Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms

P. Jonathon Phillips[a,1], Amy N. Yates[a], Ying Hu[b], Carina A. Hahn[b], Eilidh Noyes[b], Kelsey Jackson[b], Jacqueline G. Cavazos[b], Géraldine Jeckeln[b], Rajeev Ranjan[c], Swami Sankaranarayanan[c], Jun-Cheng Chen[d], Carlos D. Castillo[d], Rama Chellappa[c], David White[e], and Alice J. O'Toole[b]

[a]Information Access Division, National Institute of Standards and Technology, Gaithersburg, MD 20899; [b]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX 75080; [c]Department of Electrical and Computer Engineering, University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20854; [d]University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20854; and [e]School of Psychology, The University of New South Wales, Sydney, NSW 2052, Australia

Edited by Thomas D. Albright, The Salk Institute for Biological Studies, La Jolla, CA, and approved April 30, 2018 (received for review December 13, 2017)

Achieving the upper limits of face identification accuracy in forensic applications can minimize errors that have profound social and personal consequences. Although forensic examiners identify faces in these applications, systematic tests of their accuracy are rare. How can we achieve the most accurate face identification: using people and/or machines working alone or in collaboration? In a comprehensive comparison of face identification by humans and computers, we found that forensic facial examiners, facial reviewers, and superrecognizers were more accurate than fingerprint examiners and students on a challenging face identification test. Individual performance on the test varied widely. On the same test, four deep convolutional neural networks (DCNNs), developed between 2015 and 2017, identified faces within the range of human accuracy. Accuracy of the algorithms increased steadily over time, with the most recent DCNN scoring above the median of the forensic facial examiners. Using crowd-sourcing methods, we fused the judgments of multiple forensic facial examiners by averaging their rating-based identity judgments. Accuracy was substantially better for fused judgments than for individuals working alone. Fusion also served to stabilize performance, boosting the scores of lower-performing individuals and decreasing variability. Single forensic facial examiners fused with the best algorithm were more accurate than the combination of two examiners. Therefore, collaboration among humans and between humans and machines offers tangible benefits to face identification accuracy in important applications. These results offer an evidence-based roadmap for achieving the most accurate face identification possible.

face identification | forensic science | face recognition algorithm | wisdom-of-crowds | machine learning technology

Societies rely on the expertise and training of professional forensic facial examiners, because decisions by professionals are thought to assure the highest possible level of face identification accuracy. If accuracy is the goal, however, the scientific literature in psychology and computer vision points to three additional approaches that merit consideration. First, untrained "superrecognizers" from the general public perform surprisingly well on laboratory-based face recognition studies (1). Second, wisdom-of-crowds effects for face recognition, implemented by averaging individuals' judgments, can boost performance substantially over the performance of a person working alone (2–5). Third, computer-based face recognition algorithms over the last decade have steadily closed the gap between human and machine performance on increasingly challenging face recognition tasks (6, 7).

Beginning with forensic facial examiners, remarkably little is known about their face identification accuracy relative to people without training, and nothing is known about their accuracy relative to computer-based face recognition systems. Independent and objective scientific research on the accuracy of forensic facial practitioners began in response to the National Research Council report *Strengthening Forensic Science in the United States: A Path Forward* (8; cf. ref. 9). In the most comprehensive study to date (3), forensic facial examiners were superior to motivated control participants and to students on six tests of face identity matching. However, image pairs in these tests appeared for a maximum of 30 s. Identification decisions in a forensic laboratory typically require days or weeks to complete and are made with the assistance of image measurement and manipulation tools (10). Accordingly, the performance of forensic facial examiners in ref. 3 represents a lower-bound estimate of the accuracy of examiners in practice.

Superrecognizers are untrained people with strong skills in face recognition. Multiple laboratory-based face recognition tests of these individuals indicate that highly accurate face identification can be achieved by people with no professional training (1). Superrecognizers contribute to face recognition decisions made in law enforcement (11, 12) but have not been compared with forensic examiners or machines.

The term wisdom-of-crowds refers to accuracy improvements achieved by combining the judgments of multiple individuals to make a decision. Face recognition accuracy by humans can be boosted substantially by crowd-sourcing responses (2–5),

Significance

This study measures face identification accuracy for an international group of professional forensic facial examiners working under circumstances that apply in real world casework. Examiners and other human face "specialists," including forensically trained facial reviewers and untrained superrecognizers, were more accurate than the control groups on a challenging test of face identification. Therefore, specialists are the best available human solution to the problem of face identification. We present data comparing state-of-the-art face recognition technology with the best human face identifiers. The best machine performed in the range of the best humans: professional facial examiners. However, optimal face identification was achieved only when humans and machines worked in collaboration.

Author contributions: P.J.P., A.N.Y., D.W., and A.J.O. designed research; R.R., S.S., J.-C.C., C.D.C., and R.C. contributed new reagents/analytic tools; P.J.P., A.N.Y., Y.H., C.A.H., E.N., K.J., J.G.C., G.J., and A.J.O. analyzed data; R.R., S.S., J.-C.C., C.D.C., and R.C. implemented and ran the face recognition algorithms; and P.J.P. and A.J.O. wrote the paper.

