# Fusing Face-Verification Algorithms and Humans

Alice J. O'Toole, Hervé Abdi, Fang Jiang, and P. Jonathon Phillips, *Senior Member, IEEE*

***Abstract*—It has been demonstrated recently that state-of-the-art face-recognition algorithms can surpass human accuracy at matching faces over changes in illumination. The ranking of algorithms and humans by accuracy, however, does not provide information about whether algorithms and humans perform the task comparably or whether algorithms and humans can be fused to improve performance. In this paper, we fused humans and algorithms using partial least square regression (PLSR). In the first experiment, we applied PLSR to face-pair similarity scores generated by seven algorithms participating in the Face Recognition Grand Challenge. The PLSR produced an optimal weighting of the similarity scores, which we tested for generality with a jackknife procedure. Fusing the algorithms' similarity scores using the optimal weights produced a twofold reduction of error rate over the most accurate algorithm. Next, human-subject-generated similarity scores were added to the PLSR analysis. Fusing humans and algorithms increased the performance to near-perfect classification accuracy. These results are discussed in terms of maximizing face-verification accuracy with hybrid systems consisting of multiple algorithms and humans.**

***Index Terms*—Face and gesture recognition, human information processing, performance evaluation of algorithms and systems.**

## I. Introduction

THE FIELD of automatic face-recognition algorithms has expanded in the past decade from consisting of simple algorithms that operate on highly controlled images of faces to more sophisticated algorithms aimed at operating in the natural conditions that characterize most security applications. One particularly difficult challenge in advancing algorithms from controlled to natural environments has been the problem of operating over substantial changes in illumination. The computational difficulties posed by the illumination problem have been well documented in the automatic face-recognition (cf. [1]–[3]) and human-perception literature [4]–[6].

In more practical terms, the performance of face-recognition algorithms in controlled and uncontrolled illumination environments was assessed recently in the Face Recognition Grand Challenge (FRGC), a U.S. Government-sponsored test of face-recognition algorithms aimed at fostering algorithm development [7], [8]. The FRGC (2004–2006) included academic, industrial, and research laboratory competitors. Competitors participated in the program by volunteering to have their algorithms tested on one or more of six face-matching experiments varying in difficulty. The set of experiments included both a controlled-illumination face-matching experiment and a more difficult experiment where algorithms matched face identity in images taken under different illumination conditions. Because the FRGC tested multiple algorithms simultaneously using a standardized evaluation protocol and a common image set, it provides a useful time-locked look at the performance of state-of-the-art face-recognition algorithms.


Fig. 1. Sample pair of face images from a "match" trial. Participants responded by rating the likelihood that the pictures were of the same person using a five-point scale ranging from "1) sure they are the same person" to "5) sure they are not the same people."

The difficulty of the illumination problem can be seen clearly by comparing the performance of the algorithms in the controlled and uncontrolled illumination experiments of the FRGC. In both cases, the task of the algorithms was to decide for each of a large number of face pairs (>128 million), whether the images were of the same person or of different people. In the controlled-illumination experiment, the illumination conditions were the same for both images in the pair. In the uncontrolled-illumination experiment, one image was taken under controlled-illumination conditions, and the other was taken under uncontrolled illumination (see Fig. 1 for a sample image pair).

Twenty algorithms competed in the controlled-illumination experiment and achieved an average verification rate of 0.91 at 0.001 false-accept rate. By contrast, in the uncontrolled-illumination experiment, only seven algorithms participated, achieving an average verification rate of 0.41 at 0.001 false-accept rate. The difference in participant numbers and average performance in these experiments is evidence that the illumination problem continues to challenge face-recognition algorithms.

A rather different perspective on the relatively poor performance of algorithms in the uncontrolled-illumination experiment comes from comparing the algorithms to humans

Manuscript received May 25, 2006. The work of A. O'Toole and H. Abdi was supported by a contract from the Technical Support Working Group. The work of P. J. Phillips was supported in part by the National Institute of Justice. This paper was recommended by Guest Editor K. Bowyer.

A. J. O'Toole, H. Abdi, and F. Jiang are with the School of Behavioral and Brain Sciences (GR4.1), The University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: otoole@utdallas.edu; herve@utdallas.edu; fxj018100@utdallas.edu).

