Image and Vision Computing 32 (2014) 74–85

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Editor's Choice Article
Review Article

# Comparison of human and computer performance across face recognition experiments ☆

P. Jonathon Phillips [a,*], Alice J. O'Toole [b]

[a] National Institute of Standards and Technology, 100 Bureau Drive MS 8490, Gaithersburg, MD 20899, USA
[b] University of Texas at Dallas, School of Behavioral and Brain Sciences, GR4.1, Richardson, TX 75083-0688, USA

ARTICLE INFO

*Article history:*
Received 15 December 2012
Received in revised form 3 September 2013
Accepted 4 December 2013
Available online 14 December 2013

*Keywords:*
Face recognition
Algorithm performance
Human performance
Challenge problem

ABSTRACT

Since 2005, human and computer performance has been systematically compared as part of face recognition competitions, with results being reported for both still and video imagery. The key results from these competitions are reviewed. To analyze performance across studies, the cross-modal performance analysis (CMPA) framework is introduced. The CMPA framework is applied to experiments that were part of face a recognition competition. The analysis shows that for matching frontal faces in still images, algorithms are consistently superior to humans. For video and difficult still face pairs, humans are superior. Finally, based on the CMPA framework and a face performance index, we outline a challenge problem for developing algorithms that are superior to humans for the general face recognition problem.

Published by Elsevier B.V.

**Contents**

1. Introduction . . . 74
2. Review of human and machine comparisons . . . 75
   2.1. Methods . . . 75
   2.2. Still frontal face images . . . 76
      2.2.1. Benchmark on matching studio against ambient illumination images . . . 77
      2.2.2. Benchmark on matching ambient illumination images only . . . 77
   2.3. Video challenge . . . 78
3. Cross experiment comparison . . . 79
   3.1. Analysis . . . 79
   3.2. Conclusions . . . 81
4. Insights from structural comparisons . . . 81
5. Future directions . . . 84
Acknowledgments . . . 84
References . . . 84

## 1. Introduction

Overall, humans are the most accurate face recognition systems. People recognize faces as part of social interactions, at a distance, in still and video imagery, and under a wide variety of poses, expressions, and illuminations. A holy grail in automatic face recognition is developing an algorithm that has performance equivalent to humans—this is equivalent to solving the general face recognition problem. While it is easy to state the problem, accuracy equivalent to humans, it is not obvious how to determine if an algorithm's recognition accuracy is better than a human. One of the key challenges is establishing a measurable goal line and knowing when the goal line is crossed.

Since 2005, human and computer performance has been systematically compared as part of face recognition competitions conducted by

☆ Editor's Choice Articles are invited and handled by a select rotating 12member Editorial Board committee. This paper has been recommended for acceptance by Ioannis A. Kakadiaris.

* Corresponding author.
*E-mail addresses:* jonathon@nist.gov (P.J. Phillips), otoole@utdallas.edu (A.J. O'Toole).

0262-8856/$ – see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.imavis.2013.12.002

**Table 1**
Camera size measured in megapixels and average face-size measured in pixels between centers of the eyes broken out by competition and illumination condition.

| Illumination | Competition | Camera size (megapixels) | Average face size (pixels) |
|---|---|---|---|
| Studio | FRGC | 4 | 261 |
| Ambient | FRGC | 4 | 144 |
| Studio | FRVT—Notre Dame | 6 | 400 |
| Ambient | FRVT—Notre Dame | 6 | 190 |
| Studio | FRVT—Sandia | 4 | 350 |
| Ambient | FRVT—Sandia | 4 | 110 |
| Ambient | GBU | 6 | 175 |

the National Institute of Standards and Technology (NIST) [1–4]. The comparisons provided an assessment of accuracy for both humans and machines for each competition. However, there has not been a systematic analysis of these results across the competitions.

To analyze the results across experiments, we introduce the cross-modal performance analysis (CMPA) framework, which is demonstrated on the NIST competitions. CMPA was adapted from techniques in neuroscience that were developed to compare output from different sensing modalities of brain activity; e.g., functional magnetic resonance imaging (fMRI) and human perceptual judgments [5,6]. These techniques can measure concordance between experimental data and computational models. In our study, the modalities compared are human and algorithm performance. In the psychology and neuroscience literature, face recognition algorithms can be referred to as computational models. The computational model can be designed to optimize performance or to model the human face recognition processes. The framework is sufficiently general that it provides a goal line for determining when machine performance reaches human levels.

On frontal faces in high quality still images, our analysis shows that machine performance is superior to humans. For these images, machines represent a person's identity primarily by encoding information extracted from the face; information from the body, hair, and head is generally ignored. For video and extremely difficult-to-recognize face pairs, experiments show that humans take advantage of all available identity cues when recognizing people [7,8]. CMPA quantifies the potential for improving machine performance if all possible identity information is encoded by algorithms.

Comparing machine and human performance started with independent experiments in NIST competitions. The synthesis of the results across experiments gives a greater understanding of the relative strengths of machines and humans. The CMPA framework provides a goal line for determining if algorithm and human performance is comparable on the general face recognition problem.