Conflict of interest statement: The University of Maryland is filing a US patent application that will cover portions of algorithms A2017a and A2017b. R.R., C.D.C., and R.C. are coinventors on this patent.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: jonathon@nist.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1721355115/-/DCSupplemental.

including for forensic examiners in a time-restricted laboratory experiment (3). Combining human and machine face identification judgments also improves accuracy over either one operating alone (5). The effect of fusing the judgments of professionals and algorithms has not been explored.

Computer-based face recognition systems now assist forensic face examiners by searching databases of images to generate potential identity matches for human review (13). Direct comparisons between human and machine accuracy have been based on algorithms developed before 2013. At that time, algorithms performed well with high-quality frontal images of faces with minimal changes in illumination and expression. Since then, deep learning and deep convolutional neural networks (DCNNs) have become the state of the art for face recognition (14–18). DCNNs can recognize faces from highly variable, low-quality images. These algorithms are often trained with millions of face images of thousands of people.

Our goal was to achieve the most accurate face identification using people and/or machines working alone or in collaboration. The task was to determine whether pairs of face images showed the same person or different people. Image pairs were prescreened to be highly challenging based on data from humans and computer algorithms. Images were taken with limited control of illumination, expression, and appearance. Fig. 1 shows two example pairs (all pairs are shown in *SI Appendix*, Figs. S8–S14). To provide a comprehensive assessment of human accuracy, we tested three face specialist groups (forensic facial examiners, forensic facial reviewers, and superrecognizers) and two control groups (fingerprint examiners and undergraduate students). Humans responded on a 7-point scale that varied from high confidence that the pair showed the same person (+3) to high confidence that the pair showed different people (−3). We also tested four face recognition algorithms based on DCNNs developed between 2015 and 2017. Algorithm responses were real-valued similarity scores indicating the likelihood that the images showed the same person. The five subject groups and four algorithms were tested on the same image pairs. Facial examiners, reviewers, superrecognizers, and fingerprint examiners had 3 mo to complete the test. Students took the test in a single session.

Forensic facial experts are professionals trained to identify faces in images and videos using a set of tools and procedures (10) that vary across forensic laboratories (19). We tested two classes of forensic facial professionals. Examiners ($n = 57$, 28 females, from five continents) have extensive training, and their identity comparisons involve a rigorous and time-consuming process. Their identification decisions can be presented in written documents that can be used to support legal actions, prosecutions, and expert testimony in court. Reviewers ($n = 30$, 17 females, from two continents) are trained to perform faster and less rigorous identifications that may be used in law enforcement and can assist in generating leads in criminal cases. We also tested superrecognizers ($n = 13$, 8 females, from two continents) (20), defined here as a person who had taken a standard face recognition test that qualified them as a superrecognizer (1) or as a person used professionally as a superrecognizer (e.g., the London Metropolitan Police) (*SI Appendix*, *SI Text*).

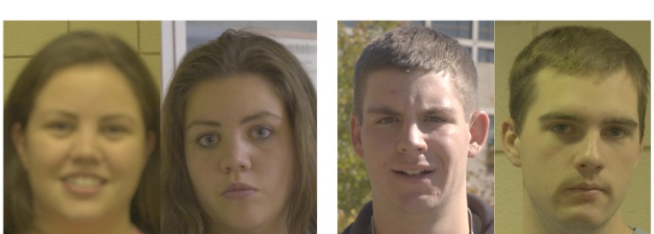

Fig. 1. Examples highlighting the face region in the images used in this study (all image pairs are shown in *SI Appendix*, Figs. S8–S14). (*Left*) This pair is a same identity pair, and (*Right*) this pair shows a different identity pair.

Professional fingerprint examiners and undergraduate students served as control groups. Fingerprint examiners ($n = 53$, 41 females, from two continents) are trained forensic professionals who perform fingerprint comparisons. They provide a baseline for forensic ability and training that excludes expertise in facial forensics. Fingerprint examiners complete extensive training for professional certification. Undergraduate students ($n = 31$, 24 females, from one continent) were tested as a proxy for the general population.

To compare humans with face recognition algorithms, four DCNNs were tested on the same stimuli judged by humans. We refer to the algorithms as A2015 (14), A2016 (15), A2017a (16), and A2017b (17). The inclusion of multiple algorithms provides a robust sample of the state of the art for automatic face recognition. To make the test comparable with humans as an "unfamiliar" face matching test, we verified that none of the algorithms had been trained on images from the dataset used for the human test. Note that A2015 can be downloaded from the web and therefore, provides a public benchmark algorithm.