P. J. Phillips is with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: jonathon@nist.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2007.907034

performing a comparable task. In a recent study [9], human-face-matching performance was compared to the performance of the seven algorithms participating in the uncontrolled-illumination matching experiment of the FRGC. We describe this previous study in some details, here, because it provides the fusion data used in this paper.

### A. Source of Fusion Data

Algorithms in the FRGC uncontrolled-illumination experiment (experiment 4 in FRGC nomenclature) matched face identities in all possible pairs of 16 028 target images and 8014 probe images, with target images taken under controlled-illumination conditions and probe images taken under uncontrolled-illumination conditions (see Fig. 1 for a sample pair). The output for each algorithm was a matrix of similarity scores for all possible pairs of faces. For each algorithm, a receiver operating characteristic (ROC) curve was generated from the similarity score matrix. The performance of the seven algorithms was compared using these ROC curves (cf. [9] for complete results).

The primary difficulty in comparing the performance of humans to algorithms in the FRGC is the implausibly large number of face-pair comparisons required for an exhaustive comparison. Therefore, to compare the performance of humans to algorithms, face pairs were sampled from the matrix by selecting a set of the easiest and most difficult pairs [9]. In this paper, we concentrate on the most difficult image pairs. In both cases, however, the sampling was done with the help of a control algorithm based on a principal component analysis (PCA) of the aligned and scaled face images. Using this algorithm, easy match pairs were defined based on similarity scores that were substantially greater than the mean for the distribution of matched face pairs, i.e., highly similar images of the same person. Difficult match pairs were those with similarity scores substantially lower than the match mean, i.e., highly dissimilar images of the same person. Easy and difficult nonmatch pairs were defined inversely.

Human subjects matched the identity of 240 sample face pairs by rating their certainty that the pairs were of the same person. Human responses ranged on a five-point scale from "certain the two images are of the same person" to "certain that two images are not of same person." The rating data allowed for the generation of a ROC curve for human performance that was comparable to the ROC curves derivable from the performance of the algorithms.

The human–machine comparison was conducted by extracting the algorithms' similarity scores for the same face pairs tested in the human-face-matching experiment. These were plotted on ROC curves along with human match-accuracy data [9]. The results demonstrated clearly that three algorithms [10]–[12] surpassed human performance on the difficult face pairs. Of these, the algorithm from The New Jersey Institute of Technology [10] and the algorithm from Carnegie Mellon University [11] have been published. Details on the third algorithm, from the Viisage Corporation,[1] are only partially available [12].

[1]See Acknowledgment.

In addition to the finding that three algorithms were competitive with humans on the difficult pairs of faces, all but one algorithm surpassed human performance on the easy face pairs. Combined, these findings suggest that, although the algorithm performance on the uncontrolled-illumination experiment in the FRGC may be poor in absolute terms, it is nonetheless competitive with the human performance. This comparison is of interest due to the fact that humans are currently performing this task in most applied situations. This previous study forms the base of this paper.

### B. Rationale for Fusion

Although the quantitative ranking of human performance relative to a set of algorithms provides a useful benchmark, this ranking does not offer any insight into whether algorithms recognize faces in ways that are similar to humans. The FRGC showed that algorithms performed poorly on face recognition in uncontrolled-illumination environments. Our previous work showed the same result for humans. If algorithms and humans take diverse approaches to the problem of face matching, it is possible that an appropriate fusion of algorithms and humans can yield better performance than a single algorithm or the fusion of multiple algorithms. Indeed, previous work has shown that fusing the multiple face-recognition algorithms improves performance over a single algorithm (cf. [13]–[15]). However, no previous studies have fused human and algorithm performance.

In the majority of applications for face recognition, a human operator is present and involved in the decision process. Thus, it may be of general value to optimize system performance by explicitly incorporating human-face-recognition capabilities into the decision process. Toward this end, we present a methodology for fusing algorithm and human performance.

In this paper, we asked two questions. First, can performance be improved by fusing algorithms from the FRGC uncontrolled-illumination experiment? Second, does fusing humans and algorithms improve performance above the level achieved by the algorithm fusion? The availability of multiple algorithm estimates of face similarity, in conjunction with analogous human estimates of similarity, offers the possibility of exploring these questions in a more systematic way than generally possible. Here, we investigated the possibility of fusing face-similarity estimates from algorithms and humans to improve face-matching performance.

