An Other-Race Effect for Face Recognition Algorithms

P. JONATHON PHILLIPS, National Institute of Standards and Technology
FANG JIANG, ABHIJIT NARVEKAR, JULIANNE AYYAD, ALICE J. O'TOOLE, The University of Texas at Dallas

Psychological research indicates that humans recognize faces of their own race more accurately than faces of other races. This "other-race effect" occurs for algorithms tested in a recent international competition for state-of-the-art face recognition algorithms. We report results for a Western algorithm made by fusing eight algorithms from Western countries and an East Asian algorithm made by fusing five algorithms from East Asian countries. At the low false accept rates required for most security applications, the Western algorithm recognized Caucasian faces more accurately than East Asian faces and the East Asian algorithm recognized East Asian faces more accurately than Caucasian faces. Next, using a test that spanned all false alarm rates, we compared the algorithms with humans of Caucasian and East Asian descent matching face identity in an identical stimulus set. In this case, both algorithms performed better on the Caucasian faces—the "majority" race in the database. The Caucasian face advantage, however, was far larger for the Western algorithm than for the East Asian algorithm. Humans showed the standard other-race effect for these faces, but showed more stable performance than the algorithms over changes in the race of the test faces. State-of-the-art face recognition algorithms, like humans, struggle with "other-race face" recognition.

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:

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., and O'Toole, A. J. 2011. An other-race effect for face recognition algorithms. ACM Trans. Appl. Percept. 8, 2, Article 14 (January 2011), 11 pages.
DOI = 10.1145/1870076.1870082 http://doi.acm.org/10.1145/1870076.1870082

1. INTRODUCTION

The other-race effect for face recognition has been established in numerous human memory studies [Malpass and Kravitz 1969] and in meta-analyses of these studies [Bothwell et al. 1989; Meissner and Brigham 2001; Shapiro and Penrod 1986]. The effect for human perceivers can be summed up in the oft-heard phrase, "They all look alike to me." This anecdote suggests that our ability to perceive

This work was supported by funding from the Technical Support Working Group of the Department of Defense and A. O'Toole from. 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: P. J. Phillips, National Institute of Standards and Technology, 100 Bureau Dr., MS 8940 Gaithersburg MD 20899; email: jonathon@nist.gov; F. Jiang, A. Narvekar, J. Ayyad and A. O'Toole, School of Behavioral and Brain Sciences, GR4.1 The University of Texas at Dallas Richardson, TX 75083-0688; email: otoole@utdallas.edu.

DOI 10.1145/1870076.1870082 http://doi.acm.org/10.1145/1870076.1870082

the unique identity of other-race faces is limited relative to our ability to perceive the unique identity of faces of our own race. Although humans have additional social prejudices that impact our ability to recognize other-race faces [Levin 1996, 2000; Slone et al. 2000], perceptual factors seem to be the primary cause of the other-race effect in humans [Bryatt and Rhodes 1998; O'Toole et al. 1994; Walker and Tanaka 2003]. These factors begin to develop early in infancy and stem from the amount and quality of experience we have with faces of different races [Kelly et al. 2007]. In fact, the other-race effect in humans can be measured in infants as a decrease in their ability to detect differences in individual other-race faces as early as 3 to 9 months of age [Kelly et al. 2007]. The decrease in other-race face perception occurs at the same time that infants are making impressive gains in distinguishing faces of their own race. Thus, it has been argued that human deficiencies in perceiving other-race faces may be a consequence of neural feature selection processes that begin early in infant development. These processes serve to optimize the encoding of unique features for the types of faces we encounter most frequently—usually faces of our own race. The cost of this optimization is a perceptual filter that limits the quality of representations that can be formed for faces that are not well described by these features [Nelson 2001; Sangrigoli et al. 2005; Kuhl et al. 1992].

The rationale for testing face recognition algorithms for an other-race effect is based on the following premises. First, many face recognition algorithms include training procedures aimed at optimally representing individual faces [Furl et al. 2002; Moon and Phillips 2001]. Second, the databases used for training different algorithms vary in the extent to which they represent human demographic categories. Thus, there is reason to be concerned that some of the underlying causes of the other-race effect in humans might apply to algorithms as well. Although face recognition algorithms have been tested extensively for performance stability across environmental context variables including viewpoint, illumination, and image resolution (e.g., Gross et al. [2005] and Phillips et al. [2005, 2000, 2003]), the question of performance stability over population demographics has received much less attention [Furl et al. 2002; Givens et al. 2004; Beveridge et al. 2008]. Furthermore, no studies have examined algorithm performance as a function of the interaction between the demographic origin of the algorithm (i.e., where it was developed) and the demographics of the population to be recognized. Understanding the stability of algorithm performance for populations of faces that vary in demographics is critical for predicting face recognition accuracy when application venues vary in their demographic structure.