Results

Accuracy. Fig. 2 shows performance of the subject groups and algorithms using the area under the receiver operating characteristic curve (AUC) as a measure of accuracy. The groups are ordered by AUC median from the most to least accurate: facial examiners (0.93), facial reviewers (0.87), superrecognizers (0.83), fingerprint examiners (0.76), and students (0.68). Algorithm performance increased monotonically from the oldest algorithm (A2015) to the newest algorithm (A2017b). Comparing the algorithms with the human groups, the publicly available algorithm (A2015) performed at a level similar to the students (0.68). Algorithm A2016 performed at the level of fingerprint examiners (0.76). Algorithm A2017a performed at a level (0.85) comparable with the superrecognizers (0.83) and reviewers (0.87). The performance of A2017b (0.96) was slightly higher than the median of the facial examiners (0.93).

More formally, all face specialist groups surpassed fingerprint examiners (facial examiners, $P = 2.14 \times 10^{-6}$; facial reviewers, $P = 0.004$; superrecognizers, $P = 0.017$). The face specialist groups also surpassed students (facial examiners, $P = 2.53 \times 10^{-8}$; facial reviewers, $P = 4.01 \times 10^{-6}$; superrecognizers, $P = 0.0005$) (*SI Appendix*, *SI Text*). Performance across the face specialist groups did not differ statistically. Summary statistics for accuracy, however, should be interpreted in the context of the full performance distributions within each group.

Performance Distributions. Individual accuracy varied widely in all groups. All face specialist groups (facial examiners, reviewers, and superrecognizers) had at least one participant with an AUC below the median of the students. At the top of the distribution, all but the student group had at least one participant with no errors. To examine specialist groups in the context of the general population (students), we fit a Gaussian distribution to the student AUCs (*SI Appendix*, *SI Text*). Next, we computed the fraction of participants in each group who scored above the 95th percentile (Fig. 2, dashed line). For the facial examiner group, 53% were above the 95th percentile of students; for the facial reviewers, this proportion was 36%. For superrecognizers, it was 46%, and for fingerprint examiners, it was 17%. For the algorithms, the accuracy of A2017b was higher than the majority (73%) of participants in the face specialist groups. Conversely, 35% of examiners, 13% of reviewers, and 23% of superrecognizers were more accurate than A2017b. Compared with students, the accuracy of A2017b was equivalent to a

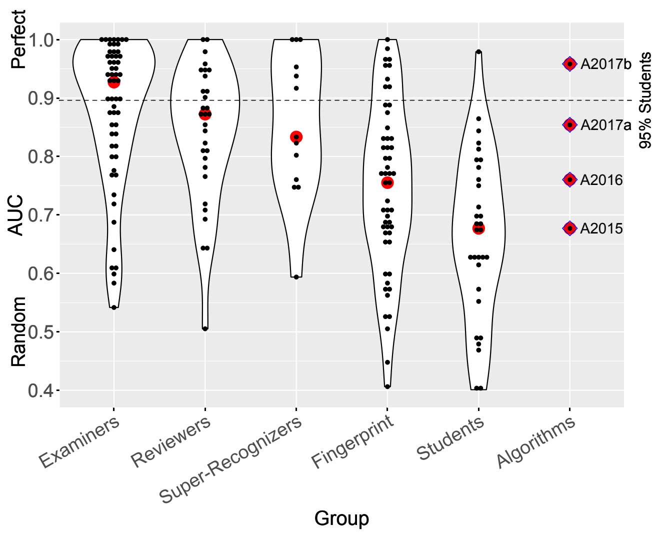

Fig. 2. Human and machine accuracy. Black dots indicate AUCs of individual participants; red dots are group medians. In the algorithms column, red dots indicate algorithm accuracy. Face specialists (facial examiners, facial reviewers, and superrecognizers) surpassed fingerprint examiners, who surpassed the students. The violin plot outlines are estimates of the density for the AUC distribution for the subject groups. The dashed horizontal line marks the accuracy of a 95th percentile student. All algorithms perform in the range of human performance. The best algorithm places slightly above the forensic examiners' median.

student at the 98th percentile (z score = 2.090), A2017a was at the 91st percentile (z score = 1.346), A2016 was at the 76th percentile (z score = 0.676), and A2015 was at the 53rd percentile (z score = 0.082). These results show a steady increase in algorithm accuracy from a level comparable with students in 2015 to a level comparable with the forensic facial examiners in 2017.

Fusing Human Judgments. In forensic practice, it is common for multiple examiners to review an identity comparison to assure consistency and consensus (3, 5). To examine the effects of fusion on accuracy, we combined individual participants' judgments in each group. We began with one participant and increased the number of participants' judgments fused from 2 to 10. To fuse n participants, we selected n participants randomly and averaged their rating-based judgments for each image pair. For fusing judgments, averaging is generally the most effective fusion strategy (21). An AUC was then computed from these average judgments. The sampling procedure was repeated 100 times for each value of n.

