Fusing Individual Algorithms and Humans Improves Face Recognition Accuracy

Alice J. O'Toole, Fang Jiang, Hervé Abdi & P. Jonathon Phillips*

The University of Texas at Dallas
*National Institute of Standards and Technology

Abstract. Recent work indicates that state-of-the-art face recognition algorithms can surpass humans matching identity in pairs of face images taken under different illumination conditions. It has been demonstrated further that fusing algorithm- and human-derived face similarity estimates cuts error rates substantially over the performance of the best algorithms. Here we employed a pattern-based classification procedure to fuse individual human subjects and algorithms with the goal of determining whether strategy differences among humans are strong enough to suggest particular man-machine combinations. The results showed that error rates for the pairwise man-machine fusions were reduced an average of 47 percent when compared to the performance of the algorithms individually. The performance of the best pairwise combinations of individual humans and algorithms was only slightly less accurate than the combination of individual humans with all seven algorithms. The balance of man and machine contributions to the pairwise fusions varied widely, indicating that a one-size-fits-all weighting of human and machine face recognition estimates is not appropriate.
...

1 Introduction

Face recognition algorithms have improved dramatically over the last decade and are now available commercially for security applications.The most common applications involve face verification, where a presented image of a person must be compared to a stored representation and be verified or rejected as an identity match. To achieve this task, algorithms produce an estimate of the likelihood that two images are of the same person. A match criterion is then set for making the decision.

According to a recent US Government sponsored test of state-of-the-art face recognition algorithms [1, 2], current algorithms are impressively accurate at this task when the images to-be-matched are taken under controlled illumination conditions. Specifically, the Face Recognition Grand Challenge

⋆ In G. Bebis et al. (Eds):Advances in Visual Computing. ISVC 2006. LNCS4292. New York: Springer Verlag. pp. 447-456, 2006.

(FRGC), conducted at the National Institute of Standards and Technology (NIST) between 2004 and 2006, tested 17 algorithms on a task of matching face identity in roughly 128 million pairs of images taken under controlled illumination. The results showed an average verification rate of .91 at the .001 false acceptance rate.

In an analogous FRGC test of algorithms on the task of matching face identity in images taken under different illumination conditions, however, the performance of algorithms was less impressive. Only seven algorithms volunteered to participate in this experiment. These systems scored an average verification rate of only .42 at the .001 false acceptance rate. This suggests that current face recognition algorithms may not be ready for application environments that have natural variations in illumination. These kinds of environments are typical in airports and other public places where some part of the illumination comes from natural light, (e.g., sunlight filtered through windows).

How well must a face recognition algorithm perform to be useful for a security application? Although human performance on face recognition is often considered the standard to which algorithms should aspire, there are few direct comparisons between humans and algorithms. A recent exception to this general rule is a study comparing the performance of humans with algorithms competing in the uncontrolled illumination experiment of the FRGC [3]. In that study, "easy" and "difficult" face pairs were sampled from the FRGC test set, using a control algorithm based on principal components analysis (PCA) of the aligned and scaled images. Human subjects rated the likelihood that the pairs of face images were of the same person. ROC curves computed for the humans and for the algorithms revealed that three algorithms [4–6] performed more accurately than humans on the difficult face pairs. Nearly *all* algorithms performed more accurately than humans on the easy face pairs. Post-hoc analyses of the human subject data gave no indication that subject attention waned toward the end of the experiment. Moreover, both humans and algorithms were more accurate on face pairs prescreened by the PCA to be easy than on the face pairs prescreened to be difficult. The results indicate, therefore, that although algorithms appear to perform poorly on the task, they are nonetheless competive with the performance of human subjects.