Fusion was performed by partial least square regression (PLSR), a statistical technique that generalizes and combines features from the PCA and multiple regression [16], [17]. The technique is used to predict a set of dependent variables from a set of independent variables (predictors). The choice of PLS is, in part, arbitrary, because other pattern classification or neural network techniques will give comparable results. We used the PLS because it has the advantage of providing easily interpretable weights for individual predictors (see as follows). Although the PLS is less well known in pattern-recognition literature, it is widely used in chemometrics, sensory evaluation, and for neuroimaging data analysis (cf. [16], [18], [19],

and [21]). We give complete algorithm details for the PLSR algorithm in the Appendix.

In this paper, algorithm and human estimates of face similarity were the predictors, and the match status of individual face pairs (i.e., same person or different people) was the dependent variable. The PLSR gives a set of orthogonal factors, sometimes called latent vectors $\{\mathbf{t}_1, \ldots, \mathbf{t}_l\}$, 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 $\mathbf{B}_{\mathrm{pls}}$ in the PLSR literature [16]. To fuse algorithms, the weights prescribed in the latent vector(s) are used to combine the similarity scores from each of the seven algorithms to produce an estimate of the match status for the face pairs. When fusing humans and algorithms, there are eight predictors: seven from the algorithms and one from the averaged human data.

The predictive power of these factors is generally assessed 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 $\mathbf{B}_{\mathrm{pls}}$. The larger the number of factors kept, the better the prediction of the "learning set." In general, however, a smaller number of factors is optimal for robust prediction (i.e., for test-set predictions).

In the first experiment, we applied the PLS to the similarity scores generated by seven algorithms that participated in the FRGC uncontrolled-illumination experiment. We tested the generality of the optimal weights found in the analysis for predicting face-match status using a jackknife procedure. In the second experiment, we added human-generated similarity scores to the algorithms' scores and measured the contribution human estimates make to the fusion.

## II. Procedure

### A. Stimuli

Face stimuli were chosen from a large database developed for the FRGC study [7], [8]. The uncontrolled-illumination probe faces had a resolution of 2272 $\times$ 1704 pixels. The controlled-illumination target faces had a resolution of 1704 $\times$ 2272 pixels. For the present analyses, we used the same set of difficult face pairs sampled for the previous quantitative comparison between humans and algorithms [9]. These were sampled from the 128 448 392 pairs available, which included 407 352 (0.32%) match pairs (i.e., image pairs of the same person) and 128 041 040 (99.68%) nonmatch pairs (i.e., image pairs of different people). To eliminate the possibility that humans could base identity comparisons on the surface facial characteristics associated with race or age, all images in the study were of faces of Caucasian males and females in their twenties. All pairs were matched by sex. Although these demographic choices have consequences for the comparison of humans versus algorithms in absolute terms, these choices were best suited with the goal of the previous study [9].

In this paper, only "difficult face pairs" were included. These were chosen using a control algorithm based on the PCA of the aligned and scaled images. Specifically, difficult match face pairs ($n = 60$) were sampled randomly from match pairs that had similarity scores less than two standard deviations below the match mean. Difficult nonmatch face pairs ($n = 60$) were sampled randomly from nonmatch pairs that had similarity scores greater than two standard deviations above the nonmatch mean.

The validity of the PCA as a prescreening algorithm for humans and algorithms was supported in the previous study [9]. The PCA algorithm reliably predicted "easy" and "difficult" sets of face pairs for humans in three experiments [9]. All seven algorithms were likewise more accurate on the PCA-screened easy face pairs than on the PCA-screened difficult faces [9]. The PCA, therefore, can serve as a useful sampling tool, even though it is not considered "state-of-the-art." We did not use the algorithms available from the FRGC, which perform more accurately than PCA, because of the potential to bias the success of particular algorithms in the algorithm–human evaluation [9].