In this study, performance for algorithms and humans was assessed on an identity matching task. Specifically, an algorithm or human is presented with two face images and must respond with a measure of confidence to indicate whether the faces are the same person or different people. In biometrics, this is referred to as a verification task. We compared the performance of an East Asian algorithm and a Western algorithm matching identity in pairs of Caucasian and East Asian faces. The East Asian algorithm was a fusion of five algorithms from East Asian countries, and the Western algorithm was a fusion of eight algorithms from Western countries. The Face Recognition Vendor Test 2006 (FRVT 2006) (see later discussion) [Phillips et al. 2010] served as the source of the algorithms that contributed to the fusions.

The present study consists of two experiments that probe the ability of algorithms developed in East Asian and Western countries to identify East Asian and Caucasian faces. In Experiment 1, the East Asian and Western algorithms matched face identity in all available East Asian and Caucasian face pairs from the FRVT 2006 database [Phillips et al. 2010]. In this first test, we focused on a range of low false accept rates typical for security applications. In Experiment 2, we benchmarked the performance of the East Asian and Western fusion algorithms against the performance of humans of Caucasian and East Asian descent. This comparison was carried out using a smaller number of face pairs that allowed for a direct comparison among humans and the two algorithms. The face pairs were selected to control for demographic factors other than race. The second experiment measured performance using A', a

nonparametric statistic, which is a more general measure that characterizes performance over the full range of false accept rates. The test of humans serves as a control condition to confirm the other-race effect for the face dataset used in these experiments. It also provides a baseline measure of human accuracy and recognition stability over a change in the race of the test population. This measure can be used to benchmark algorithm stability over demographic change.

2. EXPERIMENT 1

The purpose of this first experiment is to determine whether the algorithms tested in the FRVT 2006 show an other-race effect. Specifically, we ask whether the geographic origin of the algorithm (i.e., where it was developed) affects its accuracy in recognizing faces of different races. We hypothesize that algorithms will show a performance advantage for faces that characterize the majority race from the geographic region of their origin.

2.1 Methods

The FRVT 2006 was the source of the algorithm data and face images for this comparison [Phillips et al. 2010]. The National Institute of Standards and Technology (NIST) sponsored, U.S. Government test of face recognition algorithms was open to academic and corporate researchers worldwide [Phillips et al. 2010]. Algorithms in the FRVT 2006 competition were required to match facial identity in 568,633,560 pairs of images over five experiments on still face images. *Match pairs* consisted of two images of the same person and *nonmatch* pairs consisted of two images of different people. In this study, we focused on face pairs from one of the FRVT 2006 experiments where the images varied in illumination conditions and where there was a sufficient number of Caucasian and East Asian faces.[1] Specifically, one image in the pair was taken under controlled illumination (e.g., under studio lighting) and the other image was taken under uncontrolled illumination (e.g., in a corridor). The uncontrolled illumination images are 2,272 $\times$ 1,704 pixels, and the controlled illumination images are 1,704 $\times$ 2,272 pixels. Example image pairs for the East Asian and Caucasian faces appear in Figure 1.

2.2 Algorithms

Algorithms participating in the FRVT 2006 could be divided into those submitted by research groups from East Asia and those submitted by research groups from Western countries (Western Europe and North America). Five algorithms were submitted by research groups in East Asia (two algorithms from China, two algorithms from Japan, and one algorithm from Korea) and eight algorithms were from research groups in Western countries (two algorithms from France, four algorithms from Germany, and two algorithms from the United States). We report performance for the average of the East Asian algorithms—an *East Asian fusion algorithm*, and for the average of the Western algorithms—a *Western fusion algorithm*. (Details on the performance of individual algorithms are available elsewhere [Phillips et al. 2010]).

