

# Humans versus Algorithms: Comparisons from the Face Recognition Vendor Test 2006

Alice J. O'Toole  
School of Behavioral & Brain Sciences, GR 4.1  
Richardson, TX 75080-0688 U SA

[otoole@utdallas.edu](mailto:otoole@utdallas.edu)

P. Jonathon Phillips  
National Institute of Standards and  
Technology, 100 Bureau Dr., MS 8940  
Gaithersburg MD 20899, USA

[jonathon@nist.gov](mailto:jonathon@nist.gov)

Abhijit Narvekar  
School of Behavioral & Brain Sciences, GR 4.1  
Richardson, TX 75080-0688 U SA

[aln061000@utdallas.edu](mailto:aln061000@utdallas.edu)

## Abstract

*We present a synopsis of results comparing the performance of humans with face recognition algorithms tested in the Face Recognition Vendor Test (FRVT) 2006 and Face Recognition Grand Challenge (FRGC). Algorithms and humans matched face identity in images taken under controlled and uncontrolled illumination. The human-machine comparisons include accuracy benchmarks, an error pattern analysis, and a test of human and machine performance stability across data sets varying in image quality. The results indicate that: 1.) machines can compete quantitatively with humans matching face identity across changes in illumination; 2.) qualitative differences between humans and machines can be exploited to improve identification by fusing human and machine match scores; and 3.) recognition skills for humans and machines are comparably stable across changes in image quality. Combined the results suggest that face recognition algorithms may be ready for applications with task constraints similar to those evaluated in the FRVT 2006.*

## 1. Introduction

Automatic face recognition algorithms have been developed for decades to be used in security and identity verification applications [1]. To encourage the development of face recognition technology and to provide an independent assessment of algorithm performance, the U.S. Government has sponsored a series of *challenge problems* and *evaluations* [2-5]. A challenge problem can be considered a “homework assignment” meant to assist developers in improving algorithm performance on the types of images to be used in the

subsequent evaluation. An evaluation is considered a “final exam” that takes the form of an objective test of face recognition technology with sequestered images (i.e., images not available in the challenge problem). The most recent cycle of tests consisted of the Face Recognition Grand Challenge (FRGC) [4], which spanned 2004 to 2006, and the accompanying Face Recognition Vendor Test (FRVT) 2006 [5].

The FRGC and the FRVT 2006 are the first large scale face recognition algorithm tests to include a systematic comparison of human and machine performance. In this paper, we provide a synopsis of results comparing human and machine performance from the FRGC and the FRVT 2006 tests. This synopsis is based both on previously published results on the FRGC data [6,7] and new results on the FRVT 2006 data. The FRGC and the FRVT 2006 experiments constitute the most comprehensive study to date that directly compares human and machine performance on face recognition.

The primary rationale for comparing face recognition performance for humans and machines is that humans are currently the most widely deployed face recognition system. For access control, humans match a face in front of them with the photo on an identity card such as a passport or driver’s license. For fraud detection, humans compare photos on applications with photos in databases. For surveillance applications, they compare people to previously available photos of individuals on a “suspect” list or missing persons list. There is also reason to believe that face recognition applications will include a human operator who is augmented or assisted by an algorithm. It is therefore useful to know how accurate humans are relative to algorithms and to know if the pattern of errors for humans and machines is comparable.

All of the experiments we present compared human and machine performance on matching identity in frontal face

images across illumination changes. In an identity matching task, a human or machine answers the question: how likely is it that the two faces are the same person? The response is usually a number reflecting confidence that faces are the same person. Matching across changes in illumination remains a difficult problem both for face recognition algorithms [5,8] and for humans [9]. Although humans have strong capacity limits on the number of faces they can remember, it is generally assumed that human face recognition skills can be remarkably robust to changes in viewing and illumination conditions [9]. One of the goals of automatic face recognition is to develop algorithms that work in more natural environments with limited control of illumination and viewpoint variation. From a technology development perspective, an important goal is to design a sequence of challenge problems of increasing difficulty that will lead to a solution for the general face recognition problem. In the FRGC and the FRVT 2006, in terms of performance, the most challenging problem was matching faces over changes in indoor illumination conditions [4,5]. This is the starting point for the human-machine comparisons we consider.