The comparison between algorithms and humans on the face matching task gives an indication of the quantitative ranking of algorithms by accuracy, but does not offer any information about how similarly humans and algorithms perform the task. To gain insight into the qualitative aspects of the performance of humans and algorithms, and to see if performance could be

improved, a fusion study of the algorithms and humans was carried out [7]. If algorithms and humans employ different recognition strategies, it should be possible to fuse the their estimates of face similarity to improve performance. The fusion was carried out in two parts. In the first part, the similarity estimates of the seven algorithms from the FRGC were fused with partial least squares regression (PLS)[8, 9] and used to predict the match status of individual pairs of faces. The robustness of the PLS was tested with cross-validation. The results indicated that fusing the algorithms cut the error rate of the best-performing algorithm by a factor of two [7].

The second part of the study examined whether algorithm performance could be improved by fusing human similarity estimates with the estimates of the seven algorithms. Specifically, human similarity estimates for the difficult face pairs [3] were averaged across 49 subjects and fused with the algoritms estimates using PLS. This reduced error rate to near perfection and indicated that human face recognition strategies differ sufficiently from algorithms to make a substantial contribution to recognition performance through fusion.

Although it is of general interest to know that humans and algorithms can be fused to increase face matching accuracy, it is of more practical value to know how to combine *individual* algorithms with *individual* humans. Which algorithms combine most beneficially with which humans? Can a general rule be established, or do humans and algorithms differ sufficiently in recognition strategy to suggest that individual human-machine fusions be done on a case-by-case basis. The purpose of the present study was to explore the benefits of fusing individual humans with indidvidual algorithms. This is of practical value given that in most real-world applications, one algorithm works under the supervision of a single human operator. We assessed the suitability of particular man-machine combinations using a fusion approach.

We approached this problem in two parts. First, we fused the similarity scores produced by the seven available algorithms and individual humans. This provides data on the best-case scenario, (i.e. where humans can benefit from all available algorithm expertise). This fusion is similar to the one described previously [7], with the exception that humans, in this case, are treated as *individuals*, rather than being represented by their global average. Next, we fused individual human subjects with individual algorithms to assess performance of particular man-machine hybrids.

2 Methods

The methods for this study make use of human subject and FRGC algorithms' estimates of face pair similarity that were collected previously [3]. For com-

pleteness we provide a sketch of the methods used to collect the relevant data in that study and then proceed to describe the fusion methods applied in the present work.

2.1 Stimuli

The face images used to test algorithms in the face matching task were drawn from a database developed for the FRGC [1]. Face pairs consisted of a target image, taken under controlled illumination conditions, and a probe image, taken under uncontrolled illumination conditions, (e.g., in a corridor). Target images had a resolution of 1704 x 2272 pixels and probe images had a resolution of 2272 x 1704. A sample face pair appears in Figure 1.

Fig. 1. An example non-match pair with the probe image on the left and the target probe image on the right

To make the task as challenging as possible, we sampled the images from a homogenous face population that included only male and female Caucasians in their twenties and thirties. Each face pair was matched by sex. Combined these constraints eliminated the possibility that humans could base identity judgments on surface facial characteristics associated with sex, race, or age.

In the present study, we used the 120 difficult pairs of faces (60 male and 60 female) presented to subjects in the previous study [3]. Half of the face pairs were of matched identity, (i.e., the target and probe images were of the same person), and half were non-match pairs (i.e., the target and probe images were of different people). As noted, face pairs were prescreened by a PCA of the aligned and scaled images before sampling for the human experiment. Difficult match pairs had similarity scores that were less than two stan-

dard deviations below the match mean, (i.e., two dissimilar pictures of the same person). Difficult non-match pairs had similarity scores greater than two standard deviations above the non-match mean, (i.e., similar images of two different people).

2.2 Human estimates of face similarity

Humans judged the similarity of the 120 face pairs by rating each pair as follows, "1.) sure they are the same person; 2.) think they are the same person; 3.) don't know; 4.) think they are not the same person; 5.) sure they are not the same person." The ratings of 49 subjects on the 120 face pairs served as the human input data for the fusion analysis.

