PNAS
www.pnas.org

Supplementary Information for

Face recognition accuracy of forensic examiners, superrecognizers, and algorithms

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

Corresponding Author: P. Jonathon Phillips.
E-mail: jonathon@nist.gov

This PDF file includes:

Supplementary text
Figs. S1 to S14
Tables S1 to S3
Captions for Databases S1 to S2
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Databases S1 to S2

Supporting Information Text

Superrecognizers

There were two ways to qualify as a superrecognizer in this study. First, a previous face recognition test score that placed a person above a superrecognizer benchmark qualified the person for the superrecognizer group (1). The qualifying test could have been any in the superrecognizer literature; e.g., Cambridge Face Memory Test extended (CFMT) (2) or Glasgow Face Matching Test (GFMT) (3). Second, being employed currently to use their skills as a superrecognizer also qualified the person for superrecognizer status. Additionally, for the present study, formal experience or training in forensic comparisons disqualified the subject from the superrecognizer group.

In this study, one subject qualified by being employed professionally as a superrecognizer. The remaining 12 subjects qualified as superrecognizers by achieving accuracy of 0.90 or higher on the GFMT. These superrecognizers were screened in another study and were recruited from a pool of 18 subjects. A recruitment email for this study was sent out and 12 subjects responded. For the pool of 18 subjects, the mean accuracy was 0.96 with a standard deviation of 0.0365. Human subjects' regulations prevented sharing of individual scores.

The superrecognizers in our study had GFMT scores comparable to superrecognizers in previous studies (4, 5). Specifically, Davis et al. (4) reports GFMT scores for superrecognizers employed by the London Metropolitan Police (the Met). They also report on subjects identified as superrecognizers by taking the CFMT. In that study, the mean accuracy for the screened superrecognizers was 0.96 ($n = 10$) and for the Met superrecognizers was 0.94 ($n = 36$). Likewise, Robertson et al. (5) report results for the Met superrecognizers on the GFMT, with a mean = 0.96, ($n = 4$).

Analysis of Subject Groups

Mann-Whitney U test for between subject group analysis. In the **Accuracy** portion of the **Results** section, the P-values from comparing the AUC distributions between two subject groups are presented. The results of the comparison were based on Mann-Whitney U test (Table S1). The non-parametric Mann-Whitney U test was selected because the distribution of the AUCs for the face specialist groups was not Gaussian, see Fig. 2.

Table S1. Mann-Whitney U test from comparing the AUC distributions between the five subject groups. The comparison column lists the two subject groups being compared. The middle three columns report the Mann-Whitney U test and the sample sizes of the two distributions. The last column lists the *P*-values. Statistically significant differences are highlighted with an asterisk.

Comparison	Mann-Whitney U test	n_1	n_2	*P*-value
Facial Examiners vs. Facial Reviewers	669	57	30	0.097
Facial Examiners vs. Superrecognizers	331	57	13	0.56
Facial Examiners vs. Fingerprint Examiners	718	57	53	2.14×10^{-6} *
Facial Examiners vs. Students	246	57	31	2.53×10^{-8} *
Facial Reviewers vs. Superrecognizers	181	30	13	0.72
Facial Reviewers vs. Fingerprint Examiners	488	30	53	0.0037 *
Facial Reviewers vs. Students	145	30	31	4.01×10^{-6} *
Superrecognizers vs. Fingerprint Examiners	196	13	53	0.017 *
Superrecognizers vs. Students	66	13	31	4.87×10^{-4} *
Fingerprint Examiners vs. Students	570	53	31	0.020 *

Modeling Distribution of the Students. In the **Performance Distributions** section, we modeled the distribution of the AUCs for the students as a Gaussian distribution. The analysis here justifies this model. Fig. S1 shows the quantile-quantile plot comparisons between the students' AUC quantiles and the Gaussian quantiles. The close fit between the quantile-quantile points and the diagonal line shows that the Gaussian distribution is a reasonable model for the distribution of the students' AUCs.

Fusing Humans

Fig. 3 showed the distributions of the AUCs for the simulations that fused human ratings for the facial examiners. The fusion simulations were repeated for facial reviewers, superrecognizers, fingerprint examiners, and students. The results appear in Fig. S2. Median accuracy peaked at 1.0 (no errors) for the fusion of four forensic facial examiners or three superrecognizers. The performance of all groups increased with fusion. For reviewers, the median peaked at 0.98 with 10 participants fused. Fingerprint examiners peaked at a median of 0.97 with 10 participants. Median AUC for fusing 10 students was 0.88.

Fusing Humans and Algorithms

Fusing an algorithm with humans requires scaling the algorithm similarity score distribution to the human rating distribution. For each algorithm, the similarity scores were scaled to have the same mean and standard deviation as the ratings from all

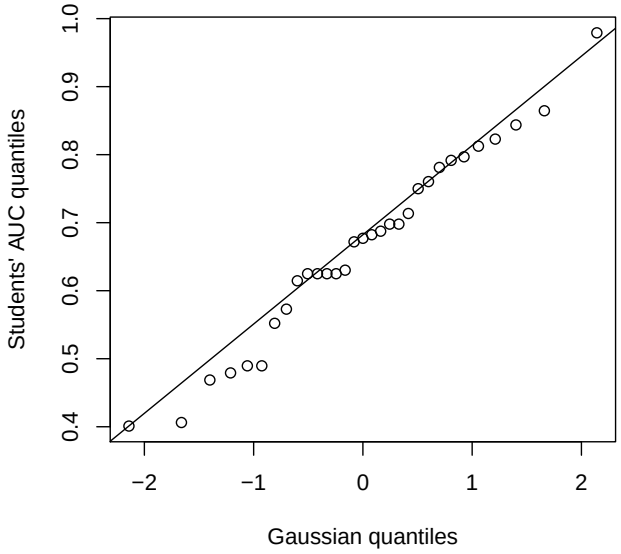


Fig. S1. A quantile-quantile plot that compares the students' AUC quantiles against the Gaussian quantiles. The close fit between the points and the diagonal line justifies modeling the distribution of the students' AUCs as a Gaussian.