### B. Human-Subject Judgments of Face Similarity

The human-subject data for this experiment were collected in an experiment in which subjects viewed the image pairs and rated the likelihood that the images were of the same person or of different people [9]. For completeness, we sketch out the methods used in that study. There are 49 subjects (25 males and 24 females) that viewed the 120 pairs of faces for 2 s each and responded by rating each pair on the following scale: 1) sure that the pictures are of the same person; 2) think that the pictures are of the same person; 3) do not know; 4) think that the pictures are not of the same person; and 5) sure that the pictures are not of the same person. Of the 120 pairs, half were match pairs and half were nonmatch pairs. Equal numbers of male and female pairs were included in the match and nonmatch conditions. The subjects were instructed to examine the face images and to determine whether the images were of the same person or of different people. Subjects were not informed about the proportion of match versus nonmatch trials nor were they given practice trials. The image pairs were presented for 2 s, but there was no time limit for entering a response.

For each pair of faces, the average rating was computed across the 49 subjects. This average served as the human similarity score for that pair of faces in the PLSR.

### C. Algorithms' Judgments of Face Similarity

The similarity scores of the 120 difficult face pairs presented to participants in the human experiment were extracted from each algorithm's 16 028 $\times$ 8014 similarity matrix. These scores served as the algorithm data for the PLSR.

## III. Results

### A. Experiment 1—Algorithm Fusion by PLSR

The similarity scores for the seven algorithms for the 120 difficult face pairs (60 match and 60 nonmatch) were combined in a columnwise matrix. The dependent variable was a 120-element vector containing the match status (+1 for match

TABLE I
WEIGHTS FOR ALGORITHM FUSION DIFFICULT FACE PAIRS

| | *NJIT* | *Viisage* | *CMU* | *A* | *B* | *C* | *D* |
|---|---|---|---|---|---|---|---|
| *weights* | -2.2 | -1.81 | -.05 | .00 | -.15 | .00 | .16 |
| *error rates* | .12 | .20 | .14 | .37 | .23 | .25 | .36 |

The table shows the weights for combining algorithm similarity estimates for optimal match performance (top row). Large absolute values indicate the most useful predictors, which in this case are the NJIT algorithm [10] and Viisage [12]. The bottom row gives the proportion of classification errors for the algorithms *individually*. The fusion cuts the best algorithm's error rate by a factor of two, from .12 to .059.

and −1 for nonmatch) for each face pair. PLSR was applied simultaneously to the combined similarity score and match-status data matrices.

We varied the number of PLSR factors retained from one to five and found a three-factor solution to be optimal. Retaining three factors indicates that the first three latent vectors, which are ordered according to the proportion of variance explained in the covariance matrix, are combined linearly to specify the weights for combining the similarity scores.

A robust performance estimate was determined with a jackknife simulation. We started with the 120 face pairs available and systematically deleted each face pair in turn, recomputing the PLSR with the remaining 119 pairs of faces. We tested the match-status predictions for the PLSR solutions derived from 119 pairs of faces on the "left-out" face pair. This yielded 120 generalized match-prediction tests. The error rate we report is the fraction of left-out face pairs incorrectly classified according to match status.

Error rates for classification with one through five factors were 0.067, 0.075, 0.059, 0.067, and 0.083, respectively. These error rates are all lower than the minimum error rate achieved by any single algorithm operating alone (cf. Table I for error rates for each individual algorithm). Specifically, the data indicate that fusion, following the optimal weighting derived with the PLSR, cuts the error rate of the best performing algorithm (NJIT [10] with a 0.12 error rate) by a factor of two.

For purposes of interpretation, the weights for combining similarity scores appear in Table I. These weights are used to combine the similarity scores from the seven algorithms to achieve a maximal separation between the match and nonmatch face-pair distributions. Algorithms with weights that have large absolute values are the most useful in improving the performance with fusion.

Using this as an interpretation guide, it is clear that most of the improvement in accuracy comes from combining just two algorithms, NJIT [10] and Viisage [12], whose weights have the largest absolute values. This might be due to these algorithms having maximally diverse strategies in computing the face similarity. This interpretation seems likely given that the CMU algorithm [11] performed somewhat better than the algorithm of Viisage [12]. Thus, more benefit can be derived from combining lesser performing algorithms that operate in different fashions than by combining higher performing similar algorithms.