Median accuracy peaked at 1.0 (no errors) with the fusion of four examiners or three superrecognizers (Fig. 3). The performance of all of the groups increased with fusion (*SI Appendix, SI Text*). For reviewers, the median peaked at 0.98 with 10 participants fused. Fingerprint examiners peaked at a median of 0.97 for 10 participants. For superrecognizers, the median increased from 0.83 to 0.98 when two superrecognizers were fused and to 1.0 when three or more superrecognizers were fused. Using a fusion perspective in comparing accuracy across participant groups, the data indicate that the median examiner (0.93) performs at a level roughly equal to two facial reviewers (median = 0.93) and seven fingerprint examiners (median = 0.94). Notably, the median of individual judgments by examiners is superior to the combination of 10 students (median = 0.88).

Fusing Humans and Machines. We examined the effectiveness of combining examiners, reviewers, and superrecognizers with algorithms. Human judgments were fused with each of the four algorithms as follows. For each face image pair, an algorithm returned a similarity score that is an estimate of how likely it is that the images show the same person. Because the similarity score scales differ across algorithms, we rescaled the scores to the range of human ratings (*SI Appendix, SI Text*). For each face pair, the human rating and scaled algorithm score were averaged, and the AUC was computed for each participant–algorithm fusion.

Fig. 4 shows the results of fusing humans and algorithms. The most effective fusion was the fusion of individual facial examiners with algorithm A2017b, which yielded a median AUC score of 1.0. This score was superior to the combination of two facial examiners (Mann–Whitney U test = 2.82×10^4, $n_1 = 1{,}596$, $n_2 = 57$, $P = 8.37 \times 10^{-7}$). Fusing individual examiners with A2017a and A2016 yielded performance equivalent to the fusion of two examiners (Mann–Whitney U test = 4.53×10^4, $n_1 = 1{,}596$, $n_2 = 57$, $P = 0.956$; Mann–Whitney U test = 4.33×10^4, $n_1 = 1{,}596$, $n_2 = 57$, $P = 0.526$, respectively). Fusing one examiner with A2015 did not improve accuracy over a single examiner (Mann–Whitney U test = 1,592, $n_1 = 57$, $n_2 = 57$, $P = 0.86$). Fusing one examiner with A2017b proved more accurate than fusing one examiner with either A2017a or A2016 (Mann–Whitney U test = 1,054, $n_1 = 57$, $n_2 = 57$, $P = 7.92 \times 10^{-4}$; Mann–Whitney U test = 942, $n_1 = 57$, $n_2 = 57$, $P = 7.28 \times 10^{-5}$, respectively). Finally, fusing one examiner with both A2017b and A2017a did not improved accuracy over fusing one examiner with A2017b (Mann–Whitney U test = 1,414, $n_1 = 57$, $n_2 = 57$, $P = 0.21$). This analysis was repeated for fusing algorithms and facial reviewers and for fusing algorithms and superrecognizers. Similar results were found for both groups (*SI Appendix, SI Text*).

Error Rates for Highly Confident Decisions

In legal proceedings, the conclusions of greatest impact are identification errors made with high confidence. These can lead to

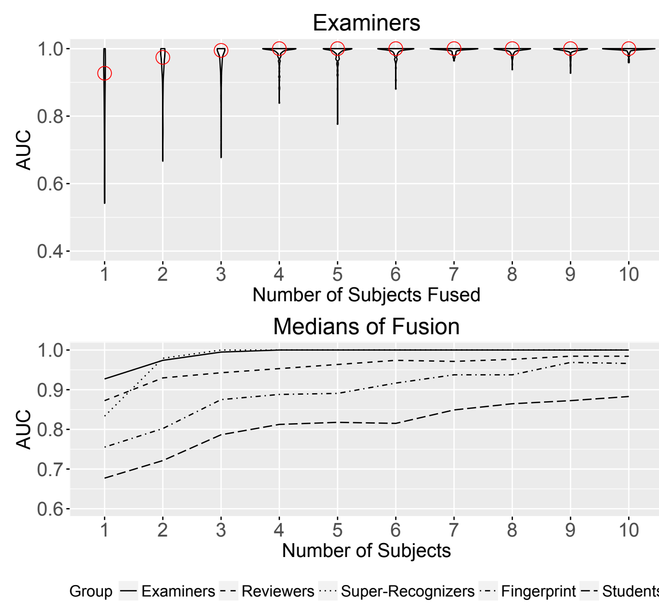

Fig. 3. Plots illustrate the effectiveness of fusing multiple participants within groups. For all groups, combining judgments by simple averaging is effective. The violin plots in *Upper* show the distribution of AUCs for fusing examiners. Red circles indicate median AUCs. In *Lower*, the medians of the AUC distributions for the examiners, reviewers, superrecognizers, fingerprint examiners, and students appear. The median AUC reaches 1.0 for fusing four examiners or fusing three superrecognizers. The median AUC of fusing 10 students was 0.88, substantially below the median AUC for individual examiner accuracy.