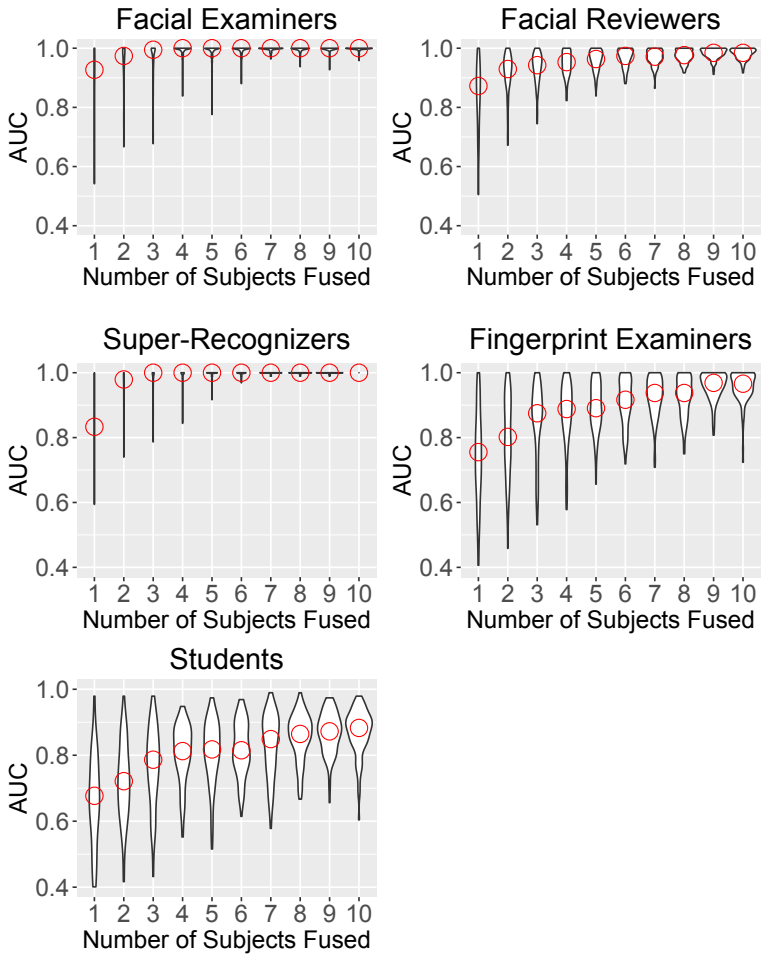


Fig. S2. Plots illustrate the effectiveness of fusing multiple participants. Results are presented for fusing facial examiners, facial reviewers, superrecognizers, fingerprint examiners, and students. Each plot shows the effect of fusing from 2 to 10 subjects. The violin plots characterize the range of accuracy based on 100 fusion simulations (median accuracy is indicated with a red circle). For the case of 1 subject, the violin plot shows the AUC range across all participants. The results show that for all five groups, combining judgments by simply averaging is effective. The median AUC reaches 1.0 for fusing four examiners or three superrecognizers. By comparison, the median AUC of fusing 10 students is 0.88, which is below the median AUC for individual facial examiner accuracy.

participants in the study. Each of the four algorithms was scaled separately. An algorithm's similarity scores were scaled to all participants to avoid over fitting to any one of the subject groups.

To scale the algorithm scores, we used the mean, μ_H, and the standard deviation, σ_H, of all human ratings. Let μ_A and σ_A be the mean and standard deviation, respectively, of the original algorithm similarity scores. For image pair i, let s_i be the original algorithm score and $\widehat{s_i}$ be the scaled algorithm score. The algorithm similarity scores were then scaled by

$$\widehat{s_i} = \left(\frac{s_i - \mu_A}{\sigma_A} + \mu_H \right) * \sigma_H.$$

Fusing Reviewers and Algorithms. In the main text section **Fusing Humans and Machines**, we present results for fusing algorithms and forensic facial examiners. Here we present the same analysis for fusing algorithms and facial reviewers. In the section that follows, the analysis is applied to fusing superrecognizers and algorithms.

Fig. S3 shows violin plots for fusing facial reviewers with each of the four algorithms. For reviewers, the most effective fusion was the combination of one reviewer with algorithm A2017b. This yielded a median AUC score of 0.99. This score was superior to the combination of two facial reviewers (Mann-Whitney U test = 2555, $n_1 = 435$, $n_2 = 30$, $P = 2.4 \times 10^{-8}$). Fusing individual reviewers with A2017a and A2016 yielded performance equivalent to the fusion of two reviewers (Mann-Whitney U test = 5224, $n_1 = 435$, $n_2 = 30$, $P = 0.068$; Mann-Whitney U test = 5718, $n_1 = 435$, $n_2 = 30$, $P = 0.257$). Fusing one reviewer with the oldest algorithm (A2015) showed no improvement over a single reviewer (Mann-Whitney U test = 431, $n_1 = 30$, $n_2 = 30$, $P = 0.78$). Fusing one reviewer with A2017b proved more accurate than fusing one reviewer with either A2017a or A2016 (Mann-Whitney U test = 264, $n_1 = 30$, $n_2 = 30$, $P = 0.005$; Mann-Whitney U test = 217, $n_1 = 30$, $n_2 = 30$, $P = 4.8 \times 10^{-4}$). Finally, fusing one reviewer and both A2017b and A2017a did not increase accuracy over fusing one reviewer with A2017b (Mann-Whitney U test = 358, $n_1 = 30$, $n_2 = 30$, $P = 0.16$).

