests that to a first approximation, the judgments generated by

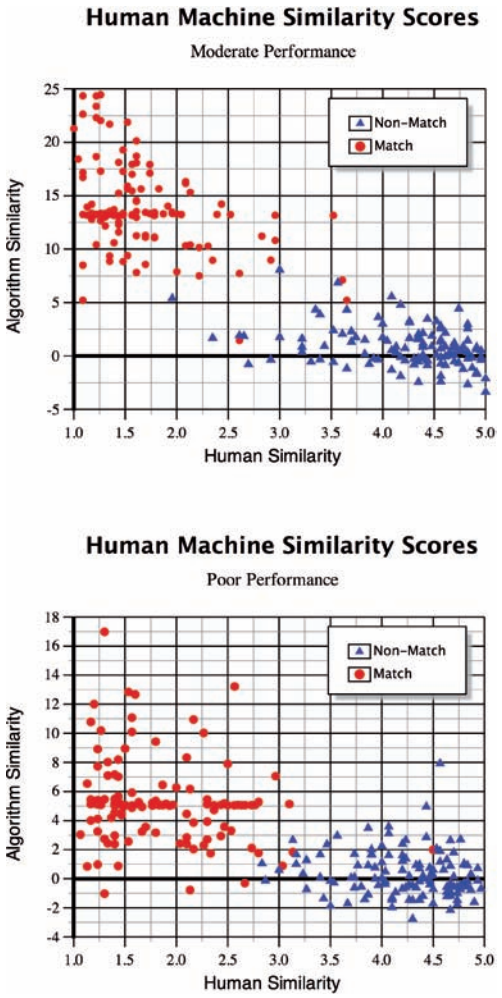

Fig. 5. This figure shows a strong but imperfect relationship between human and algorithm-generated similarity scores for moderate (top) and poor (bottom) performance conditions. Note that for humans, high numbers indicate low similarity, whereas for the algorithm, the inverse is true.

the two "systems" are related. Notwithstanding, the remaining scatter in the human-machine similarity scores may indicate strategic differences in the way humans and machines judged the similarity of the faces. As noted, although these differences can be difficult to characterize generally, they are likely to be related to combinations of quality factors that are differentially problematic for humans and machines. Concomitantly, it is also possible that humans also make some use of non-face configural information from the combination of the face and body (e.g., neck, shoulders). This information is unlikely to be used by the face recognition algorithms. In recent work, we show that humans do use some external body information in cases where the face provides limited information [Rice et al. 2012]. In a previous study based on data from the earlier FRGC algorithm competition, strategic differences between humans and machines were examined using a statistical-based fusion of the human similarity scores and the scores from seven face recognition algorithms. The fusion improved accuracy to a

level better than either the machines or humans operating alone [O'Toole et al. 2007], suggesting at least partial independence in the processes used by humans and machines to judge face similarity.

A striking aspect of the present data was the substantial accuracy advantage for the machines over the humans in the two less challenging conditions. Again, this result is consistent with the finding we noted previously from a human-machine comparison in the FRGC face recognition algorithm competition. In that earlier competition, the best algorithms were better than humans on the "difficult identity pairs," but nearly all of the algorithms were better than humans on the "easy" pairs [O'Toole et al. 2007]. One possible explanation of these two findings is that the face recognition algorithms tested in the FRCG and FRVT 2006 are highly tuned to operate on faces imaged from the frontal view. This may make it possible for the algorithms to exploit information in the frontal view that is of limited general value, as the face pose varies away from the front. Small variations in face texture or facial markings (and their configuration) are examples of these features. We speculate that humans may attend less to this kind of detailed information in making their judgments. Although this may limit their performance potential when faces are viewed frontally, it may make for more stable and reliable coding over changes in viewing conditions. This does not suggest that humans create a three-dimensionally invariant representation of the face. Rather, it is consistent with the possibility that humans can tune their perceptual systems to multiple views with some inherent flexibility around canonical views.