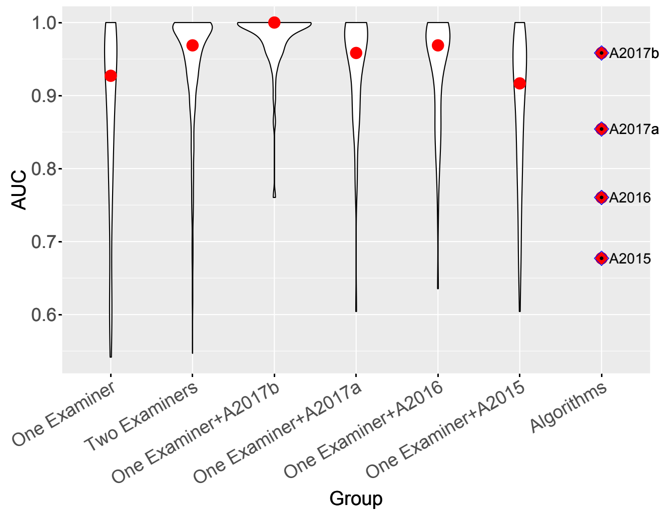

Fig. 4. Fusion of examiners and algorithms. Violin plots show the distribution of AUCs for each fusion test. Red dots indicate median AUCs. The distribution of individual examiners and the fusion of two examiners appear in columns 1 and 2. Also, algorithm performance appears in column 7. In between, plots show the forensic facial examiners fused with each of the four algorithms. Fusing one examiner and A2017b is more accurate than fusing two examiners, fusing examiners and A2017a or A2016 is equivalent to fusing two examiners, and fusing examiners with A2015 does not improve accuracy over a single examiner.

miscarriages of justice with profound societal implications. In this study, the two responses that expressed high confidence were "the observations strongly support that it is the same person" (+3) and "the observations strongly support that it is not the same person" (−3). To examine the error rates associated with judgments of +3 and −3, we computed the fraction of high-confidence same-person (+3) ratings made to different identity face pairs and estimated the error rate as a Bernoulli distribution. The Bernoulli parameter $\hat{q}$ is the fraction of different identity pairs that were given a rating of +3. Fig. 5 shows the estimated parameter $\hat{q}$ with 95% confidence intervals by participant group. (*SI Appendix*, Table S2 shows estimated Bernoulli parameters and the confidence intervals.) The analysis was also conducted on the probability of same identity pairs being assigned a −3 rating.

For facial examiners, the error rate for judging with high confidence that two different faces were the same was 0.009 (upper limit of the confidence interval, 0.022). The corresponding error rate on judging the same person as two different people was 0.018 (upper limit of confidence interval, 0.030). For facial reviewers, the corresponding error rates and confidence intervals were similar to those for the facial examiners (*SI Appendix, SI Text*). For superrecognizers, although their error rate for the rating of +3 on two different faces was comparable with that of examiners and reviewers, their error rate for −3 ratings assigned to same face image pairs was higher. Student error rates for high-confidence decisions were substantially higher than those of the facial examiners, reviewers, and superrecognizers. Notably, we found that fusion reduced high-confidence errors for facial examiners, facial reviewers, and superrecognizers (*SI Appendix, SI Text*). Specifically, fusing one individual and A2017b was superior to fusing two individuals, and fusing two individuals was superior to one individual.

One possible explanation for these results is that forensic professionals avoid extreme ratings at both ends of the scale. To test this, we examined whether forensic professionals (facial examiners, facial reviewers, fingerprint examiners) overall made fewer high-confidence responses than nonprofessionals (superrecognizers, students). For each participant, the number of high-confidence responses was computed. Analysis showed that forensic professionals made fewer high-confidence decisions than nonforensic professionals (Mann–Whitney U test = 1,966.5, $n_1 = 140$, $n_2 = 44$, $P = 2.83 \times 10^{-4}$). This is consistent with a result obtained in a previous study by Norell et al. (22), which tested police detectives and students on face identity matching experiments. The result suggests that forensic training of any kind may affect the use of the response scale to avoid errors made with high confidence.

Discussion

The results of the study point to tangible ways to maximize face identification accuracy by exploiting the strengths of humans and machines working collaboratively. First, to optimize the accuracy of face identification, the best approach is to combine human and machine expertise. Fusing the most accurate machine with individual forensic facial examiners produced decisions that were more accurate than those arrived at by any pair of human and/or machine judges. This human–machine combination yielded higher accuracy than the fusion of two individual forensic facial examiners. Computational theory indicates that fusing systems works best when their decision strategies differ (21, 23). Therefore, the superiority of human–machine fusion over human–human fusion suggests that humans and machines have different strengths and weaknesses that can be exploited/mitigated by cross-fusion.

Second, for human decisions, the highest possible accuracy is obtained when human judgments are combined by simple averaging. The power of fusing human decisions to improve accuracy is well-known in the face recognition literature (3, 4). Our results speak to the tangible benefits of putting fusion formally into the process of a forensic decision-making process. Collaborative peer review of decisions is a common strategy in facial forensics. This study suggests that, in addition to social collaboration, computationally combining multiple independent decisions made in isolation also produces solid gains in accuracy (24). Although fusing student judgments improves accuracy, we show that there are limits to the gains possible from fusion. A fusion of student judgments will not approach the accuracy of fusing facial examiners or reviewers. This suggests that a strategy for achieving optimal accuracy is to fuse people in the most accurate group of humans.