TABLE II
WEIGHTS FOR HUMAN–ALGORITHM FUSION

| | *Human* | *NJIT* | *Viisage* | *CMU* | *B* | *D* |
|---|---|---|---|---|---|---|
| *weights* | .47 | -1.29 | -.71 | -.03 | -.12 | .20 |

The table shows the weights for combining human and algorithm similarity estimates for optimal match performance. Algorithms A and C had weights of zero and are not included in the table. The addition of humans to the PLSR decreased the error rate from .059, for algorithm fusion, to .008 for human-algorithm fusion.

### B. Experiment 2—Human and Algorithm Fusion by PLSR

Can fusing humans and algorithms add to the accuracy of the match estimates and further improve classification over that obtained with the fused algorithms? In this experiment, we added human similarity estimates to the PLSR model. The analysis proceeded as before but with a column vector containing the averaged human similarity data appended to the predictor matrix.[2]

Again, we varied, from one to five, the number of PLSR factors we retained. In this case, we found a two-factor solution to be most robust, using the jackknife procedure described previously. The weights for combining human and algorithm similarity estimates are shown in Table II. Performance with one factor through five factors yielded classification error rates of 0.042, 0.008, 0.033, 0.033, and 0.042, respectively.

These results illustrate that it is possible to obtain nearly perfect classification, when humans are added into the predictor matrix. This suggests that human strategies for assigning similarities to faces add usefully to those employed by the best algorithms. In particular, this result shows that human similarity ratings provide specific information about the face-pair comparisons that are not available from any of the algorithms.

It is worth noting from previous work [9] that the accuracy of humans was found to be below that of NJIT [10], CMU [11], and Viisage [12] but above the accuracy of algorithms A, B, C, and D. In that study, similarity ratings from individual subjects were collapsed across the 120 face pairs to create an ROC curve for each subject. These individual ROC curves were then averaged to give an overall estimate of human accuracy. Here, we averaged the similarity ratings for 120 face pairs, collapsing across the individual subjects. Interestingly, although perhaps not surprisingly, we found that by averaging across the 49 human subjects' estimates of face similarity for each face pair individually, human error rate was 0.12, comparable to NJIT, which is the best algorithm. This suggests that individual subjects, like algorithms, may employ diverse strategies for judging the similarity of the face pairs. By consequence, combining the similarity estimates of individual subjects by fusion could likewise benefit accuracy.

## IV. DISCUSSION

Fusing humans and algorithms is a reasonable goal for face-recognition researchers and corporations with hopes of

[2]The direction of the similarity scores for the humans was inverted as compared to the algorithms, so for interpretation purposes, attention should be paid only to the absolute values of the PLSR weights.

applying their systems to real applications. Knowing how accurately algorithms and humans are by themselves is a start in trying to estimate how well combinations of algorithms and humans will work. However, quantitative measures of accuracy for individual algorithms and humans are not sufficient in guiding the development of hybrid systems. This paper illustrates that the most useful fusions of algorithms and humans are likely to come from combining face-recognition systems (algorithms or humans) with diverse face-recognition strategies.

In this paper, we demonstrated that fusing algorithms and humans substantially improved performance on a difficult face-matching task. The use of PLSR to fuse the algorithms and humans also yielded a precise indication of how to combine the individual components of the fusion optimally. This weight vector serves simultaneously as a recipe for fusing systems and as an indicator of the similarity of algorithm and human strategies for face verification.

Given that neither algorithms nor humans perform face recognition well in uncontrolled environments and that a majority of applications have a human operator in the loop, a reasonable goal of researchers should be to design face-recognition strategies that optimally combine algorithms and humans. Fusion of algorithms and humans to create good hybrids can, therefore, be a useful and practical approach to improving face-matching performance in important applications.

## APPENDIX

In this Appendix, we give a brief description of the PLSR. A more complete presentation can be found in previous works [16], [20]. MATLAB programs can be downloaded from www.utdallas.edu/~herve. The PLSR generalizes and combine, features from PCA and multiple regression. Its goal is to optimally predict a set of dependent variables from a set of predictors. Specifically, PLSR searches for a set of components (called *latent* vectors) that performs a simultaneous decomposition of $\mathbf{X}$ and $\mathbf{Y}$ with the constraint that these components explain as much as possible of the *covariance* between $\mathbf{X}$ and $\mathbf{Y}$. This step is followed by a regression step where the decomposition of $\mathbf{X}$ is used to predict $\mathbf{Y}$.