Fusing Superrecognizers and Algorithms. Fig. S4 shows violin plots for fusing superrecognizers with each of the four algorithms. Fusing one superrecognizer with algorithm A2017b yielded a median AUC score of 0.99. Fusing two superrecognizers was equivalent to fusing one superrecognizer with A2017b, A2017a or A2016 (Mann-Whitney U test = 388, $n_1 = 78$, $n_2 = 13$, $P = 0.16$; Mann-Whitney U test = 368, $n_1 = 78$, $n_2 = 13$, $P = 0.11$; Mann-Whitney U test = 336, $n_1 = 78$, $n_2 = 13$, $P = 0.05$). Fusing one superrecognizer with the oldest algorithm (A2015) showed no improvement over a single superrecognizer (Mann-Whitney U test = 83, $n_1 = 13$, $n_2 = 13$, $P = 0.94$). Fusing one superrecognizer with A2017b proved more accurate than fusing one superrecognizer with either A2017a or A2016 (Mann-Whitney U test = 35, $n_1 = 13$, $n_2 = 13$, $P = 0.011$; Mann-Whitney U test = 32, $n_1 = 13$, $n_2 = 13$, $P = 0.007$). Finally, fusing one superrecognizer and both A2017b and A2017a did not increase accuracy over fusing one superrecognizer with A2017b (Mann-Whitney U test = 60, $n_1 = 13$, $n_2 = 13$, $P = 0.20$).

Error Rates for Highly Confident Decisions

The error rates for +3 and −3 judgments were presented graphically in Fig. 5. Table S2 lists the statistical tests reported in Fig. 5.

Table S2. Statistics supporting Fig. 5. The table shows the estimate $\hat{q}$ for the error rate and the upper and lower limits of the 95% confidence interval. These statistics are reported for facial examiners, facial reviewers, superrecognizers, fingerprint examiners, and students.

Group	Upper limit	Lower limit	$\hat{q}$	Error type
Facial Examiners	0.022	0.002	0.009	+3 on different faces
Facial Reviewers	0.036	0.003	0.012	+3 on different faces
Superrecognizers	0.052	0.0002	0.010	+3 on different faces
Fingerprint Examiners	0.061	0.022	0.038	+3 on different faces
Students	0.112	0.044	0.073	+3 on different faces
Facial Examiners	0.030	0.009	0.018	−3 on same faces
Facial Reviewers	0.032	0.005	0.014	−3 on same faces
Superrecognizers	0.099	0.022	0.051	−3 on same faces
Fingerprint Examiners	0.050	0.021	0.033	−3 on same faces
Students	0.185	0.111	0.145	−3 on same faces

Fusion for Highly Confident Decisions

Here we look at the impact of fusion on error-rates of facial examiner's highly confident decisions. This supplements results presented in the main text section **Error Rates for Highly Confident Decisions**.

Human ratings are made on a seven point scale with highly confident decisions assigned ratings of +3 and −3. Because algorithms use a continuous scale, the fusion of humans and algorithms must be compared with reference to a common scale. Here, we used receiver operator characteristic (ROC) analysis to compare the accuracy of highly confident human decisions with decisions based on the two fusion strategies. ROCs in this analysis report the trade-off between false accept rate (FAR)

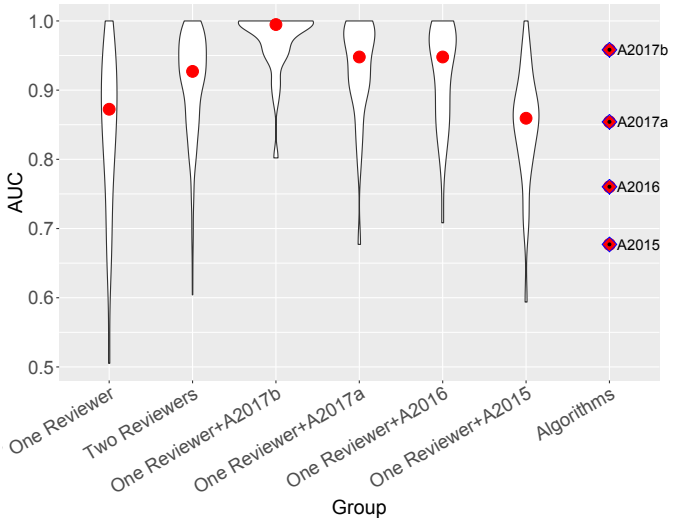


Fig. S3. Demonstration of the effectiveness of fusing reviewers and algorithms. Violin plots show the distribution of AUCs for each fusion simulation (red dots indicate the medians). The distribution of individual reviewers and the distribution of the fusion of two reviewers appear in the two leftmost violin plots. Also, the performance of the four algorithms appears in the right column. In between, plots show the distribution of facial reviewers fused with each of the four algorithms. Fusing one reviewer and A2017b produced more accurate identifications than fusing two reviewers; fusing one reviewer and A2017a or A2016 was equivalent to fusing two reviewers.

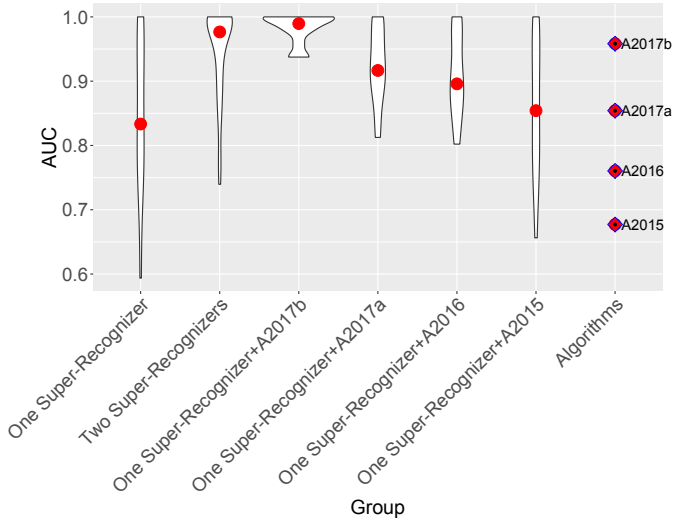


Fig. S4. Demonstration of the effectiveness of fusing superrecognizers and algorithms. Violin plots show the distribution of AUCs for each fusion simulation (red dots indicate the medians). The distribution of individual superrecognizers and the distribution of the fusion of two superrecognizers appear in the two leftmost violin plots. Also, the performance of the four algorithms appears in the right column. In between, plots show the distribution of superrecognizers fused with each of the four algorithms. Fusing one superrecognizer and A2017b did not produce more accurate identifications than fusing two superrecognizers; fusing one superrecognizer and A2017a or A2016 was equivalent to fusing two superrecognizers.

and false reject rate (FRR). A false accept occurs when two different faces are classified as the same person; a false reject occurs when two images of the same face are classified as two different people. To focus on the error rates of highly confident decisions, both the FAR and FRR axes are plotted on logarithmic scales. Note: this type of ROC plots one type of error against another type of error. For these ROCs, superior performance is towards the lower left corner—smaller false accept and false reject rates

To compare highly confident decisions, we start with three full ROCs that show performance for: a.) single facial examiners; b.) the fusion of two facial examiners; and c.) the fusion of facial examiners and A2017b, see Fig. S5. For facial examiners, high confidence decisions of +3 are indicated with a blue diamond and high confidence decisions of −3 are indicated with a blue circle.