2.3 Algorithm estimates of face similarity

To participate in the FRGC uncontrolled illumination experiment, algorithm developers were asked to compute a matrix containing similarity scores between all possible pairs of 16,028 target images and 8,014 probe images. The resulting matrix, therefore, contained 128,448,392 similarity scores, each representing the likelihood that a target and probe images in the pair were of the same person. These matrices were scored by NIST and complete performance results for the algorithms are available elsewhere [1, 2]. For present purposes, the similarity scores for the 120 difficult face pairs presented to subjects were extracted from each of seven participating algorithms' similarity matrices. These scores served as the algorithms' input to the fusion analysis.

2.4 PLS fusion

Fusion was performed by partial least squares (PLS) regression, a statistical technique that generalizes and combines features from principal component analysis and multiple regression [8, 9]. The technique is used to predict a set of dependent variables from a set of independent variables (predictors). Though less known in the pattern recognition literature, PLS is widely used in chemometrics, sensory evaluation, and neuroimaging data analysis, (cf. [9]).

The predictors for the present study were the human- and algorithm-generated similarity scores for the 120 pairs of face images. We define this more specifically in the context of particular fusion analyses. The dependent variable was the match status of the face pair (i.e., same versus different person). with match pairs assigned a value of 1 and non-match pairs a value

of −1. PLS regression gives a set of orthogonal factors, sometimes called latent vectors, from the covariance matrix of predictors and dependent variables. These can be used to predict the dependent variable(s), by appropriately weighting the predictors. This set of weights is called $\mathbf{B}_{\text{pls}}$ in the PLS-regression literature [8, 9].

The predictive power of a PLS solution is assessed generally with cross-validation techniques such as a bootstrap or jackknife procedure. All factors, or only a subset of them, can be used to compute the prediction of the dependent variable(s), which are obtained as a weighted combination of the original predictors given by $\mathbf{B}_{\text{pls}}$. The larger the number of factors kept, the better the prediction of the "learning set" but, in general, a smaller number of factors is optimal for robust prediction (i.e., for test set predictions).

3 Results

3.1 Baseline Human Performance

We first assessed the baseline performance of individual humans in a way that is comparable with the output of the fusion test. Accuracy in the fusion test is given as the error rate for match status classifications determined in a jack-knife procedure. As noted, humans rated each pair of faces on a 5-point scale varying from "sure the same person" to "sure different people". Given that one of the possible human responses was "don't know", we computed the number of errors in two ways by: a.) assigning the "don't know" to a match response; and b.) assigning the "don't know" response to a non-match response. We computed the number of errors for each subject in both ways and averaged these two values. Across all subjects, the human error rate averaged .141 (see column one of Table 1), indicating good, but not perfect performance.

Table 1. The error rate information for individual human performance appears in the first column. Analogous information appears in the second column for the fusion of individual humans combined with the seven algorithms. The data for the best-paired individual human-algorithm fusions appear in the third column

Error rate	Human	Human+7 Algorithms	Human+Algorithm Best-paired Fusions
average	.141	.057	.078
minimum	.041	.033	.033
maximum	.258	.108	.117

3.2 Fusing Individual Humans with All Algorithms

Next, we fused each individual with all seven algorithms using PLS. The purpose of this analysis was to assess the relative importance of algorithms when combined with individual humans. We also use the performance of this simulation as "best case" control for the individual human-algorithm fusions that follow.

A predictor matrix for the PLS was created for each of the 49 subjects as follows. Estimates of face similarity for the 120 face pairs from the human subject were combined in a column-wise matrix with the estimates taken from the seven algorithms for the same face pairs. A jack-knife procedure operated by deleting each of the face pairs in turn and computing the PLS on the remaining 119 pairs. Accuracy was measured as the number of correct classifications for the deleted, "left-out" face pairs. In all cases, jack-knife procedures tested the accuracy of human algorithm fusion with 1 through 5 factors PLS solutions. The optimal number of factors for the 49 subjects varied from 1 to 5, with a median of 1, and a mean of 1.36. For each subject, only the solution that yielded highest accuracy is reported.