This paper is organized as follows. In Section 2, we present a brief sketch of the FRGC and the FRVT 2006 (full details of these evaluations appear elsewhere [4,5]). In Section 3, we detail the human-machine comparisons in three parts. First, we present the quantitative comparisons of human and machine performance taken from the FRGC. Second, we consider the results of fusing identity match judgments from the FRGC algorithms with match judgments from humans. This fusion is done to assess the qualitative accord in error patterns for humans and machines. Third, we present human and machine performance data from the FRVT 2006 on two datasets that differ in image quality and demographic composition. This third comparison provides a look at the stability of algorithm and human performance across different types of image sets. In Section 4, we offer some conclusions and discuss future challenges for algorithms.

## 2. FRGC and FRVT 2006

The primary goal of the FRGC was to motivate the development of face recognition algorithms to achieve an order of magnitude improvement in performance over the preceding FRVT 2002 evaluation [3,5]. The baseline for measuring the order of magnitude improvement was assessed on the task of matching frontal face images taken under controlled illumination in the FRVT 2002.

In the FRGC challenge problem (2004-2006), participating researchers were provided with a corpus of images, a set of experiments, ground truth, and code for scoring the performance of their algorithm. In the subsequent FRVT 2006 independent evaluation of face recognition technology, participants submitted executables

to the organization conducting the test (the U.S. National Institute of Standards and Technology, NIST). NIST measured the performance of the submitted algorithms on a set of sequestered images. Participation in the FRVT 2006 required algorithm developers to agree to be identified by name in all reported results. Both the FRGC and the FRVT 2006 were open to algorithm developers from industry and academics, worldwide.

*FRGC Image set.* The images in the FRGC dataset were taken with a 4 Megapixel Canon PowerShot G2<sup>1</sup>. The size of the face in the images was measured as the number of pixels between the centers of the eyes. For the controlled illumination images the average size of the face was 261 pixels between the centers of the eyes and for the uncontrolled illumination images the average size was 144 pixels. The FRGC data contained a very large number of images (approximately 100) of each of 466 people.

*FRVT 2006 Image Set.* Two datasets were used in the FRVT 2006: a *very high-resolution* image set and a *high-resolution* image set. The *very high-resolution* image set was collected at the same institution as the FRGC dataset. It therefore had a similar demographic composition to the FRGC dataset, but with none of the same subjects. The cameras and collection protocols were modified somewhat between the data collection for the FRGC and the FRVT 2006. The very-high resolution images for the FRVT 2006 were taken with a 6 Mega-pixel Nikon D70 camera. The average face size for the controlled images was 400 pixels between the centers of the eyes and 190 pixels for the uncontrolled images. The very high-resolution data contained many images of 335 people.

The FRVT 2006 *high-resolution* dataset was collected at a different institution than the FRGC and FRVT 2006 very high-resolution datasets. The demographics differed substantially from these previous datasets. Full details on the demographic composition of both datasets are available elsewhere [5]. For present purposes, it is worth noting that the very high-resolution dataset was composed primarily of college age subjects, with a strong Caucasian majority. The high-resolution dataset was taken at a workplace and consisted of mostly middle-aged and older adults again with a strong Caucasian majority.

The high-resolution images were taken with a 4 Megapixel Canon PowerShot G2. The average face size for the controlled images was 350 pixels between the centers of the eyes and 110 pixels for the uncontrolled images. The high-resolution data contained images of 257 people. In all other respects, the high-resolution dataset collection protocol was similar to the protocols used for the other datasets.

*Algorithm Task.* The FRGC and FRVT 2006 included several challenge problems and experiments. We focused

<sup>1</sup> The identification of any commercial product or trade name does not imply endorsement or recommendation by the authors or their institutions.



**Fig. 1.** Controlled and uncontrolled illumination face image of the same person.

the human comparisons on the *uncontrolled illumination* problem in which algorithms computed a similarity (i.e., match) score between a pair of frontal face images. In both the FRGC and the FRVT 2006, algorithms matched identity for a large number of face image pairs (see below). In all experiments, one image was taken under controlled illumination and the other image was taken under uncontrolled illumination conditions (Fig. 1). In terms of performance, this was the most difficult problem in the FRGC and the FRVT 2006.