If we look at the cases where single facial examiners assign a rating of +3 to two different facial identities (high confidence false accepts), we find a FAR = 0.009, (see Table S2) and a FRR of 0.77. When we fuse two examiners, the FRR falls to 0.52. When we fuse examiners and A2017b, the FRR falls further to 0.27. Therefore, for the FAR corresponding to a +3 rating, both fusion strategies yield substantially lower FRRs. Another strategy is to maintain the same FRR with the goal of operating at a lower FAR.

We now turn our attention to the other highly confident error, when single examiners assign a rating of −3 to two face images of the same person (high confidence false rejects). The −3 error for the same face corresponds to a FRR = 0.018 and FAR = 0.77, see Table S2. When we fuse two examiners, the FAR falls to 0.46; when we fuse examiners and A2017b the FAR falls further to 0.23. For all operating points including highly confident decisions, the best performance is obtained by fusing examiners and A2017b, followed by fusing two examiners. Both are superior to one examiner.

This ROC analysis was repeated for facial reviewers and superrecognizers. Fig. S6 compares the ROCs for the facial reviewers, fusing two facial reviewers, and fusing facial reviewers and A2017b. For all operating points including highly confident decisions, the best performance is obtained by fusing reviewers and A2017b, followed by fusing two reviewers. Both are superior to one reviewer.

The third case we consider is superrecognizers. Fig. S7 compares the ROCs for the superrecognizers, fusing two superrecognizers, and fusing superrecognizers and A2017b. For superrecognizers, both fusion strategies are superior to one superrecognizer. Because the ROCs for the two fusion strategies intersect, the best fusion strategy depends on the operating point. This difference from the facial examiner and facial reviewer cases could be due to the small number of superrecognizers in the study.

Algorithms

The four algorithms in this study were chosen because: a.) the details of the algorithms are documented in the literature; b.) performance is reported on established benchmark datasets in the face recognition literature; c.) the details of the construction of the training sets are documented in the literature; and d.) the training sets are independent of the face images used in the test we conducted in this study.

Algorithm A2015 is publicly available and has been adopted by the automatic face recognition community as a benchmark DCNN-based algorithm. A2016, A2017a, and A2017b are a sequence of algorithms designed to improve face identification accuracy when there is significant variability in pose, illumination, and size of the face. The four algorithms in this study reflect the rapid evolution of DCNN-based algorithms and show improvement in DCNNs performance between 2015 and 2017.

Three common benchmarks for DCNN-based face recognition algorithms are the Labeled Faces in the Wild (LFW) (6), the YouTube* Faces Database (7), and the IARPA Janus Benchmark-A (IJB-A) (8). LFW consists of face images of famous people downloaded from the Web. The majority of the faces in the LFW images are frontal, and there is a wide range of illumination conditions. The YouTube Faces Database consists of YouTube videos and is modeled on the LFW. Similar to LFW, most of the faces in the videos are frontal and there is variation in the illumination between videos. The IJB-A consists of face images downloaded from the Web and contains a wide variation in pose, illumination, and size of the face. It is the most challenging of the three benchmarks. The accuracy metric for LFW and YouTube Faces Database is $1 - \text{EER}$, where EER is the equal error rate. For IJB-A, we report the true accept rate at a false accept rate of 1 in 100. It is common practice for computer-based face identification algorithms to be tested at low false alarm rates, because they are required to operate in most applications with low false alarm rates. A false accept occurs when two different faces are classified as the same person; a true accept occurs when two images of the same face are classified as the same person. The algorithms in this study reported state-of-the-art performance at the time of their respective publication dates, see Table S3.

Table S3. Benchmark accuracy of algorithms A2015, A2016, A2017a, and A2017b on three publicly available benchmarks. All four algorithms reported state-of-the-art performance at time of publication. The benchmarks are the Labeled Faces in the Wild (LFW), the YouTube Faces Database (YouTube), and the IARPA Janus Benchmark-A (IJB-A). The performance metric for LFW and YouTube is 1−EER, where EER is the equal error rate. For IJB-A, we report the true accept rate at a false accept of 0.01.

	LFW	YouTube	IJB-A
A2015	0.9727	0.928	
A2016	0.9745		0.838
A2017a			0.893
A2017b	0.9978	0.9608	0.95