Fusing individual humans with the seven algorithms cut the human error rate by more than a factor of two (cf., Table 1). For reference, the error rates for the individual algorithms from [7] appear in first column of Table 2. As can be seen, the average error rate achieved by fusing individual subjects with the combination of expertise found in the seven algorithms is less than half of that achieved by the best algorithm.

To interpret the role of individual algorithms and humans in tbe fusion, we looked at the PLS-derived weights for combining similarity estimates in the PLS. These weights emerge from the PLS as a recipe for optimally combining algorithm and human similarity estimates to predict match status. Averaged across the 49 subject-based fusions, the strongest contributor to the fusion was the algorithm from Viisage Corporation [6]. This was followed by the algorithm from the New Jersey Institute of Technology (NJIT) [4]. Anonymous algorithms B and D also played a role in the fusion. The weights for humans were well below these algorithm weights indicating that the role of humans in the fusion was minor. In previous work [7], we fused the *average* human with the seven algorithms and found that the human role in improving performance in the fusion was more substantial. Here we found that fusing *individual* subject performance with the seven algorithms and *averaging* the weights afterwards revealed a much smaller role for humans. This suggests that individual human performance may vary strategically and that particular person-tailored mixtures of individuals and algorithms may provide for better balance in human-machine fusions. This finding motivates the next

Table 2. Results compilation for human-algorithm fusions. The first column contains error rates obtained previously [7] for each algorithm operating alone. The human error rate from Table 1 appears at the top of the column for comparison. The second column shows the average error rates achieved by fusing individual humans with each of the seven algorithms. The percentage reduction in error rate from the individual algorithms to human-algorithm pair fusion appears in column 3.

Source	individual error rate[7]	paired fusion error rate	% error rate reduction
human	.14	–	
NJIT	.12	.08	33%
Viisage	.20	.12	40%
CMU	.14	.09	36%
Algorithm A	.37	.13	65%
Algorithm B	.23	.11	52%
Algorithm C	.25	.12	52%
Algorithm D	.26	.13	50%

analysis in which we fused all possible pairs of individual humans and individual algorithms.

3.3 Fusing Individual Humans and Individual Algorithms

The fusion was carried out as described previously, but this time combining each of the 49 subjects inidvidually with each of the 7 algorithms. By definition, only 1 and 2-factor solutions are possible, and so we carried out a pretest of all human-model combinations using PLS to determine the optimal number of factors for each pairing. In what follows, only the better of these two solutions is analyzed.

The average performance for the best pair-wise fusion between individual human subjects and algorithms appears in column 3 of Table 1. Performance for this fusion is reduced only slightly from the performance we found when combining individuals with all seven algorithms. This indicates that the performance of pair-wise fusions can come close to that achieved by fusing humans with all available algorithm experitise. For reference, the average performance of possible pairs of human-algorithm fusions was 0.111, indicating that the best pair-wise fusions were substantially better than the overall average of paired-fusions. This suggests that human-machine fusions are not all equally beneficial.

We note also that in previous work [7], fusing the 7 algorithms (without humans) produced an error rate of .059. This is comparable to the error rate

of individual humans fused with the seven algorithms, but is substantially larger than the near-zero error rate achieved when fusing the average human with the 7 algorithms [7].

Overall, the NJIT algorithm [4] produced the best performance when paired with individual human subjects. Forty of the 49 subjects produced their best fused performance with this algorithm. An additional 8 subjects paired best with Carnegie Mellon University's algorithm [5] and one person paired best with Algorithm B. That said, when comparing the performance of the algorithms individually (column 1 of Table 2) with the performance of the paired human-algorithm fusions (column 3 of Table 2, the error rate reductions are remarkable(column 4 of Table 2 shows percentage error rate reductions for the seven algorithms). On average, the error rate is reduced by 47 percent. Clearly, much of this improvement comes from the reduction of error rate for lesser performing algorithms. It bears noting, however, that even with the large error rate reductions, the average performance of a human alone is only slightly worse than the paired fusions for the lesser performing algorithms. Given that the paired fusions are worthwhile only when they better the error rates of the algorithm or human operating alone, the pairwise fusions of CMU[6] and NJIT [4] are the most promising candidates for fusion.