In the machine version of the FRGC face matching experiment, each algorithm matched identity in all pairs of face images between a target set of 16,028 face images and a query set of 8,014 face images. Thus, each algorithm produced a similarity matrix of roughly 128 million similarity scores. The similarity scores indicated an algorithm’s estimate of the likelihood that the faces in the two images are the same person. The similarity matrix was delivered to NIST to be scored. Performance for each algorithm was measured with a receiver operating characteristic curve (ROC). In FRGC, participants’ performance results were reported anonymously unless permission was obtained from a participant.

In the machine version of the FRVT 2006, the uncontrolled illumination experiment was carried out separately for each dataset. For the very high-resolution dataset, 4.3 million similarity scores were computed from a set of 5402 images. For the high-resolution experiment, 7.3 million similarity scores were computed from a set of 7192 images.

### 3. Human Machine Comparisons

#### 3.1. Performance Accuracy Comparison

The first comparison of human and algorithm accuracy was carried out in conjunction with the FRGC uncontrolled illumination experiment [6]. Because it was impossible to collect match data from human subjects on millions of image pairs, the human-machine comparison focused on a subset of the face pairs that were classified as *easy* or *difficult*. Easy and difficult were defined using a

baseline principal components analysis (PCA) algorithm applied to the scaled and aligned face images. This algorithm was used as a baseline because it is well understood and easily available, although it is not considered state-of-the-art. *Easy match pairs* were defined as pairs with similarity scores that were greater than 2 or more standard deviations above the mean for the match scores (i.e., images of the same person that were highly similar); *Difficult match pairs* had similarity scores that were 2 or more standard deviations below the match mean (i.e., images of the same person that were highly dissimilar); *Easy no-match pairs* had similarity scores that were 2 or more standard deviations below the no-match mean (i.e., images of the different people that were highly dissimilar); *Difficult no-match pairs* had similarity scores that were 2 or more standard deviations above the no-match mean (i.e., images of the different people that were highly similar). From the thousands of face pairs that met these criteria, 240 face pairs (half difficult; half easy) were selected randomly for the human experiments. Half of the difficult and easy pairs were *match pairs* (same identity) and half were *no-match pairs* (different identity). We also included equal numbers of male and female faces in each set of face pairs.

Human subjects ( $n = 49$ ) viewed the face pairs for 2 seconds and rated them on the following scale: “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; and 5.) sure they are not the same person.” A human ROC was constructed from these ratings. Similarity scores for the 7 algorithms participating in the FRGC were extracted for the same 240 face pairs judged by humans.

The human and machine ROC curves for the difficult face pairs (Fig. 2) show that three algorithms were more accurate than humans [10,11,12] and four algorithms were less accurate. For the easy face pairs [6], the algorithms and machines were highly accurate, with all but one algorithm performing more accurately than humans.

These results indicate that the best algorithms in the FRGC can compete with humans matching face identity across changes in illumination. The comparability of human and machine performance is especially interesting given that in FRGC, the uncontrolled illumination task was the most difficult in terms of performance [4, 5]. The finding suggests that even with performance that is far from perfect, algorithms may actually improve security in some applied settings where humans are currently performing the task.

#### 3.2. Qualitative Comparisons via Fusion

The finding that algorithms can compete with humans on this task opens up the question: Do humans and algorithms make similar errors? A prerequisite to addressing this

question is to determine whether different algorithms show similar error patterns. We did this by fusing the performance of algorithms and humans to determine the extent to which combining the similarity estimates of humans and machines can improve performance. If error patterns are highly similar, there is little to be gained by fusion. If error patterns differ, however, fusion may be able to exploit these differences to improve performance.

First, to address the prerequisite question, we applied a fusion algorithm to the face similarity estimates for the 120 difficult face pairs generated by the 7 algorithms tested in the FRGC [6]. Only the difficult face pairs were used in the fusion due to the low error rates found for the easy face pairs in this study. The second step was to include human-generated similarity estimates for the difficult face pairs as “an additional algorithm”.

Fusion was performed by partial least squares (PLS) regression, a statistical learning technique that generalizes and combines features from principal component analysis and multiple regression [13]. The technique is used to predict a set of dependent variables from a set of independent variables (predictors). In this application, PLS acts like a classifier that learns to predict the match status of face pairs (same or different person) by an optimal combination of the input similarity estimates from algorithms and/or humans.