*The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

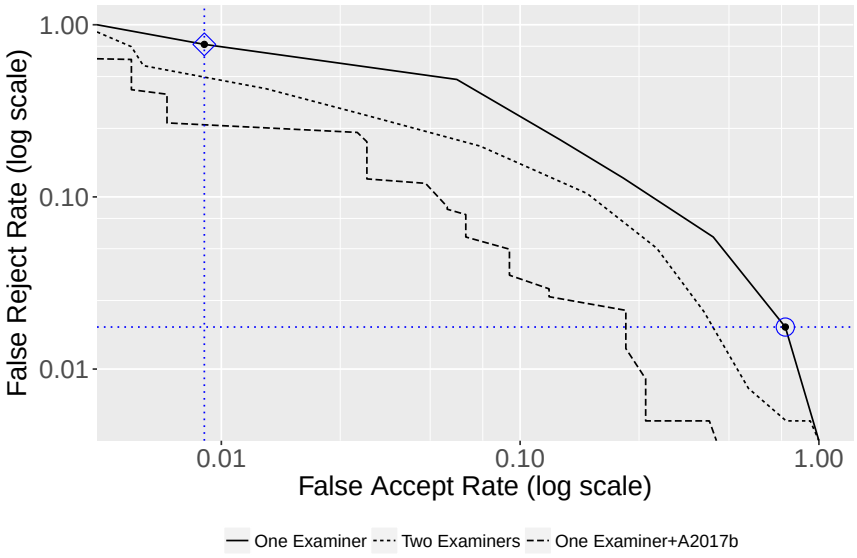


Fig. S5. Demonstrating the effectiveness of fusing two facial examiners and of fusing one facial examiner and algorithm A2017b. ROCs are plotted for facial examiners, fusion of two facial examiners, and fusion of one facial examiner and algorithm A2017b. To emphasize the error rates for highly confident decisions, both the false accept and false reject axes are plotted on a logarithmic scale. The point with the blue diamond is the operating point corresponding to the error rate for +3 highly confident decisions and the vertical blue dotted line corresponds to the false accept rate (FAR) for examiners' +3 rating. If the FAR is set to 0.009, the decrease in the false reject rate (FRR) for the two fusion strategies is shown at the intersection of the vertical line and the fusion ROCs. The black dot with the blue circle is the operating point for -3 highly confident decisions and the horizontal blue dotted line corresponds to the FRR of the -3 operating point. If the FRR is set to 0.018, the decrease in the FAR for the two fusion strategies is shown at the intersection of the horizontal line and the fusion ROCs. For all operating points, fusing one examiner and A2017b is superior to both one examiner and to fusing two examiners. For these ROCs, superior performance is towards the lower left corner—smaller false accept and false reject rates.

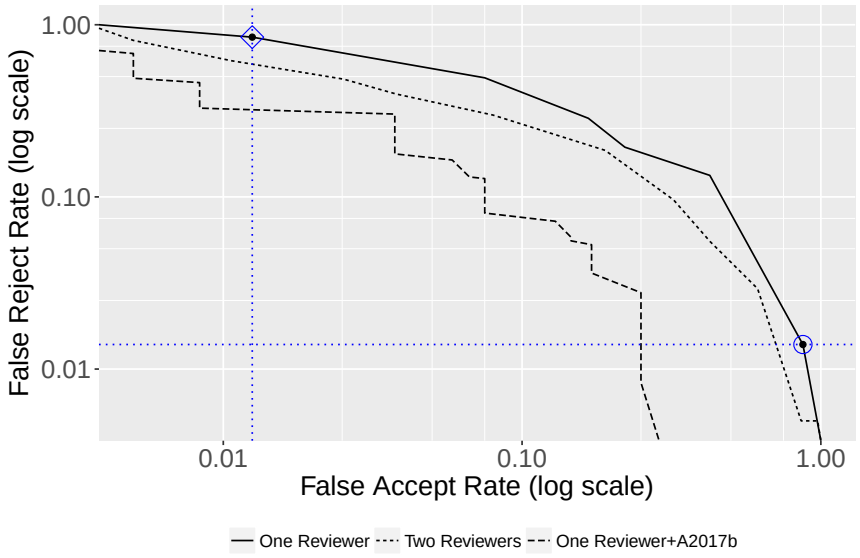


Fig. S6. Demonstrating the effectiveness of fusing two facial reviewers and of fusing one facial reviewer and algorithm A2017b. ROCs are plotted for facial reviewers, fusion of two facial reviewers, and fusion of one facial reviewer and algorithm A2017b. To emphasize the error rates for highly confident decisions, both the false accept and false reject axes are plotted on a logarithmic scale. The point with the blue diamond is the operating point corresponding to the error rate for +3 highly confident decisions and the vertical blue dotted line corresponds to the false accept rate (FAR) for examiners' +3 rating. If the FAR is set to 0.012, the decrease in the false reject rate (FRR) for the two fusion strategies is shown at the intersection of the vertical line and the fusion ROCs. The black dot with the blue circle is the operating point for -3 highly confident decisions and the horizontal blue dotted line corresponds to the FRR of the -3 operating point. If the FRR is set to 0.014, the decrease in the FAR for the two fusion strategies is shown at the intersection of the horizontal line and the fusion ROCs. For all operating points, fusing one reviewer and A2017b is superior to both one reviewer and fusing two reviewers. For these ROCs, superior performance is towards the lower left corner—smaller false accept and false reject rates.

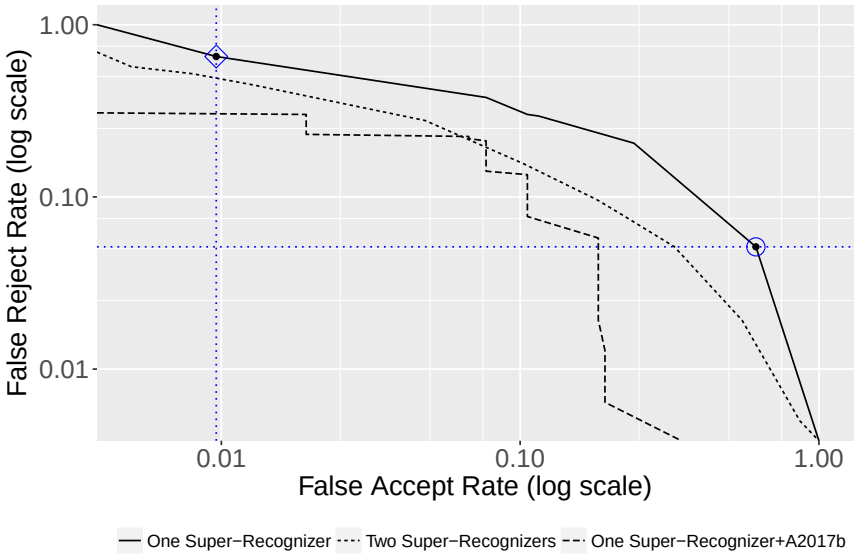


Fig. S7. Demonstrating the effectiveness of fusing two superrecognizers and of fusing one superrecognizer and algorithm A2017b. ROCs are plotted for superrecognizers, fusion of two superrecognizers, and fusion of one superrecognizer and algorithm A2017b. To emphasize the error rates for highly confident decisions, both the false accept and false reject axes are plotted on a logarithmic scale. The point with the blue diamond is the operating point corresponding to the error rate for +3 highly confident decisions and the vertical blue dotted line corresponds to the false accept rate (FAR) for examiners' +3 rating. If the FAR is set to 0.010, the decrease in the false reject rate (FRR) for the two fusion strategies is shown at the intersection of the vertical line and the fusion ROCs. The black dot with the blue circle is the operating point for -3 highly confident decisions and the horizontal blue dotted line corresponds to the FRR of the -3 operating point. If the FRR is set to 0.051, the decrease in the FAR for the two fusion strategies is shown at the intersection of the horizontal line and the fusion ROCs. For all operating points, both fusion strategies are superior to one superrecognizer. For these ROCs, superior performance is towards the lower left corner—smaller false accept and false reject rates.