The task of the individual algorithms was to compare identity in pairs of face images consisting of a controlled and an uncontrolled illumination image. Identity comparisons were based on the computed similarity scores between the controlled and uncontrolled illumination images. The similarity score for any given algorithm on a pair of faces represents the algorithm's estimate of whether the images are of the same person or of different people. High similarity scores indicate higher confidence that the two images are of the same person, and low similarity scores indicate more likelihood that the people are different. Each algorithm generates a similarity score for all possible pairs of controlled and

[1]The algorithm results used in this study are from the very–high-resolution still face images in the uncontrolled illumination experiment (Section 5.3, [Phillips et al. 2010]).

Fig. 1. Example of controlled (left) and uncontrolled (right) illumination images.

uncontrolled illumination images. This yields a matrix of similarity scores where element $s_{i,j}$ of the matrix contains the similarity between the i^{th} controlled illumination image and the j^{th} uncontrolled illumination image. The source data for Experiment 1 were based on each algorithms' matrix of similarity scores for all available East Asian face pairs age 18 to 35 ($n = 205{,}114$; 4,858 match pairs and 200,256 nonmatch pairs) and all available Caucasian face pairs age 18 to 35 ($n = 3{,}359{,}404$; 13,812 match pairs and 3,345,592 nonmatch pairs). These scores were extracted from the similarity score matrix computed by each algorithm for the FRVT 2006.

The algorithms from East Asian and Western countries were fused separately in a two-step process. In the first step, for each algorithm, the median and the median absolute deviation (MAD) were estimated from 6,849 out of 7,007,032 similarity scores ($median_k$ and MAD_k are the median and MAD for algorithm k).[2] The median and MAD were estimated from 6,849 similarity scores to avoid overtuning the estimates to the data. The similarity scores were selected to evenly sample the images in the experiment. The fused similarity scores are the sum of the individual algorithm similarity scores after the median has been subtracted and then divided by the MAD. If s_k is a similarity score for algorithm k and s_f is a fusion similarity score, then $s_f = \sum_k (s_k - \text{median}_k)/\text{MAD}_k$.

[2]The parameters for the fusion formula were computed from a subset of the similarity scores rather than on the complete set of similarity scores. This was done with the goal of generating a fusion formula that would generalize to additional faces or algorithm data, rather than being overly tuned to this particular dataset. In the algorithm evaluations carried out by NIST, the commonly applied procedure is to combine data with a method that has the ability to generalize.