To fuse the algorithms, the input consisted of face pair similarity estimates generated by the 7 algorithms and the output of the classifier was the match status of the face pair. The PLS classifier method was applied with a robustness simulation as follows. PLS solutions were derived from 119 of the face pairs and the match status prediction was tested with the “left-out” pair. This was done 120 times rotating the left-out face pair, with error rate defined as the fraction of left-out pairs incorrectly classified.

For the pre-requisite problem, fusing the 7 algorithms reduced the error rate by a factor of two over the best algorithm operating alone [7]. Specifically, the fused error rate was 0.059, whereas the best-performing single algorithm [10] achieved an error rate of 0.12. The pattern of errors differed enough across algorithms, therefore, for fusion to benefit the overall performance.

In evaluating the similarity of the error patterns, an advantage of using PLS for the fusion is that it yields a set of weights for each component (each algorithm) in the success of the fusion. These weights illustrate where there may be qualitative differences in strategy among the algorithms. Of note, it was not necessarily the best algorithms that were weighted most strongly in the fusion. Instead, algorithms that performed less well on their own can contribute to fusion by succeeding on different face pairs than the better performing algorithms [cf., 7 for details of which algorithms combined best to produce the performance improvements].

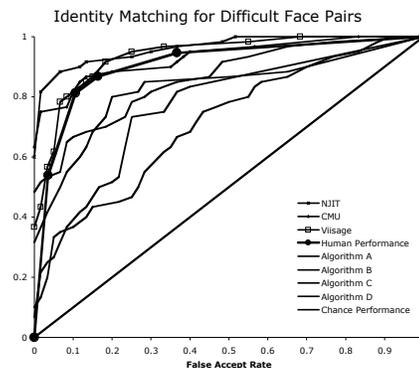


Fig. 2. Human and machine performance in FRGC.

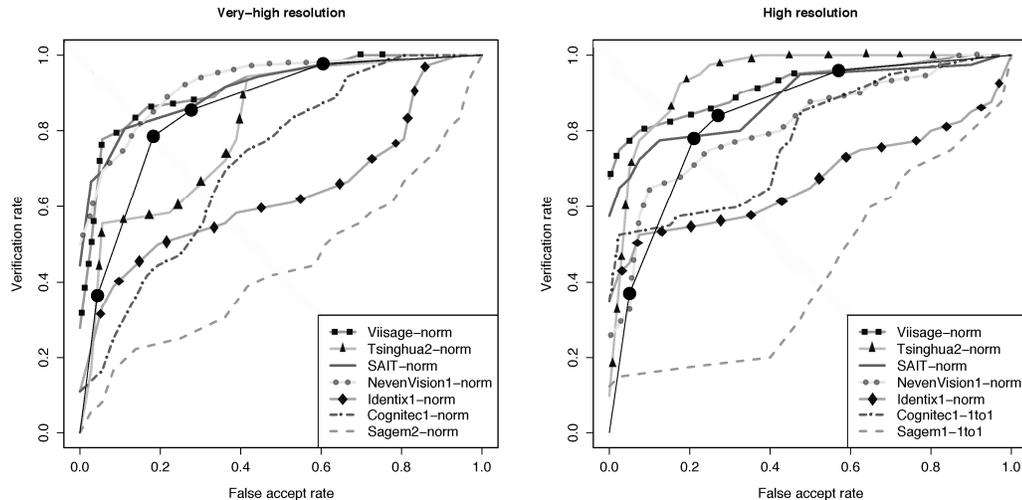
Next, we fused the algorithm face similarity estimates with a human-generated estimate of face similarity. The human similarity estimate was calculated using the match ratings supplied by the 49 subjects in the previous study (cf., Section 3.1 [6]). As noted, subjects in that experiment rated the likelihood that the people in the controlled and uncontrolled face pairs were the “same person” or “different people” using a 5-point scale. The average rating for each pair across the 49 human subjects served as its human similarity estimate. In the human-machine fusion, the human similarity estimate for each of the face pairs was included with the similarity estimates of the 7 algorithms and input to the PLS. Match status predictions were evaluated with a jack-knife procedure.

The results indicated that fusing the human similarity score with the 7 algorithm scores reduced error rate to nearly zero (.008) from 0.12 for humans only. This finding suggests that the pattern of errors made by humans diverges enough from the algorithm error patterns for fusion to exploit the best of both types of strategies.