Training DCNNs requires millions of face images with 100s to 1000s of images per person. Because DCNNs are a supervised learning method, in the training set, all images of each individual have the same label. The predominate method to obtain training sets of the required size is to scrape the Web for images of celebrities and famous people. This method has been adopted because: a.) images of celebrities are readily available, b.) there are a large numbers of images of celebrities taken under different conditions, and c.) search engines can locate face images for each celebrity. This search procedure produces a set images associated with each celebrity. However, because images "associated with a celebrity" may not actually contain the celebrity, or they contain multiple faces, or contain an image of extremely poor-quality, a second screening phase is required to remove non-face images and poor-quality face images, and to correct identity label errors. This is done typically by a combination of human and automated steps. The human steps consist of crowd sourcing by lay people. The specific steps implemented here differed for A2015 (9) and A2016, A2017a, A2017b (10), (11). However, in all cases, the final product consisted of a training set of cropped faces and identity labels for all faces. It is never possible to remove all errors. The goal of this procedure is, therefore, to reduce the labeling and cropping errors so that DCNNs could be effectively trained. The training set for A2015 consisted of 2.6 million face images of 2,622 people. Algorithms A2016, A2017a and A2017b were all trained from a training set that consisted of 3.7 million face images of 58,207 people.

Clearly, the more accurate the labels in the training set, the more accurate the DCNN algorithm will be. One potential method for increasing the efficacy of a training set is to employ facial examiners or superrecognizers to inspect the labels of the most difficult cases.

Methods

Facial examiners, facial reviewers, superrecognizers, and fingerprint examiners were given three months to complete the comparisons. Because the participants completed the comparison as their schedules permitted, we were not able to collect data on the length of time it took each subject to perform each comparison or the total amount of time a subject spent performing all 20 comparisons.

There was some overlap between the facial examiners who participated in the present experiment and those who participated in White et al. (3). That study was administered in a single half-day session to 27 facial examiners on 4 May 2014. In the current study, consenting of facial examiners started in March 2016. The test was administered to 57 facial examiners, with all participants allotted 3 months from their respective start date. In (3), examiners viewed the facial pairs for 30 seconds and they did not receive feedback on their performance. Because White et al. did not distinguish between facial examiners and facial reviewers, the corresponding number of subjects for this study amounts to 87 (the combination of 57 examiners and 30 reviewers). Consequently, we do we do not know the exact number of reviewers and examiners who also participated in (3). Because of the time between the two studies and 30-second viewing time in the first study, examiner participation in (3) is highly unlikely to have affected performance in the current study.

Same-Identity Pair 1

Same-Identity Pair 2

Same-Identity Pair 3

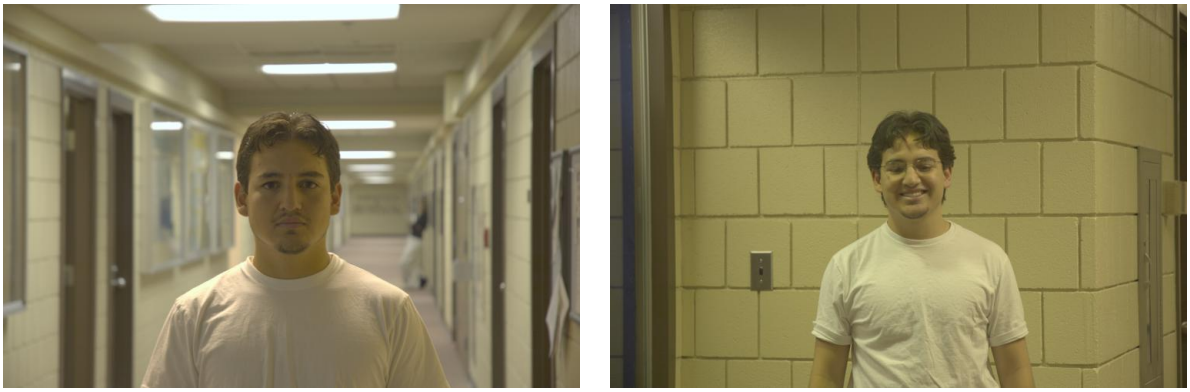
Fig. S8. Same-identity pairs 1, 2 and 3.