Finally, the superiority of machines over humans in this study reminds us of what face recognition algorithms have not yet accomplished – recognition robustness over large changes in viewpoint. Although human recognition for unfamiliar faces is affected by changes in viewpoint [Hancock et al. 2000], familiar face recognition is not [Burton et al. 2011; Johnston and Edmonds 2009]. More generally, differences between the performance of humans (at their best) and state-of-the-art face recognition algorithms are analogous to differences between humans recognizing familiar versus unfamiliar people. We would argue that machines now perform at an accuracy level comparable to humans when they recognize "unfamiliar faces". Changes in illumination, expression, viewpoint, and appearance, which are problematic for unfamiliar faces, are much less challenging for recognizing people we know well (e.g., friends, family, famous people) [Johnston and Edmonds 2009]. In the recent study by Jenkins and colleagues, recall that many identities were perceived from a large set of images of two people [Jenkins et al. 2011]. Notably, Jenkins et al. also compared Dutch and U.K. observers on a comparable test made with two well-known Dutch personalities. The results indicated that almost all of the Dutch participants performed perfectly on the task, while the U.K observers "found" significantly more identities among the set of images. The behavioral disparities for familiar versus unfamiliar faces are accompanied by a variety of differences in the way these faces are represented neurally [Gobbini and Haxby 2011; Natu and O'Toole 2011]. Ultimately, the elaborated visual and neural codes that support familiar face recognition may give insight into computational processes that may overcome photometric changes in viewing conditions.

To compete successfully with humans, the next generation of face recognition algorithms will have to operate with levels of robustness comparable to those humans show in recognizing familiar faces. There is some evidence to suggest that this kind of recognition may involve representations of the face and body and may include some part of identity-specific motions, such as facial gesture and gait recognition [Haxby et al. 2000; O'Toole et al. 2002]. Notwithstanding, the present accomplishments of face recognition algorithms are impressive and, within constrained viewing conditions, may now be considered as good as, or better than, humans. This should make it possible to use these algorithms in applications that can be constrained appropriately. Moreover, the present data suggest that when humans and machines disagree in these circumstances, the human is not more likely to be correct.

ACKNOWLEDGMENTS

We would like to thank Sam Weimer for testing subjects and for stimulus preparation, as well as Allyson Rice and three reviewers for helpful comments on a previous version of the manuscript.

REFERENCES

ADINI, Y., MOSES, Y., AND ULLMAN, S. 1997. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Anal. Mach. Intell. 19*, 721–732.

BRAJE, W. 2003. Illumination encoding in face recognition: Effect of position shift. *J. Vision 3*, 161–170.

BRAJE, W., KERSTEN, D., TARR, M. J., AND TROJE, N. 1999. Illumination effects in face recognition. *Psychobiology 26*, 371–380.

BURTON, A. M., JENKINS, R., AND SCHWEINBERGER, S. R. 2011. Mental representations of familiar faces. *British J. Psychol.102*, 943–958.

GOBBINI, M. I. AND HAXBY, J. V. 2011. Neural systems for recognition of familiar faces. *Neuropsychologica 45*, 32–41.

GROSS, R., BAKER, S., MATTHEWS, I., AND KANADE, T. 2005. Face recognition across pose and illumination. In *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds., Springer, Berlin, 193–216.

HANCOCK, P. J. B., BRUCE, V., AND BURTON, A. M. 2000. Recognition of unfamiliar faces. *Trends Cognitive Sci. 4*, 330–337.

HAXBY, J., HOFFMAN, E., AND GOBBINI, M. 2000. The distributed human neural system for face perception. *Trends Cognitive Sci. 20*, 6, 223–233.

HILL, H. AND BRUCE, V. 1996. Effects of lighting on the perception of facial surface. *J. Experiment. Psychol. 22*, 986–1004.