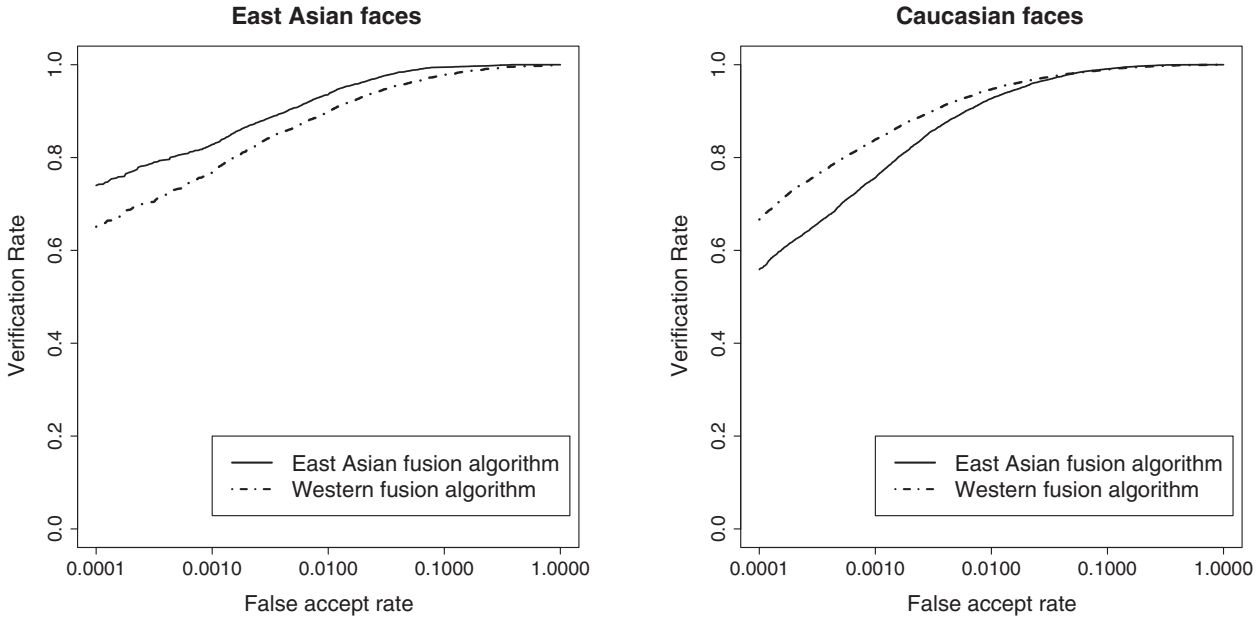

Fig. 2. ROC of the East Asian fusion and Caucasian fusion algorithms on the Experiment 1 dataset. The horizontal axis is on a logarithmic scale to emphasis low false accept rates typical of security applications. The East Asian fusion algorithm is more accurate with East Asian face pairs, and the Western fusion algorithm is more accurate with Caucasian face pairs. The effect is most pronounced at lower false accept rates, where security applications commonly operate.

2.3 Results

Figure 2 shows the receiver operating characteristic (ROC) curve for the performance of the algorithms. The ROC plots the trade-off between the verification rate and the false accept rate as a threshold is varied. A false accept occurs when an algorithm incorrectly states that the faces of two different people are the same person. A successful verification occurs when an algorithm correctly accepts that two faces are from the same person. The curve is plotted using a logarithmic scale on the horizontal axis to highlight performance in the range of the low false accept rates required for security applications. A classic "other-race effect" is evident. The East Asian fusion algorithm is more accurate at recognizing the East Asian faces, and the Western fusion algorithm is more accurate on the Caucasian faces. There is also an advantage for East Asian faces, consistent with both uncontrolled and controlled face matching studies in the literature [Givens et al. 2004; Beveridge et al. 2008; Grother 2004]. As we will see, this advantage may be primarily limited to the low false accept rate operating points common in security applications.

3. EXPERIMENT 2

In Experiment 2, we carried out a direct comparison between humans of East Asian and Caucasian descent and the East Asian and Western fusions algorithms. In the first experiment, we found an other-race effect for the East Asian and Caucasian fusion algorithms using all available pairs of East Asian and Caucasian face pairs. One limitation of using all available pairs of faces is that the people in the pairs may have differed on characteristics other than race (e.g., gender and age). In the second experiment, we used a smaller set of face pairs that were matched carefully for demographic characteristics other than race. We measured human and algorithm performance using the area under the

ROC (AUC). The AUC is a general measure of performance that summarizes a ROC across all false accept rates.

3.1 Stimuli

Forty pairs of Asians (20 match pairs and 20 nonmatch pairs; 16 female and 24 male pairs) and 40 Caucasian pairs (20 match pairs and 20 nonmatch pairs; 16 female and 24 male pairs) were used in the experiment. Following the procedure in the FRVT 2006 human performance studies, selected face pairs were rated as having medium difficulty; for example, approximately half the algorithms matched the identity correctly [Phillips et al. 2010; O'Toole et al. 2008]. Face pairs in the experiment excluded mismatched gender and retained only young adult faces (i.e., 18 to 35 years old).

3.2 Human Experimental Methods

3.2.1 *Human 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 26 students (19 females and 7 males) participated in the experiment. There were 16 Caucasians (11 female and 5 male) and 10 East Asians (8 female and 2 male).

3.2.2 *Procedure.* In the experiment, human participants were asked to match the identity of people in pairs of face images. On each trial, an image pair was displayed on the computer screen for 2 seconds, followed by a prompt asking the participant to respond as follows, "1.) sure they are the same; 2.) think they are the same; 3.) do not know; 4.) think they are not the same; and 5.) sure they are not the same." The next trial proceeded after a response was entered. Participants matched East Asian face pairs in one block of 40 trials and the Caucasian face pairs in another block of 40 trials. Half of the participants were tested with the East Asian faces first and Caucasian faces second. The remaining subjects were tested with the blocks in the reverse order.