Same-Identity Pair 4

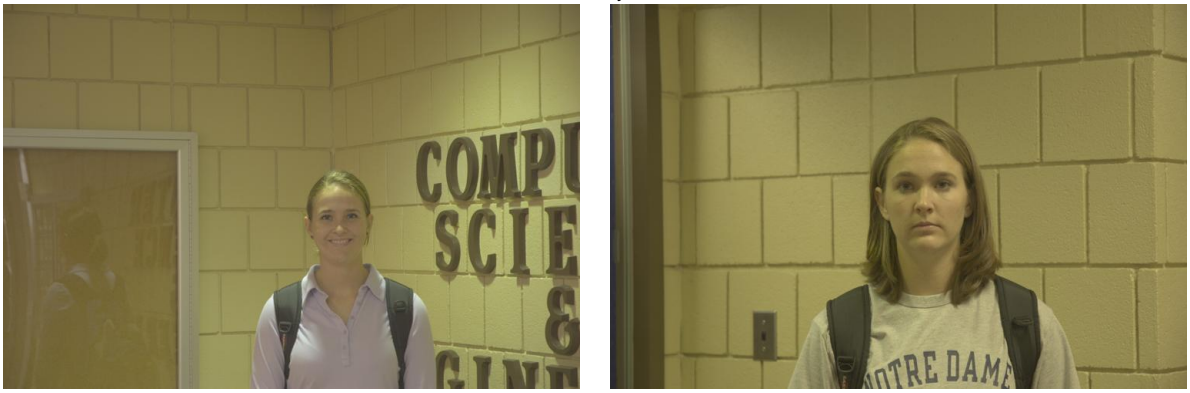
Same-Identity Pair 5

Same-Identity Pair 6

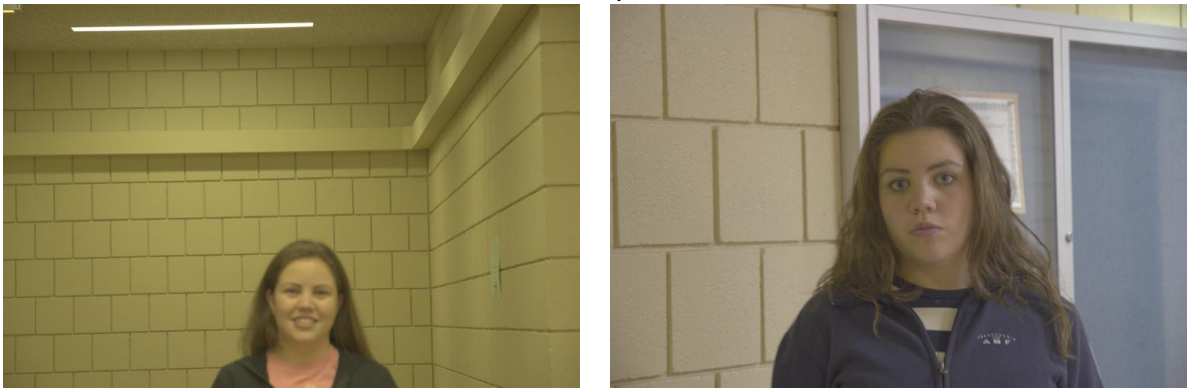
Fig. S9. Same-identity pairs 4, 5 and 6.



Same-Identity Pair 7



Same-Identity Pair 8



Same-Identity Pair 9

Fig. S10. Same-identity pairs 7, 8 and 9.

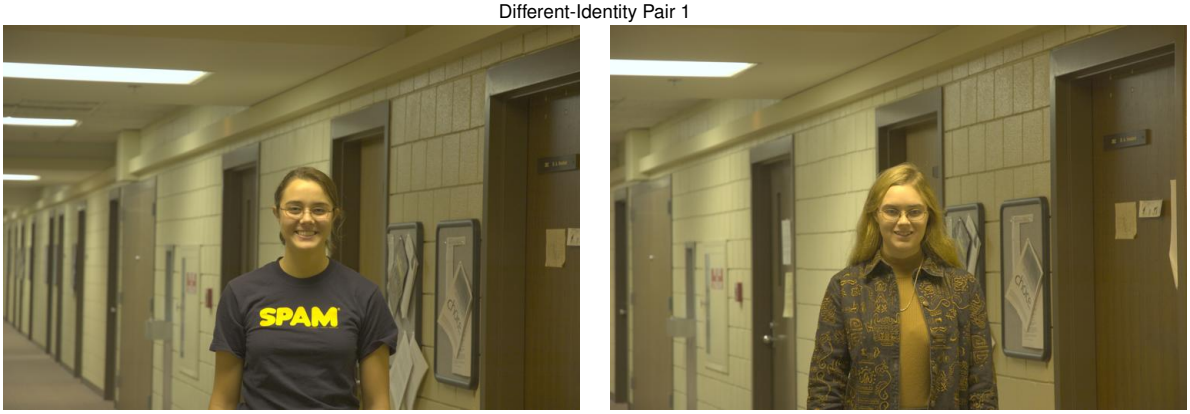
Same-Identity Pair 10

Same-Identity Pair 11

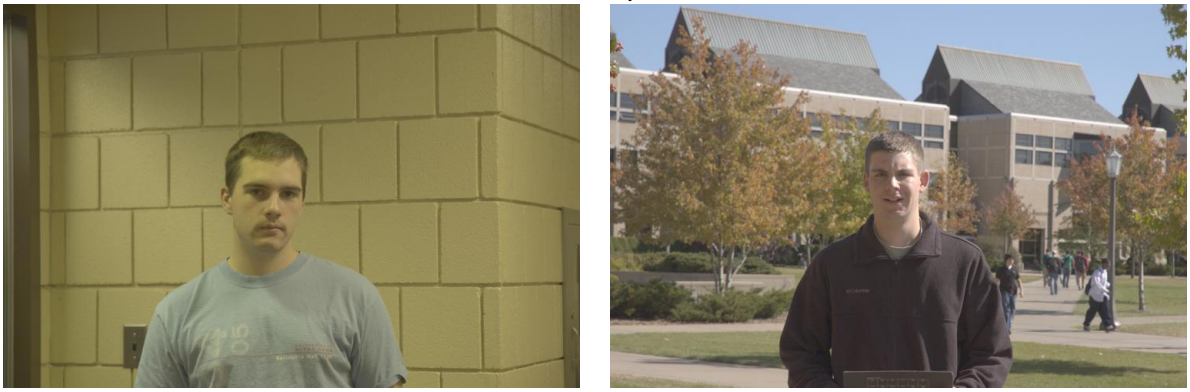
Same-Identity Pair 12

Fig. S11. Same-identity pairs 10, 11 and 12.