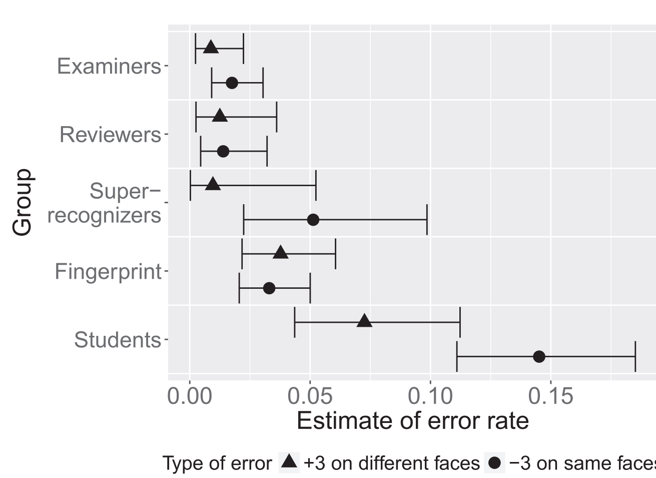

Fig. 5. Estimated probability of highly confident same person ratings (+3 judgment, strong evidence the same person) when the identities are different and estimated probability of highly confident different person ratings (−3 judgment, strong evidence different people) when the identity is the same. The 95% confidence intervals are shown.

Third, systematic differences were found for the performance of the human groups on average. Professional forensic facial examiners, professional facial reviewers, and superrecognizers were the most accurate groups. Fingerprint examiners were less accurate than the face specialists but more accurate than students. Notably, the group medians ranged from highly accurate for facial examiners ($AUC = 0.93$) to moderately above chance for students ($AUC = 0.68$). This suggests that our face matching test tapped into the entire operating range of normal human accuracy.

Fourth, the distribution of individual performance in this test was perhaps as informative as the summary data on central tendency. In particular, although the median accuracy measures strongly prescribe the use of professional facial examiners for cases where face identification accuracy is important, some individuals in this group performed poorly. Mitigating this concern to some extent, confident incorrect judgments by facial examiners were extremely rare. At the other end of the spectrum, some individuals in other groups performed with high accuracy that was well within the range of the best face specialists. Remarkably, in all but the student group, at least one individual performed the test with no errors. The range of accuracy of individuals in each group suggests the possibility of prescreening the general population for people with natural ability at face identification. The superrecognizers in our study were not trained formally in face recognition, yet they performed at levels comparable with those of the facial professionals. This suggests that both talent and training may underlie the high accuracy seen in the two groups of facial professionals.

Turning to the performance of the algorithms, the results indicate the potential for machines to contribute beneficially to the forensic process. Accuracy of the publicly available algorithm that we tested (A2015) was at the level of median accuracy of the students—modestly above chance. The other algorithms follow a rapid upward performance trajectory: from parity with a median fingerprint examiner (A2016) to parity with a median superrecognizer (A2017a) and finally, to parity with median forensic facial examiners (A2017b). There is now a decade-long effort to compare the accuracy of face recognition algorithms with humans (6). In the earliest tests (25), the face matching tasks presented relatively controlled images. As these tests progressed, algorithms and humans were compared on progressively more challenging image pairs. In this study, image pairs were selected to be extremely challenging based on both human and algorithm performance. The difficulty of these items for humans was supported by the accuracy of students, who represent a general population of untrained humans. Students performed poorly on these challenging image pairs. All four of the algorithms performed at or above median student performance. Two algorithms performed in the range of the facial specialists, and one algorithm matched the performance of forensic facial examiners.

In summary, this is the most comprehensive examination to date of face identification performance across groups of humans with variable levels of training, experience, talent, and motivation. We compared the accuracy of state-of-the-art face recognition algorithms with humans and show the benefits of a collaborative effort that combines the judgments of humans and machines. The work draws on previous cornerstone findings on human expertise and talent with faces, strategies for fusing human judgments, and computational advances in face recognition. The study provides an evidence-based roadmap for achieving highly accurate face identification. These methods should be extended in future work to test humans and machines on a wider range of face recognition tasks, including recognition across viewpoint and with low-quality images and video as well as recognition of faces from diverse demographic categories.

Materials and Methods

Test Protocol for Human Participants. To allow examiners access to their tools and methods while comparing face images, participants in all conditions, except the untrained student control group, downloaded the pairs of face images and were allowed 3 mo to complete the comparisons. For facial examiners and reviewers, comparisons were completed in their laboratory using their tools and methods. For superrecognizers and fingerprint examiners, the comparisons were done on a computer using tools available on the computer (e.g., image software tools). Students viewed the face pairs presented on a computer monitor one at a time. The size of the images was preset, and it was the same for all images. Pairs remained visible until a response was entered on the keyboard.