3.3 Results

3.3.1 *Human Behavioral Data.* The performance of the East Asian and Caucasian participants on the East Asian and Caucasian faces was measured by tallying their responses to the match and nonmatch pairs. For the purpose of measuring statistical significance, we computed each subjects' A' for discriminating match versus nonmatch pairs for the Caucasian and East Asian faces. The statistic A' is used commonly in the psychology literature as a nonparametric estimate of area under the ROC curve derived from human certainty data [Macmillan and Creelman 1991]. In each condition, A' was computed from the subjects' rating responses as follows. Responses 1 and 2 were deemed "same person" judgments and responses 3, 4, and 5 were deemed "different person" judgments.[3] For each subject, the verification rate (i.e., hit rate) was computed as the proportion of face pairs correctly judged to be "same" when the face pair presented images of the same person. The false acceptance rate (i.e., false accept rate) was calculated as the proportion of face pairs that were incorrectly judged to be the same, when they were images of two different people. The statistic A' was then computed from the hit and false accept rates as

$$\frac{1}{2} + \left[\frac{(H-F)(1+H-F)}{4H(1-F)}\right], \tag{1}$$

[3]Note that a reasonable alternative measure is to assign ratings 1, 2, and 3 to the category of hits. We computed all results in this alternative fashion and found the same pattern of results.

where H is the hit rate and F is the false accept rate [Pollack and Norman 1964].[4] Analogous to the AUC measure for algorithms, this formula gives a score of 1 for perfect performance and .50 for chance performance.

The A' values for East Asian and Caucasian face identification were then used to compute a partially repeated-measures analysis of variance (ANOVA) with race of participant (East Asian or Caucasian) as a between-subjects factor and race of the face (East Asian or Caucasian) as a within-subjects factor. Evidence for the other-race effect was found in the form of a significant interaction between the race of the subject and the race of the face, $F(1, 24) = 6.15$, $p < .02$. The pattern of this interaction appears in Figure 3 (left side) and shows a substantial Caucasian face advantage for Caucasian participants and a slight East Asian face advantage for East Asian participants.

This two-factor ANOVA also makes it possible to test for main effects of the race of the subject and the race of the face. A main effect of subject race might indicate that either the East Asian or Caucasian subjects were more accurate overall. A main effect of face race might indicate that either the East Asian or Caucasian faces were inherently easier to identify. Notably, no statistically significant effects were found either for the race of the subject or for the race of the face, indicating that there was no statistical difference in the accuracy of East Asian and Caucasian participants and no statistical difference in accuracy of participants overall for the East Asian and Caucasian faces. We note, however, that the interaction is tilted slightly (though not significantly) in favor of accuracy on Caucasian faces. We will consider this result shortly in the Conclusion, taking into account the algorithm results.

3.4 Algorithm Methods

Next, we compared algorithm and human accuracy on these face pairs. Algorithm performance was assessed by extracting similarity scores from the East Asian and Western fusion algorithms for the same set of face pairs presented to human participants. On these face pairs, the AUC was computed for the East Asian and Western fusion algorithms on the East Asian face pairs and the Caucasian face pairs.

3.5 Results

The AUC values for the algorithms are shown on the right side of Figure 3 for comparison with human performance. Both the East Asian and Western fusion algorithms were more accurate with the Caucasian face pairs than with the East Asian face pairs. However, the accuracy advantage for Caucasian face pairs is far larger for the Western fusion algorithm than for the East Asian algorithm. This result is consistent with an other-race effect, but one that is superimposed on a Caucasian face advantage. Thus, the results differ from human performance in the general advantage seen for the Caucasian faces, but are similar to human performance in the interaction seen between the demographic origin of the algorithm and the race of faces in the test set. It is also worth noting that the Western and East Asian algorithms show a larger difference in performance for the two test populations than the humans. This indicates that the performance of the algorithms is less stable over race change than the performance of humans.

4. CONCLUSION

The primary conclusion of this work is that demographic origin of face recognition algorithms and the demographic composition of a test population interact to affect the accuracy of the algorithms. At

[4] A' is the nonparametric version of d'. We use A' here as it approximates the area under the ROC statistic used for the algorithms. Note, however, that d' yielded the same pattern of results and statistical effects.