### A. Notation

The $I$ observations described by $K$-dependent variables are stored in an $I \times K$ matrix denoted by $\mathbf{Y}$, and the $I \times J$ matrix of predictors is denoted $\mathbf{X}$. Without loss of generality, both $\mathbf{X}$ and $\mathbf{Y}$ are assumed to be centered and normalized. The common set of (orthogonal) latent vectors is stored in the $I \times L$ matrix $\mathbf{T}$ (i.e., $\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{I}$). PLRS decomposes $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}}$$

where $\mathbf{P}$ is a $J \times L$ matrix called the $\mathbf{X}$-loading matrix. The matrix $\mathbf{Y}$ is estimated as

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{C}^{\mathrm{T}}$$

where $\mathbf{B}$ is a diagonal matrix with the "regression weights" as diagonal elements, and $\mathbf{C}$ is the "weight matrix" of the dependent variables.

### B. Computations of Latent Vectors, Loadings, and Weights

A latent vector is obtained by finding two sets of weights $\mathbf{w}$ and $\mathbf{c}$ in order to create (respectively) a linear combination of the columns of $\mathbf{X}$ and $\mathbf{Y}$ such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors

$$\mathbf{t} = \mathbf{X}\mathbf{w} \quad \mathbf{u} = \mathbf{Y}\mathbf{c} \tag{1}$$

under the constraint that

$$\mathbf{w}^{\mathrm{T}}\mathbf{w} = 1 \quad \mathbf{t}^{\mathrm{T}}\mathbf{t} = 1 \quad \mathbf{t}^{\mathrm{T}}\mathbf{u} \quad \text{be maximal.} \tag{2}$$

When the first latent vector has been found, it is subtracted from both $\mathbf{X}$ and $\mathbf{Y}$, and the procedure is iterated until $\mathbf{X}$ becomes a null matrix (see the algorithm section for more).

### C. Algorithm

The different components of PLSR can be found by a series of singular-value decompositions, each followed by a deflation. Specifically, the first weight vectors $\mathbf{w}$ and $\mathbf{c}$ are, respectively, the first right and left singular vectors of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. Vectors $\mathbf{t}$ and $\mathbf{u}$ are then derived using (1). With these vectors, the value of $b$ is computed as $b = \mathbf{t}^{\mathrm{T}}\mathbf{u}$ and then used to predict $\mathbf{Y}$ from $\mathbf{t}$ as $\mathbf{Y} = b\mathbf{t}\mathbf{c}^{\mathrm{T}}$. The factor loadings for $\mathbf{X}$ are computed as $\mathbf{p} = \mathbf{X}\mathbf{t}$. Now, subtract (i.e., partial out) the effect of $\mathbf{t}$ from both $\mathbf{X}$ and $\mathbf{Y}$ as follows: $\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^{\mathrm{T}}$ and $\mathbf{Y} = \mathbf{Y} - b\mathbf{t}\mathbf{c}^{\mathrm{T}}$. The vectors $\mathbf{t}$, $\mathbf{u}$, $\mathbf{w}$, $\mathbf{c}$, and $\mathbf{p}$ are then stored in the corresponding matrices, and the scalar $b$ is stored as a diagonal element of $\mathbf{B}$. If $\mathbf{X}$ is a null matrix, then the whole set of latent vectors has been found; otherwise the procedure is repeated.

### D. Prediction of the Dependent Variables

The dependent variables are predicted using the multivariate regression formula defined as

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{C}^{\mathrm{T}} = \mathbf{X}\mathbf{B}_{\mathrm{PLS}} \tag{3}$$

with

$$\mathbf{B}_{\mathrm{PLS}} = \mathbf{P}^{\mathrm{T}+}\mathbf{B}\mathbf{C}^{\mathrm{T}} \tag{4}$$

where $\mathbf{P}^{\mathrm{T}+}$ is the Moore–Penrose pseudoinverse of $\mathbf{P}^{\mathrm{T}}$.