Overall, the findings support two conclusions. First, machine performance can be improved by fusing together algorithms with different error profiles. Second, humans perform the task in ways that complement the strengths of the algorithms. The error rate reduction to near zero suggests that it is possible to exploit a human contribution to optimize human-algorithm partnerships. Note however that optimally weighted combinations of humans and algorithms must be found empirically for different algorithms and possibly different humans also.

### 3.3. Performance Stability Over Datasets

An important difference between the FRGC and the FRVT 2006 was that the FRVT 2006 tested algorithms with sequestered data from two *different* databases: a high-resolution and very high-resolution dataset. These varied also in demographic composition. As noted, the algorithms in the FRVT 2006 did not have access to any



**Fig. 3.** Performance stability for algorithms and humans (large black circles) with the datasets used in the FRVT 2006.

of the identities in the preceding FRGC test. Both the resolution differences and the relatively large demographic differences between the databases presented us with the opportunity to evaluate the stability of algorithm performance for face populations that are not well matched to the training data (i.e., the FRGC images). The use of two datasets in the FRVT 2006 allowed us to compare the stability of algorithm and human performance on different datasets.

We assessed the stability of human and algorithm performance across the *high resolution* and *very high-resolution* datasets used for the uncontrolled illumination experiment in the FRVT 2006. In the previous work, we selected *easy* and *difficult* face pairs based on the performance of a PCA baseline algorithm. Here, we used the performance of the 7 algorithms tested in the FRVT 2006 on the uncontrolled illumination problem to find moderately difficult pairs. Specifically, a difficulty score was assigned to each face pair based on the number of algorithms that incorrectly assigned the match status of the pair at the false accept rate of 0.001. We selected face pairs from the middle range of algorithm performance. These were pairs erroneously judged by between 3 and 5 of the 7 algorithms. We selected the image pairs randomly from the pairs meeting this criterion. For the *high-resolution* human experiment, 40 match pairs and 40 no-match pairs (half female and half male) were selected. For the *very high-resolution* experiment, 36 match pairs and 36 no-match pairs (half female and half male) were selected.

In both experiments, subjects viewed pairs of faces for two seconds each and rated them on the same 1-5 identity match scale used previously. Twenty-eight subjects rated face pairs in the *high-resolution* experiment and 25 subjects rated face pairs in the *very high-resolution*

experiment. Human performance was measured separately for each experiment using an ROC curve.

The results for the participating algorithms were tallied as before, by selecting the same face pairs presented to humans and creating an ROC curve for each algorithm on these face pairs. The combined human-machine results appear in Fig. 3. Several points are worth noting. Again, consistent with the previous finding with the FRGC comparison, machine performance is in the range of human performance, with the best algorithms surpassing humans. This finding replicates the previous one and extends it to “moderately difficult” face pairs from the high and very high-resolution datasets. It is further worth noting that the findings of comparable performance for machines and humans holds even with a face pair sampling procedure that was substantially different from the PCA-based procedure used in the previous work.

On the question of performance stability across the two datasets, humans were quite stable across the datasets. The performance of algorithms is also mostly stable. This human-machine experiment indicates performance stability over variation in the size of the face and across the age demographics of the two image sets.

We qualify these results in three ways. First, the differences in face image size that we considered are small in comparison to the range algorithms encounter in many applications (e.g., video surveillance).

Second, the demographic differences considered here were restricted to age change. Different challenges may apply in cases where race and ethnicity vary between databases and between intended application venues. Preliminary work looking at the FRVT 2006 performance over sets of East Asian and Caucasian faces supports this conclusion.

Third, although the performance of algorithms on this

well-controlled sample of moderately difficult face pairs was stable across datasets, in the FRVT 2006 report, algorithm stability as measured on the entire database was less impressive, with some large performance differences on the two datasets [5]. Moreover, there was no consistent advantage for algorithms on either the high- or very high-resolution datasets. Three algorithms performed better on the high-resolution database, two algorithms better on the very high-resolution database, and two performed comparably on the two datasets. The stable performance we observed for algorithms was for moderately difficult face pairs. Thus, despite the variability of algorithm performance over the two image sets, both in relative and absolute terms, it is possible to pick an image set on which performance is stable across the two sets.