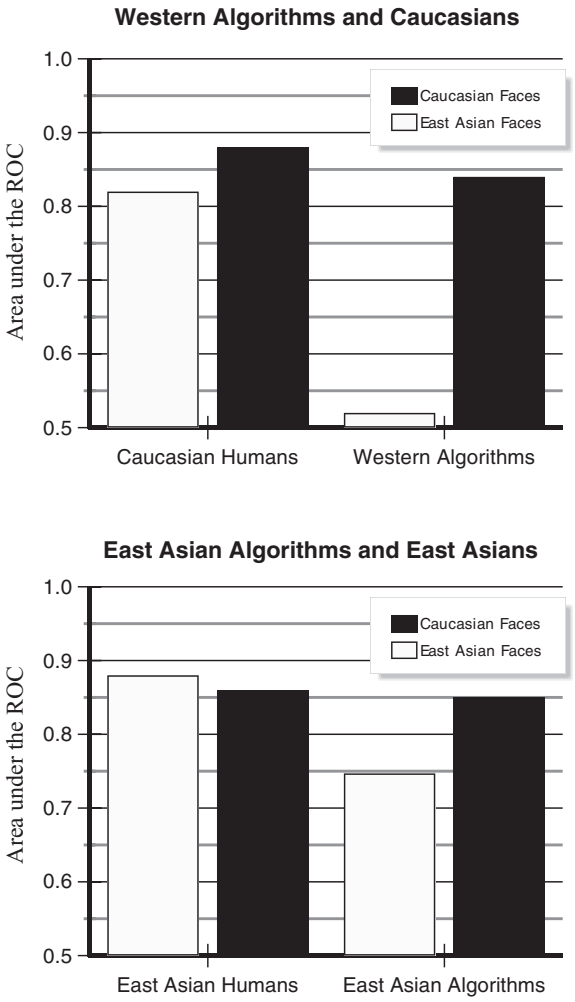

Fig. 3. Performance of humans and algorithms on the East Asian and Caucasian face pairs in Experiment 2, measured with AUC for algorithms and the A' estimate of AUC for humans. Accuracy for Caucasian subjects and the Western fusion algorithm (top) and for the East Asian subjects and the East Asian fusion algorithm (bottom) show an other-race effect.

its core, this finding indicates that algorithm performance varies over changes in population demographics. As noted, the variability of algorithm performance over changes in viewing parameters such as illumination and pose has been well studied previously. The results of the present study indicate that stability over demographics should also be included in measures of algorithm robustness. Specifically, algorithm evaluations for suitability in particular application venues should be made using a test population with comparable demographics.

Although it is clear from these experiments that algorithm performance estimates made using populations with different demographics do not converge, understanding the mechanisms behind this finding is challenging. This is due primarily to the fact that the algorithms evaluated in the FRVT 2006 were submitted to the NIST as executables, with no access to source code or to the training sets incorporated by the algorithms during development (note that several of the algorithms submitted to

the NIST test are proprietary). Using the human behavioral literature as a guide, and the human findings with this set of test faces, however, we can consider possible causes of the findings that might be common to humans and algorithms.

To understand the complete pattern of results, we start by comparing the other-race effect for the humans and for the algorithms tested. In all three cases, the effect is defined by an interaction between the race of the face and the race (demographic origin) of the participants (algorithms). A complete crossover interaction was found for the algorithms in Experiment 1 when all available face pairs were tested and when the results focused on low false accept rates. There was also an advantage for East Asian faces. By complete crossover, we mean that the East Asian algorithm was better on East Asian face pairs and the Western algorithm was better on Caucasian faces. This symmetry of crossover defines a "classic other-race effect." In Experiment 2, humans showed a lopsided interaction, with Caucasian subjects substantially more accurate with Caucasian faces and East Asian subjects slightly more accurate with East Asian faces. The algorithms in Experiment 2 showed an interaction, tilted in favor of Caucasian faces, but with a very large Caucasian face advantage for the Western algorithm and a smaller Caucasian face advantage for the East Asian algorithm. All three results indicate an other-race effect.

In addition to these other-race findings, face race, *per se*, is also an important performance factor in the two experiments. In Experiment 1, performance was reported at low false accept rates using all available face pairs (including some with mismatched demographics, for example, gender and age). This experiment showed an advantage for East Asian faces, consistent with a recent study on the FRVT 2006 data where only matched face pairs were considered [Beveridge et al. 2008]. Combined, these findings suggest that the low false accept criterion is the primary cause of the East Asian face advantage.