Finally, the balance of man and machine input in these fusions can be assessed by looking at the PLS-derived weights for the pairwise fusions. These are displayed in Table 3, with the PLS-derived weights for the human and algorithm scores in the first and second columns, respectively. The proportion of human contribution to these fusions is given in column 3. The human contribution varies from a minimum of .0439 for Viisage [6] to almost complete dominance for anonymous algorithms A and C. Notably, the man-machine balance varies widely with the algorithm.

4 Discussion

In most security applications, the use of face recognition algorithms is under the supervision of a human operator. Previous studies indicate that the performance of algorithms can compete with humans in some cases [3]. The accuracy of *individual* humans and *individual* algorithms, however, can be quite variable. In these cases, the question of whether algorithms or humans perform more accurately may be less important than understanding how particular man-machine combinations perform.

In this study, we fused individual algorithms and individual humans and show that these pairwise fusions performed substantially better than the algorithms operating alone. This suggests that humans can contribute, through

Table 3. The balance of human-algorithm contributions in the pairwise fusions. The weights for humans and algorithms appear in the first and second columns, respectively. The third column gives the proportion of human contribution to the overall fusion. This proportion is computed as $\frac{\text{Human}}{\text{Human+Algorithm}}$

Source	human weight	algorithm weight	proportion human
NJIT	−0.2683	−2.4902	0.0973
Viisage	−0.3268	−7.1175	0.0439
CMU	−0.2992	−0.0900	0.7689
Algorithm A	−0.4073	−0.0025	0.9940
Algorithm B	−0.3283	−2.4525	0.1180
Algorithm C	−0.3422	−0.0009	0.9974
Algorithm D	−0.4052	−2.3027	0.1496

fusion, to better face recognition performance. One caveat, however, is that these pairwise fusions must compete, not only with the performance of the algorithm alone, but with the performance of the human alone. From this perspective, pairwise fusions compete with both humans and algorithms only for the best performing algorithms.

Finally, we show that it is important to combine human and algorithm-generated responses in quantitatively precise ways to improve performance optimally. The strong variability in man-machine balance we found across the 7 algorithms illustrates the importance of considering algorithms and humans as individuals.

References

1. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, K., Hoffman, J., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challeng. In: Proceedings of the IEEE Computer Vision and Pattern Recognition. (2005) 947–954
2. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Worek, W.: Preliminary face recognition grand challenge results. In: Proceedings of the Seventh International Conference on Automatic Face Recognition. (2006) 15–24
3. O'Toole, A., Phillips, P., Jiang, F., Ayyad, J., Penard, N., Abdi, H.: Face recognition algorithms surpass humans. submitted (2006)
4. Liu, C.: Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. IEEE:Transactions on Pattern Analysis and Machine Intelligence (2006) 725–737
5. Xie, C., Savvides, M., Kumar, V.: Kernel correlation filter based redundant class-dependence feature analysis (kcfa) on frgc2.0 data. IEEE International Workshop on Analysis and Modeling Faces and Gestures **1** (2005) 32–43

6. Husken, M., Brauckmann, B., Gehlen, S., von der Malsburg, C.: Strategies and benefits of fusion of 2d and 3d face recognition. In 1, ed.: Proceedings of the IEEE Workshop on Face Recognition Grand Challenge Experiments. Computer Society Digital Library. Volume 3., IEEE Press (2005) 174
7. O'Toole, A., Abdi, H., F., J., Phillips, P.: Fusing face recognition algorithms and humans. submitted (2006)
8. Abdi, H.: Partial least squares regression. In Beck, M., A., B., Futing, T., eds.: Encyclopedia for Research Methods in the Social Sciences, Thousand Oaks, CA, Sage (2003) 792–795
9. McIntosh, A., Lobaugh, N.: Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage **23** (2004) 250–263