Different-Identity Pair 1



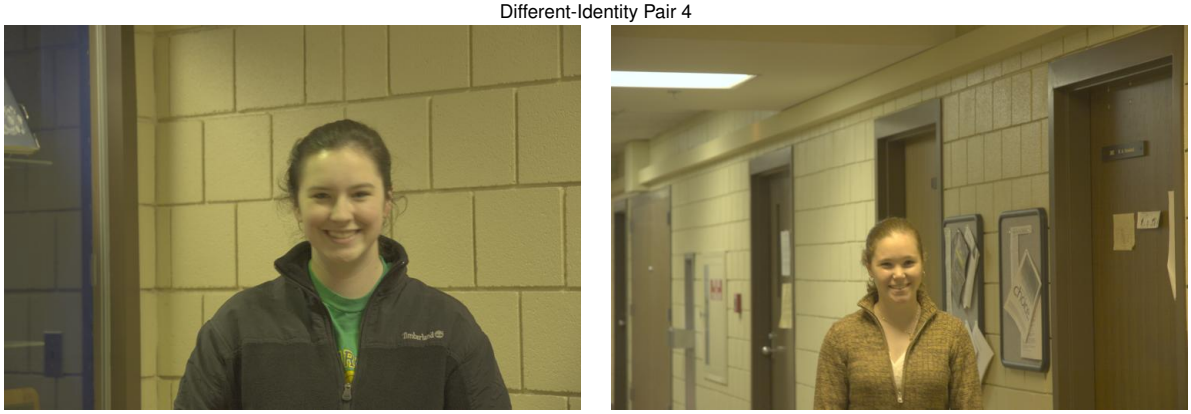
Different-Identity Pair 2



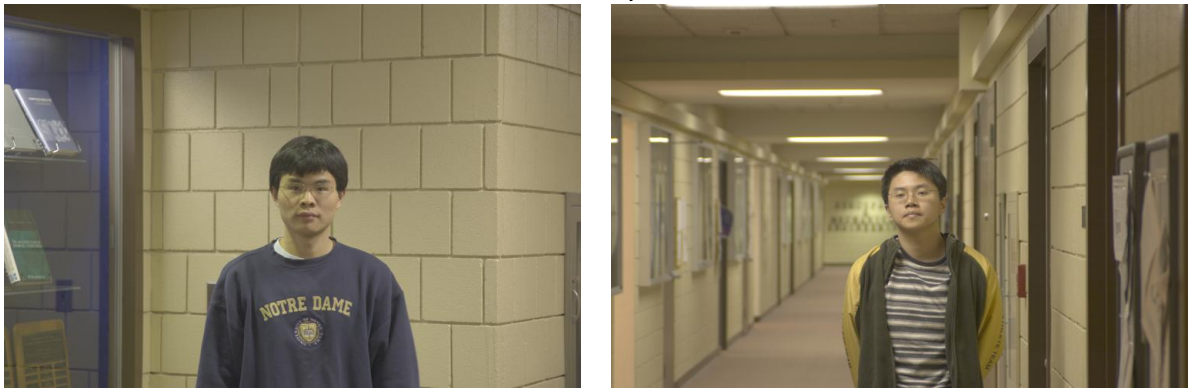
Different-Identity Pair 3

Fig. S12. Different-identity pairs 1, 2 and 3.

Different-Identity Pair 4

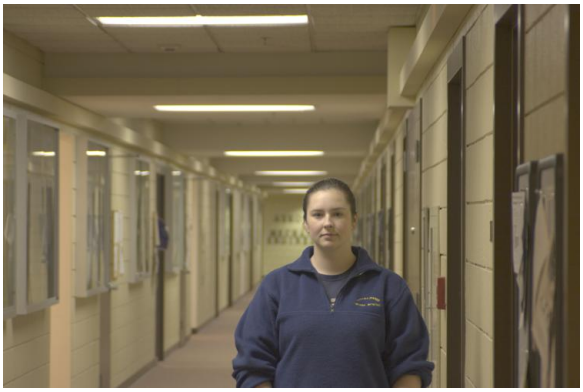


Different-Identity Pair 5

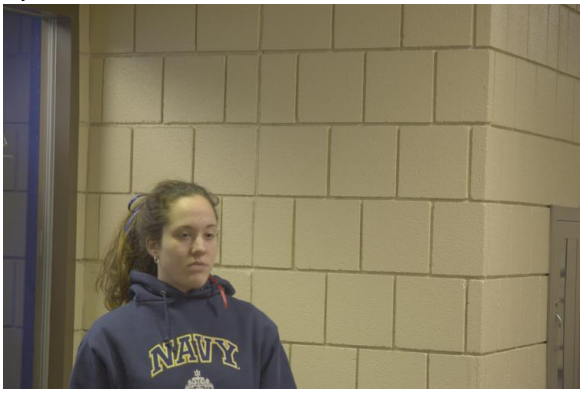


Different-Identity Pair 6

Fig. S13. Different-identity pairs 4, 5 and 6.




Different-Identity Pair 7



Different-Identity Pair 8

Fig. S14. Different-identity pairs 7 and 8.

Additional data table S1 (studentRatingsForPublication20171211.csv)
Ratings for students on all 20 image-pairs.

Additional data table S2 (algorithmsForPublication20171212.csv)
Similarity scores for algorithms A2015, A2016, A2017a, and A2017b.

References

1. Noyes E, Phillips PJ, O'Toole AJ (2017) What is a super-recogniser? in *Face processing: Systems, Disorders, and Cultural Differences*, eds. Bindermann M, Megreya AM. (Nova, New York, NY, USA), pp. 173–201.
2. Russell R, Duchaine B, Nakayama K (2009) Super-recognizers: people with extraordinary face recognition ability. *Psychon Bull Rev* 16(2):252–257.
3. Burton AM, White D, McNeill A (2010) The Glasgow Face Matching Test. *Behavior Research Methods* 42:286–291.
4. Davis JP, Lander K, Evans R, Jansari A (2016) Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology*.
5. Robertson DJ, Noyes E, Dowsett A, Jenkins R, Burton AM (2016) Face recognition by metropolitan police super-recognisers. *PLoS ONE* 11(2).
6. Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled Faces in the Wild: a database for studying face recognition in unconstrained environments, (University of Massachusetts, Amherst), Technical Report 07-49.
7. Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. pp. 529–534.
8. Klare BF, et al. (2015) Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A in *IEEE Conference on Computer Vision and Pattern Recognition*.
9. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. *Proceedings of the British Machine Vision*.
10. Lin WA, Chen JC, Chellappa R (2017) A proximity-aware hierarchical clustering of faces in *Proc. 12th IEEE International Conference on Automatic Face & Gesture Recognition*.
11. Bansal A, Nanduri A, Castillo C, Ranjan R, Chellappa R (2017) UMDFaces: an annotated face dataset for training deep networks in *Proceedings of International Joint Conference on Biometrics*.