In Experiment 2, where performance was reported over the full range of false accept rates, there is some evidence for a Caucasian face advantage. This was supported both by the Caucasian face advantage seen for the algorithms in Experiment 2 and by the larger other-race deficit Caucasian participants showed in comparison to East Asian participants. This Caucasian advantage could occur potentially if Caucasian faces are inherently more discriminable than East Asian faces. However, data from meta-analyses on human behavioral studies of the other-race effect [Shapiro and Penrod 1986; Meissner and Brigham 2001] show no evidence for inherent discriminability differences for faces of different races—making this an unlikely general explanation for the results. For individual test sets, however, there may be small differences in the discriminability of different races of faces. The slight (nonstatistically significant) advantage human participants showed for the Caucasian faces combined with the overall advantage of the algorithms on the Caucasian faces in Experiment 2 is consistent with the possibility that the Caucasian faces from this particular dataset were inherently easier to discriminate than the East Asian faces.

An equally valid alternative possibility, however, is that both the humans and algorithms had somewhat "more experience" with Caucasian faces than with East Asian faces. As noted, the human participants in this experiment were of East Asian and Caucasian descent, but were recruited from a university in the United States in a city where Caucasians comprise the majority of the local population. In fact, most behavioral studies of the other-race effect are conducted with participants of two different races from the same local venue where one of the two races is the local majority race. In these cases, the other-race effect is commonly superimposed on a small local face race advantage.

For algorithms, "experience" refers to the amount and nature of training employed. The relevant component of experience for this study is the extent to which algorithms were trained with different races of faces. On this question, we have no direct knowledge, but we know the following about data availability. All of the research groups included in the fusion algorithms tested here participated in

the Face Recognition Grand Challenge (FRGC). The FRGC preceded the FRVT 2006 by 2 years and one of the datasets used in the FRVT 2006 was collected at the same site as the FRGC dataset. The test faces for the FRGC and FRVT 2006 comprised mutually exclusive sets of the faces from the high-resolution database developed for these large-scale tests [Phillips et al. 2005]. The FRGC dataset was composed of a strong majority of Caucasian faces (70%) and a minority of East Asian faces (22%). It is probable that all of the algorithms made use of the FRGC training faces in preparing for the FRVT 2006; therefore, all of the algorithm had some experience with Caucasian faces. In addition, training procedures implemented by the developers of the East Asian algorithms prior to the FRGC and FRVT 2006 may have included more East Asian faces than training procedures implemented by the developers of the Western algorithms. Thus, analogous to the East Asian participants in the behavioral experiments, the experience of the East Asian algorithms might have been based on both Caucasian and East Asian faces. Analogous to the Caucasian participants, experience for the Western algorithm training may have strongly favored Caucasian faces.

In conclusion, the performance of state-of-the-art face recognition algorithms varies as a joint function of the demographic origin of the algorithm and the demographic structure of the test population. This result is analogous to findings for human face recognition. The mechanisms underlying the other-race effect for humans are reasonably well understood and are based in early experience with faces of different races. Although our hypotheses about the mechanisms underlying the algorithm effects are still tentative, the effects we report are not. The present results point to an important performance variable combination that has not received much attention. The results also suggest a need to understand how the ethnic composition of a training set impacts the robustness of algorithm performance. Finally, from a practical point of view, recent studies indicate that algorithms are now capable of surpassing human performance matching face images across changes in illumination [O'Toole et al. 2007, 2008] and on the task of recognition from sketches [Tang and Wang 2003; 2004]. This increases the likelihood that face recognition algorithms will find new real-world applications in the near future. In these cases, there is a pressing need to test algorithms intended for applications in venues with highly diverse target populations using face sets that match statistics of the demographics expected in these venues.

REFERENCES

BEVERIDGE, J. R., GIVENS, G. H., PHILLIPS, P. J., DRAPER, B. A., AND LUI, Y. M. 2008. Focus on quality, predicting FRVT 2006 performance. In *Proceeding of the 8th International Conference on Automatic Face and Gesture Recognition*. IEEE, Los Alamitos, CA.

BOTHWELL, R. K., BRIGHAM, J. C., AND MALPASS, R. S. 1989. Cross-racial identification. *Pers. Social Psychol. Bull. 15*, 19–25.

BRYATT, G. AND RHODES, G. 1998. Recognition of own-race and other-race caricatures: Implications for models of face recognition. *Vision Res. 38*, 2455–2468.

FURL, N., PHILLIPS, P. J., AND O'TOOLE, A. J. 2002. Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitve Sci. 26*, 797–815.