For each pair of face images, the participants in all subject groups were required to respond on a 7-point scale: +3, the observations strongly support that it is the same person; +2, the observations support that it is the same person; +1, the observations support to some extent that it is the same person; 0, the observations support neither that it is the same person nor that it is different persons; −1, the observations support to some extent that it is not the same person; −2, the observations support that it is not the same person; −3, the observations strongly support that it is not the same person. The wording was chosen to reflect scales used by forensic examiners in their daily work. A receiver operating characteristic curve and the AUC were computed from the ratings for each subject.

The experimental design was approved by the National Institute of Standards and Technology (NIST) IRB. Data collection procedures for students were approved by the IRB at the University of Texas at Dallas, and all subjects provided consent.

Test Protocol for Algorithms. Algorithms first encoded each face as a compact vector of feature values by processing the image with the trained DCNN. DCNNs consist of multiple layers of simulated neurons that convolute and pool input (face images), feeding the data forward to one or more fully connected layers at the top of the network. The output is a compressed feature vector that represents a face (algorithm A2015 uses 4,096 features, A2016 uses 320 features, and A2017a and A2017b use 512 features). For each image pair in the test, a similarity score was computed between the representations of the two faces. The similarity score is the algorithm's estimate of whether the images show the same person. To avoid response bias, performance was measured by computing an AUC directly from the similarity score distributions for same and different identity pairs, eliminating the need for a threshold. *SI Appendix*, *SI Text* has details on the algorithms.

Stimuli. Image pairs were chosen carefully in three screening steps. These steps were based on human and algorithm performance (details follow). The goal of the screening process was to select highly challenging image pairs that would test the upper limits of the participants' skills, while avoiding floor effects for the students. The starting point for pair selection was a set of 9,307 images of 507 individuals taken with a Nikon D70 6 megapixel single-lens reflex camera. Images were acquired during a single academic year in indoor and outdoor settings at the University of Notre Dame. Faces were in approximately frontal pose (Fig. 1 shows example pairs).

We screened for identity matching difficulty with a fusion of three top-performing algorithms from an international competition of algorithms [Face Recognition Vendor Test 2006 (FRVT 2006)] (26). Based on the results of the fusion algorithm, the images were stratified into three difficulty levels (27). Image pairs were further pruned using human experimental data. We began with the accuracy of undergraduate students on the two most difficult levels for the algorithm (28, 29). We selected the highest performing 25% of participants and chose the 84 same identity and 84 different identity image pairs that elicited the highest proportion of errors in this group. These pairs formed a stimulus pool of image pairs that were challenging for humans and previous generation face recognition algorithms. A second stimulus pool was created in a similar way but with the goal of finding image pairs on which previous generation algorithms failed systematically. We sampled the stimuli from those used in a recent study that compared human and computer algorithm performance on a special set of image pairs for which machine performance in the FRVT 2006 (26) was 100% incorrect (29). Specifically, similarity scores computed between same identity faces were uniformly lower than those computed for the different identity image pairs. Finally, we implemented a third level of stimulus screening for both stimulus pools. We used performance on an identity matching task with very short (30 s) stimulus presentation times (3) and sorted these stimuli according to difficulty for the forensic examiners from that test.

Discussions with facial examiners before the study indicated that they were willing to compare 20 pairs of images over a 3-mo period. This allowed them to spend the time that they would normally spend for a forensic comparison. Using the screening described, we chose 12 image pairs from the first stimulus pool and 8 pairs from the second. There were same ($n = 12$) and different identity ($n = 8$) pairs. The slight imbalance eliminated the use of a process of elimination strategy (*SI Appendix, SI Text*).

Data Availability. Deidentified data for facial examiners and reviewers, superrecognizers, and fingerprint examiners can be obtained by signing a data transfer agreement with the NIST. The images are available by license from the University of Notre Dame. Data for the students and algorithms are in Datasets S1 and S2.

ACKNOWLEDGMENTS. Work was funded in part by the Federal Bureau of Investigation (FBI) to the NIST; the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via IARPA R&D Contract 2014-14071600012 (to R.C.); Australian Research Council Linkage Projects LP160101523 (to D.W.) and LP130100702 (to D.W.); and National Institute of Justice Grant 2015-IJ-CX-K014 (to A.J.O.). The views and conclusions contained herein should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, the IARPA, or the FBI. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The identification of any commercial product or trade name does not imply endorsement or recommendation by the NIST.