#### 4. Discussion

In comparing human and machine performance in the FRVT 2006 and FRGC, we come to three conclusions. First, since FRVT 2002 there has been significant improvement in the performance of algorithms matching identity in frontal face images across changes in illumination. In 2002, algorithms were not capable of surpassing human performance on this problem. The present data indicate that the best algorithms tested in the FRGC and FRVT 2006 can outperform humans.

Second, the human-machine fusion experiment reveals at least some qualitative differences in the pattern of errors generated by the different algorithms and by humans. Optimal partnerships can be constructed by appropriately fusing algorithms and humans using parameters determined by empirical testing.

Third, the relative stability of the algorithms and humans across datasets is encouraging. This stability provides evidence that human performance could serve as a method for rating the difficulty of image sets. Concomitantly, the human rated image sets could be one factor in designing a series of challenge problems with the ultimate goal of solving the general face recognition problem.

Finally, it is important to note that the task carried out by humans in these experiments is one of “unfamiliar face recognition”. This is an appropriate comparison for algorithms, because the face recognition tasks done by human security guards are also with unfamiliar faces. Human face recognition skills are at their best for highly familiar faces. These are faces we have seen many times under many different viewing conditions. In these cases, humans can recognize faces in very poor viewing conditions. It is therefore a reasonable next step to try to understand how humans recognize familiar faces and to begin to raise machine performance to this level. This would allow for algorithms to operate with high levels of accuracy in unconstrained environments.

#### 5. References

- [1] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35:399-459, 2003.
- [2] P.J. Phillips, H. Moon, P. Rizvi, and P. Rauss. The FERET evaluation method for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Learning*, 22:1090-1104, 2000.
- [3] P.J. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J.M. Bone. FRVT 2002 evaluation report. Tech. Rep. NISTIR 6965 <http://www.frvt.org>, 2003.
- [4] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. *Proc. IEEE Computer Vision & Pattern Recognition*, 1:947-954, 2005.
- [5] P.J. Phillips, W. T. Scruggs, A.J. O’Toole, P.J. Flynn, K.W. Bowyer, C. L. Schott, and M. Sharpe. FRVT-2006 and ICE-2006 Large-Scale Results. Submitted.
- [6] A.J. O’Toole, P.J. Phillips, F. Jiang, J.J. Ayyad, N. Pénard, and H. Abdi. Face recognition algorithms surpass humans matching faces across changed in illumination. *IEEE Transactions on Pattern Analysis and Machine Learning*, 29:1642-1646, 2007.
- [7] A.J. O’Toole, H. Abdi, F. Jiang, and P.J. Phillips. Fusing face recognition algorithms and humans. *IEEE Transactions on Systems, Man & Cybernetics*, 37:1149-1155, 2007.
- [8] R. Gross, S. Baker, I. Matthews, and T. Kanade. Face recognition across pose and illumination. *Handbook of Face Recognition*, (S.Z. Li and A.K. Jain, Eds.) Springer:193-216, 2005.
- [9] A.J. O’Toole, F. Jiang, D. Roark, and H. Abdi. Predicting human face recognition. *Face Processing: Advanced methods and models*, W-Y. Zhao and R. Chellappa, Eds. Elsevier, 2006.
- [10] C. Liu. Capitalize on dimensionality increasing techniques from improving face recognition Grand Challenge performance. *IEEE Transactions on Pattern Analysis and Machine Learning*, 28: 725-737, 2006.
- [11] C.M. Xie, M. Savvides, and V. Kumar. Kernel correlation filter based redundant class-dependence feature analysis (KCFA) on FRGC2.0 Data. *IEEE International Workshop Analysis & Modeling Faces & Gestures*, 32-43, 2005.
- [12] M. Husken, B. Brauckmann, S. Gehlen, and C. von der Malsburg. Strategies and benefits of fusion of 2D and 3D face recognition. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, 3:174, 2005.
- [13] H. Abdi. Partial least squares regression (PLS-regression). In M. Lewis Beck, A. Bryman, T. Futing (Eds.) *Encyclopedia for Research Methods for the Social Sciences*. Thousand Oaks: CA, Sage, 2003. pp. 792-795.

**Acknowledgements.** This work was funded by a TSWG contract to A. O’Toole and H. Abdi. P. J. Phillips was supported in part by the National Institute of Justice. Thanks are due to Julianne Ayyad, David Raboy, and Sam Weimer for work on figures and for subject testing.