GIVENS, G. H., BEVERIDGE, J. R., DRAPER, B. A., GROTHER, P. J., AND PHILLIPS, P. J. 2004. How features of the human face affect recognition: A statistical comparison of three face recognition algorithms. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*. IEEE, Los Alamitos, CA, 381–388.

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.

GROTHER, P. 2004. Face recognition vendor test 2002: Supplemental report. Tech. rep. NISTIR 7083, National Institute of Standards and Technology. http://www.frvt.org.

KELLY, D. J., QUINN, P. C., SLATER, A. M., LEE, K., GE, L., AND PASCALIS, O. 2007. The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychol. Sci. 18*, 1084–1089.

KUHL, P. K., WILLIAMS, K. H., AND LACERDO, F. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science 225*, 606–608.

LEVIN, D. 1996. Classifying faces by race: The structure of face categories. *J. Exp. Psychol. 22*, 1364–1382.

LEVIN, D. 2000. Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *J. Exp. Psychol. 129*, 559–574.

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

MALPASS, R. S. AND KRAVITZ, J. 1969. Recognition for faces of own and other race faces. *J. Pers. Soc. Psychol. 13*, 330–334.

MEISSNER, C. A. AND BRIGHAM, J. C. 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychol. Public Policy Law 7*, 3–35.

MOON, H. AND PHILLIPS, P. J. 2001. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception 30*, 303–321.

NELSON, C. A. 2001. The development and neural bases of face recognition. *Infant Child Dev. 10*, 3–18.

O'TOOLE, A. J., DEFFENBACHER, K. A., VALENTIN, D., AND ABDI, H. 1994. Structural aspects of face recognition and the other-race effect. *Memory & Cognition 22*, 2, 208–224.

O'TOOLE, A. J., PHILLIPS, P. J., JIANG, F., AYYAD, J., P_ENARD, N., AND ABDI, H. 2007. Face recognition algorithms surpass humans matching faces across changes in illumination. *IEEE Trans. Patt. Anal. 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 8th International Conference on Automatic Face and Gesture Recognition*. IEEE, Los Alamitos, CA.

PHILLIPS, P. J., FLYNN, P. J., SCRUGGS, T., BOWYER, K. W., CHANG, J., HOFFMAN, K., MARQUES, J., MIN, J., AND WOREK, W. 2005. Overview of the face recognition grand challenge. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 947–954.

PHILLIPS, P. J., GROTHER, P. J., MICHEALS, R. J., BLACKBURN, D. M., TABASSI, E., AND BONE, J. M. 2003. Face recognition vendor test 2002: Evaluation report. Tech. rep. NISTIR 6965, National Institute of Standards and Technology. http://www.frvt.org.

PHILLIPS, P. J., MOON, H., RIZVI, S., AND RAUSS, P. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Patt. Anal. Machine Intell. 22*, 1090–1104.

PHILLIPS, P. J., SCRUGGS, W., OTOOLE, 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 Anal. Mach. Intell. 32*, 831–846.

POLLACK, I. AND NORMAN, D. 1964. A non-parametric analysis of recognition experiments. *Psychonomic Sci. 1*, 125–126.

SANGRIGOLI, S., PALLIER, C., ARGENTI, A. M., VENTUREYRA, V. A. G., AND DE SCHONEN, S. 2005. Reversibility of the other-race effect in face recognition during childhood. *Psychol. Sci. 16*, 440–444.

SHAPIRO, P. N. AND PENROD, S. D. 1986. Meta-analysis of face identification studies. *Psychol. Bull. 100*, 139–156.

SLONE, A. E., BRIGHAM, J. C., AND MEISSNER, C. A. 2000. Social and cognitive factors affecting the own-race bias in whites. *Basic Appl. Social Psychol. 22*, 71–84.

TANG, X. AND WANG, X. 2003. Face sketch synthesis and recognition. In *Proceedings of the 9th International Conference on Computer Vision*. IEEE, Los Alamitos, CA, 687–694.

TANG, X. AND WANG, X. 2004. Face sketch recognition. *IEEE Trans. Circuits Syst. Video Technol.* 50–57.

WALKER, P. M. AND TANAKA, J. W. 2003. An encoding advantage for own-race versus other-race faces. *Perception 32*, 1117–1125.

Received March 2010; revised June 2010; accepted July 2010