1. Noyes E, Phillips PJ, O'Toole AJ (2017) What is a super-recogniser? *Face Processing: Systems, Disorders, and Cultural Differences*, eds Bindermann M, Megreya AM (Nova, New York), pp 173–201.
2. White D, Burton AM, Kemp RI, Jenkins R (2013) Crowd effects in unfamiliar face matching. *Appl Cognit Psychol* 27:769–777.
3. White D, Phillips PJ, Hahn CA, Hill MQ, O'Toole AJ (2015) Perceptual expertise in forensic facial image comparison. *Proc R Soc B* 282:20151292.
4. Dowsett AJ, Burton AM (2015) Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *Br J Psychol* 106:433–445.
5. O'Toole A, Abdi H, Jiang F, Phillips PJ (2007) Fusing face recognition algorithms and humans. *IEEE Trans Syst Man Cybern B* 37:1149–1155.
6. Phillips PJ, O'Toole AJ (2014) Comparison of human and computer performance across face recognition experiments. *Image Vis Comput* 32:74–85.
7. Phillips PJ (2017) A cross benchmark assessment of deep convolutional neural networks for face recognition. *Proceedings of the 12th IEEE International Conference on Automatic Face Gesture Recognition*, pp 705–710. Available at https://ieeexplore.ieee.org/document/7961810/. Accessed May 14, 2018.
8. National Research Council (2009) *Strengthening Forensic Science in the United States: A Path Forward* (National Academies Press, Washington, DC).
9. White D, Norell K, Phillips PJ, O'Toole AJ (2017) Human factors in forensic face identification. *Handbook of Biometrics for Forensic Science*, eds Tistaerlli M, Champod C (Springer, Cham, Switzerland), pp 195–218.
10. Facial Identification Scientific Working Group (2012) Guidelines for facial comparison methods, Version 1.0. Available at https://www.fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02.pdf. Accessed May 14, 2018.
11. Davis JP, Lander K, Evans R, Jansari A (2016) Investigating predictors of superior face recognition ability in police super-recognisers. *Appl Cognit Psychol* 30:827–840.
12. Robertson DJ, Noyes E, Dowsett A, Jenkins R, Burton AM (2016) Face recognition by metropolitan police super-recognisers. *PLoS One* 11:e0150036.
13. White D, Dunn JD, Schmid AC, Kemp RI (2015) Error rates in users of automatic face recognition software. *PLoS One* 10:e0139827.
14. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. *Proceedings of the British Machine Vision Conference*, eds Xie X, Jones MW, Tam GKL, pp 41.1–41.12. Available at www.bmva.org/bmvc/2015/index.html. Accessed May 14, 2018.
15. Chen JC, Patel VM, Chellappa R (2016) Unconstrained face verification using deep cnn features. *Proceedings of the IEEE Winter Conference of Appl Computer Vis (WACV)*, pp 1–9. Available at https://ieeexplore.ieee.org/document/7477557/. Accessed May 14, 2018.
16. Ranjan R, Sankaranarayanan S, Castillo CD, Chellappa R (2017) An all-in-one convolutional neural network for face analysis. *Proceedings of the 12th IEEE International Conference on Automatic Face Gesture Recognition Gesture Recognition*, pp 17–24. Available at https://ieeexplore.ieee.org/document/7961718/. Accessed May 14, 2018.
17. Ranjan R, Castillo CD, Chellappa R (2017) L2-constrained softmax loss for discriminative face verification. arXiv:170309507.
18. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Washington, DC), pp 1701–1708.
19. Prince J (2012) To examine emerging police use of facial recognition systems and facial image comparison procedures—Israel, Netherlands, UK, USA, Canada. *The Winston Churchill Memorial Trust of Australia*. Available at https://www.churchilltrust.com.au/media/fellows/2012_Prince_Jason.pdf. Accessed May 14, 2018.
20. Russell R, Duchaine B, Nakayama K (2009) Super-recognizers: People with extraordinary face recognition ability. *Psychon Bull Rev* 16:252–257.
21. Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. *IEEE Trans Pattern Anal Mach Intell* 20:226–239.
22. Norell K, et al. (2015) The effect of image quality and forensic expertise in facial image comparisons. *J Forensic Sci* 60:331–340.
23. Hu Y, et al. (2017) Person recognition: Qualitative differences in how forensic face examiners and untrained people rely on the face versus the body for identification. *Vis Cognit* 25:492–506.
24. Jeckeln G, Hahn CA, Noyes E, Cavazos JG, O'Toole AJ (March 5, 2018) Wisdom of the social versus non-social crowd in face identification. *Br J Psychol*, 10.1111/bjop.12291.
25. O'Toole AJ, et al. (2007) Face recognition algorithms surpass humans matching faces across changes in illumination. *IEEE Trans Pattern Anal Mach Intell* 29:1642–1646.
26. Phillips PJ, et al. (2010) FRVT 2006 and ICE 2006 large-scale results. *IEEE Trans Pattern Anal Mach Intell* 32:831–846.
27. Phillips PJ, et al. (2011) An introduction to the good, the bad, and the ugly face recognition challenge problem. *Proceedings of the Ninth IEEE International Conference on Automatic Face Gesture Recognition*, pp 346–353. Available at https://ieeexplore.ieee.org/document/5771424/. Accessed May 14, 2018.
28. O'Toole AJ, An X, Dunlop J, Natu V, Phillips PJ (2012) Comparing face recognition algorithms to humans on challenging tasks. *ACM Trans Appl Perception* 9:1–13.
29. Rice A, Phillips PJ, Natu V, An X, O'Toole AJ (2013) Unaware person recognition from the body when face identification fails. *Psychol Sci* 24:2235–2243.