JENKINS, R., WHITE, D., MONFORT, X. V., AND BURTON, A. M. 2011. Variability in photos of the same face. *Cognition 121*, 313–323.

JOHNSTON, A., HILL, H., AND CARMEN, N. 1992. Recognising faces: effects of lighting direction, inversion and brightness. *Perception 21*, 365–375.

JOHNSTON, R. A. AND EDMONDS, A. J. 2009. Familiar and unfamiliar face recognition: A review. *Memory 17*, 5, 577–596.

MACMILLAN, N. A. AND CREELMAN, C. D. 1991. *Detection Theory: A User's Guide*. Cambridge University Press, Cambridge, UK.

NATU, V. AND O'TOOLE, A. J. 2011. The neural processing of familiar and unfamiliar faces: A review and synopsis. *British J. Psychol. 102*, 726–747.

O'TOOLE, A., ABDI, H., JIANG, F., AND PHILLIPS, P. J. 2007. Fusing face recognition algorithms and humans. *IEEE Trans. Syst. Man Cybern*. 37, Part B, 1149–1155.

O'TOOLE, A., ROARK, D., AND ABDI, H. 2002. Recognition of moving faces: A psychological and neural perspective. *Trends Cognitive Sci. 6*, 261–266.

O'TOOLE, A. J., PHILLIPS, P. J., AN, X., AND DUNLOP, J. 2012. Demographic effects on estimates of automatic face recognition. *Image, Vision, Comput. 30*, 169–176.

O'TOOLE, A. J., PHILLIPS, P. J., JIANG, F., AYYAD, J., PENARD, N., AND ABDI, H. 2007. Face recognition algorithms surpass humans matching faces across changes in illumination. *IEEE Trans. Pattern Analy. Machine Intell. 29*, 1642–1646.

O'TOOLE, A. J., PHILLIPS, P. J., AND NARVEKAR, A. 2008. Humans versus algorithms: Comparisons from the FRVT 2006. In *Proceedings of the Eighth International Conference on Automatic Face and Gesture Recognition.*

PHILLIPS, P. J., BEVERIDGE, J. R., DRAPER, B. A., GIVENS, G., O'TOOLE, A. J., BOLME, D., DUNLOP, J., LUI, Y. M., SAHIZADA, H., AND WEIMER, S. 2012. An introduction to the good, bad, and ugly challenge problem. *Image. Vision, Comput. 30*, 177–185.

PHILLIPS, P. J., BEVERIDGE, J. R., DRAPER, B. A., GIVENS, G., O'TOOLE, A. J., BOLME, D., DUNLOP, J., LUI, Y. M., SAHIBZADA, H., AND WEIMER, S. 2011. An Introduction to the good, the bad, and the ugly face recognition challenge problem. In *Proceedings of the 9th International Conference on Automatic Face and Gesture Recognition*.

PHILLIPS, P. J., FLYNN, P. J., SCRUGGS, W. T., BOWYER, K. W., CHANG, J., HOFFMANN, K., MARQUES, J., MIN, J., AND WOREK, W. 2005. Overview face recognition grand challenge results. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR). I:947–954.

PHILLIPS, P. J., JIANG, F., NARVEKAR, A., AND O'TOOLE, A. J. 2010. An other-race e_ect for face recognition algorithms. *ACM Trans. Appl. Percept. 8*.

PHILLIPS, P. J., SCRUGGS, W. T., O'TOOLE, A. J., FLYNN, P. J., BOWYER, K. W., SCHOTT, C. L., AND SHARPE, M. 2010. FRVT 2006 and ICE 2006 large-scale results. *IEEE Trans. Pattern Analy. Machine Intell. 32*, 5, 831–846.

RICE, A., PHILLIPS, P. J., NATU, V. S., AN, X., AND O'TOOLE, A. J. 2012. Unconscious use of the body in identifying the face. *J. Vision,* VSS Abstract.

Received November 2011; revised March 2012; accepted April 2012