## ACKNOWLEDGMENT

This work was performed for the Department of Justice in accordance with Section 303 of the Border Security Act, codified as 8 U.S.C. 1732. Specific hardware and software products identified in this paper were used in order to perform the evaluations described in this paper. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology (NIST) nor does it imply that the products and equipment identified are

necessarily the best available for the purpose. The primary goal of the FRGC is to encourage and facilitate the development of face-recognition algorithms. To provide the face-recognition research community with an unbiased assessment of state-of-the-art algorithms, research groups voluntarily submit similarity scores from prototyped experiments to the NIST for analysis. The results of the analysis by NIST are anonymous, unless otherwise agreed to by the participating algorithm developers. All participating groups were given the choice of remaining anonymous or being identified in this paper. Performance results are from January 2005 for all algorithms except Xie *et al.*, 2005, where results are from August 2005.

## REFERENCES


[1] R. Gross, S. Baker, I. Matthews, and T. Kanade, "Face recognition across pose and illumination," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds. New York: Springer-Verlag, 2005, pp. 193–216.

[2] P. J. Phillips, H. Moon, P. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[3] P. J. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. M. Bonein "FRVT 2002 Evaluation Report," Tech. Rep. NISTIR 6965, 2003. [Online]. Available: http://www.frvt.org

[4] W. J. Braje, "Illumination encoding in face recognition: Effect of position shift," *J. Vis.*, vol. 3, no. 2, pp. 161–170, 2003.

[5] W. J. Braje, D. Kersten, M. J. Tarr, and N. F. Troje, "Illumination effects in face recognition," *Psychobiology*, vol. 26, no. 4, pp. 371–380, 1999.

[6] W. J. Braje, G. E. Legge, and D. Kersten, "Invariant recognition of natural objects in the presence of shadows," *Perception*, vol. 29, no. 4, pp. 383–398, 2000.

[7] 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," in *Proc. IEEE Comput. Vis. Pattern Recog.*, 2005, vol. 1, pp. 947–954.

[8] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, and W. Worek, "Preliminary face recognition grand challenge results," in *Proc. 7th Int. Conf. Autom. Face Gesture Recog.*, 2006, pp. 15–24.

[9] A. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Pénard, and H. Abdi, *Face Recognition Algorithms Surpass Humans*, 2005. Tech. Rep. NISTIR. [Online]. Available: http://face.nist.gov

[10] C. Liu, "Capitalize on dimensionality increasing techniques from improving face recognition Grand Challenge performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 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," in *Proc. IEEE Int. Workshop Anal. Modeling Faces Gestures*, 2005, pp. 32–43.

[12] M. Husken, B. Brauckmann, S. Gehlen, and C. von der Malsburg, "Strategies and benefits of fusion of 2D and 3D face recognition," in *Proc. IEEE Comput. Soc. Conf. CVPR*, 2005, vol. 3, p. 174.

[13] P. Grotherin "Face Recognition Vendor Test 2002 Supplemental Report," Tech. Rep. NISTIR 7083, 2004. [Online]. Available: http://www.frvt.org

[14] O. Melnik, Y. Vardi, and C.-H. Zhang, "Mixed group ranks: Preference and confidence in classifier combination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 973–981, Aug. 2004.

[15] J. Czyz, J. Kittler, and L. Vanderdorpe, "Combining face verification experts," in *Proc. 16th Int. Conf. Pattern Recog. II*, 2002, pp. 28–31.

[16] H. Abdi, "Partial least squares regression (PLS-regression)," in *Encyclopedia for Research Methods for the Social Sciences*, M. Lewis Beck, A. Bryman, and T. Futing, Eds. Thousand Oaks, CA: Sage, 2003, pp. 792–795.

[17] T. Naes, T. Isaksson, T. Fearn, and T. Davis, *Multivariate Calibration and Classification*. Chichester, U. K.: NIR Publications, 2004.

[18] H. Martens and M. Martens, *Multivariate Analysis of Quality*. London, U. K.: Wiley, 2001.

[19] A. R. McIntosh and N. Lobaugh, "Partial least squares analysis of neuroimaging data: Applications and advances," *Neuroimage*, vol. 23, pp. 250–263, 2004.

[20] H. Abdi, "Partial least squares regression," in *Encyclopedia of Measurement and Statistics*, N. J. Salkind, Ed. Thousand Oaks, CA: Sage, 2007, pp. 740–744.

[21] H. Abdi, "Multivariate analysis," in *Encyclopedia for Research Methods for the Social Sciences*, M. Lewis Beck, A. Bryman, and T. Futing, Eds. Thousand Oaks, CA: Sage, 2003, pp. 699–702.


**Alice J. O'Toole** received the B.A. degree in psychology from The Catholic University of America, Washington, DC, in 1983 and the M.S. and Ph.D. degrees in experimental psychology from Brown University, Providence, RI, in 1985 and 1988, respectively.

She spent the following year and a half as a Postdoctoral Fellow with the Université de Bourgogne, Dijon, France, which was supported by the French Embassy to the U.S., and with the Ecole Nationale Superieure des Télécommunications, Paris, France. Since 1989, she has been a Professor with the School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson. In 1994, she was awarded a Fellowship from the Alexander von Humboldt Foundation for a sabbatical year at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany. Her research interests include human perception, memory, and cognition, with an emphasis on computational modeling of high-level vision. Current projects include the study of human memory for faces, the comparison of human and algorithm performance on face-recognition tasks, and the computational modeling of data from functional neuroimaging experiments.

**Hervé Abdi** was born in Belfort, France. He received the M.S. degree in psychology from the University of Franche-Comté, Besancon, France, in 1975, the M.S. (D.E.A.) degree in economics from the University of Clermond-Ferrand, Clermond-Ferrand, France, in 1976, the M.S. (D.E.A.) degree in neurology from the University Louis Pasteur, Strasbourg Cedex, France, in 1977, and the Ph.D. degree in mathematical psychology from the University of Aix-en-Provence, Aix-en-Provence, France, in 1980.

He was an Assistant Professor with the University of Franche-Comté in 1979, an Associate Professor with the University of Bourgogne, Dijon, France, in 1983, and a Full Professor with the University of Bourgogne in 1988. He is currently a Full Professor with the School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, and an Adjunct Professor of radiology with the University of Texas Southwestern Medical Center at Dallas. He was twice a Fulbright Scholar. He has also been a Visiting Professor in Brown University, Providence, RI, and with the University of Dijon, Dijon, France, Chuo University, Tokyo, Japan, and the University of Geneva, Geneva, Switzerland. His recent work is concerned with face and person perception, odor perception, and with computational modeling of these processes. He is also developing statistical techniques in analyzing the structure of large data sets (e.g., in brain imaging and sensory evaluation) with partial least square regression, STATIS, DISTATIS, discriminant correspondence analysis, multiple-factor analysis, and additive tree representations.

**Fang Jiang** was born in China. She received the M.S. degree in applied cognition and neuroscience from The University of Texas at Dallas, Richardson, in 2004 and the Ph.D. degree from the School of Behavioral and Brain Sciences.

She has worked on the modeling of MRI data and on the comparison between human and machine face recognition. Her recent work is probing the nature of high-level face representation using adaptation. She has published papers in the areas of human perception, computational comparisons between humans and face-recognition algorithms, and cognitive neuroscience.

**P. Jonathon Phillips** (SM'06) received the Ph.D. degree in operations research from Rutgers University, Piscataway, NJ.

He is a Leading Technologist in the fields of computer vision, biometrics, face recognition, and human identification. From 2000 to 2004, he was assigned to the Defense Advanced Research Projects Agency as Program Manager for the Human Identification at a Distance Program. He was with the U.S. Army Research Laboratory. He is currently with the National Institute of Standards and Technology (NIST), Gaithersburg, MD, where he is the Program Manager for the Face Recognition Grand Challenge and Iris Challenge Evaluation and the Test Director for the Face Recognition Vendor Test (FRVT) 2006. He was the Test Director for the FRVT 2002. His current research interests include computer vision, face recognition, biometrics, and computational psychophysics. His work has been reported in print media of record, including *The New York Times* and the *Economist*.

Dr. Phillips is an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He was the recipient of the Department of Commerce Gold Medal for his work on FRVT 2002.