## 2. Review of human and machine comparisons

We examine the relative performance of humans and machines for both still and video imagery. This review section presents the key details and conclusion for each study. The key details and conclusions were selected to lay the groundwork for the cross-experiment analysis in Section 3. The summary includes an overview of the images in the experiment, how the images were selected for measuring human performance, the key receiver operating characteristics (ROCs) comparing machine and humans, and the headline conclusions for each experiment. References are provided for full details on each experiment and the associated competitions.

### 2.1. Methods

To encourage the development of face recognition technology and to provide an independent assessment of algorithm performance, since 1993 the U.S. Government has sponsored a series of competitions [9]. The competitions came in two varieties: challenge problems and evaluations. A challenge problem can be considered a homework assignment meant to assist developers in improving algorithm performance. 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 goal of challenge problems is to encourage and facilitate the development of new face recognition technology. In a challenge problem, participants were provided with a large set of face images, a protocol for performing a set of experiments, and the code for scoring algorithm performance. The experiments were designed so that multiple algorithms could be compared on exactly the same images. In addition, participants were given the answers, also known as ground truth. By providing the answers, it allowed participants in a challenge problem to improve their algorithms or develop new algorithms. Participants could submit their raw results from matching images in an experiment to NIST for analysis. From the raw results, NIST performed analysis across participants on a common set of images. The analyses were included in summaries, presentations, and papers on a challenge problem.

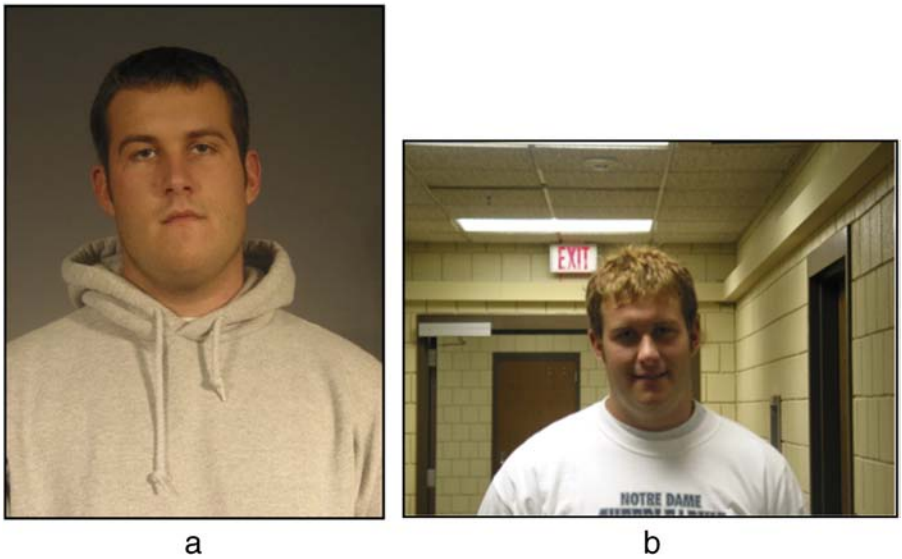


**Fig. 1.** Example of a pair of images used in experiments comparing identities in images captured in a studio environment (a) and an ambient environment (b).

The goal of an evaluation is to provide an independent assessment of algorithm performance. For evaluations, participants submitted their algorithms to NIST, and the algorithms were tested in the NIST Biometric Testing Laboratory. Performance of algorithms was measured on sequestered data. In the statistical learning domains, this is the standard methodology to measure the ability of an algorithm to generalize to novel data and mitigate the effects of over-tuning to the development data.

Although the NIST competitions measured algorithm performance on a number of recognition tasks, the analysis in this paper is limited to a verification task. In the perception community, this is known as face-identity matching of unfamiliar faces. In our verification task, humans and machines were given a pair of images or videos, with each image or video containing one face. The humans and machines had to respond how likely the two faces were of the same person. For machines, the response is a number called a similarity score. Each algorithm has its own similarity score distribution. From the similarity scores, receiver operating characteristics (ROCs) can be computed.

Human performance was measured on normal (i.e., untrained) people who had no professional experience with face recognition. Performance was measured by presenting two face images or videos on a computer screen. They were asked to judge the similarity between a face pair on the following scale:

- 1.) You are sure they are the same person;
- 2.) You think they are the same person;
- 3.) You don't know;
- 4.) You think they are different people;
- 5.) You are sure they are different people.

From the human generated ratings, ROCs were computed. For consistency, in our analysis of still image experiments, all face pairs were presented on the computer screen for 2 s. The presentation time of 2 s was chosen based on experiments in O'Toole et al. [1] showing that human accuracy was stable between 2 s and unlimited time. However, subsequent experiments showed that a slight improvement in performance is possible when a subject has unlimited time to make a decision [8]. The number of subjects judging the similarity between face pairs varied by experiment. Because this is a review article, we provide citations to the original papers that have the full experimental details including the number of subjects.

Because one of the main goals of an evaluation was to test the ability of algorithms to generalize to novel faces, faces in evaluations were sequestered; e.g., images of the faces in an evaluation were not released to the face recognition community. Thus, the faces in evaluations were "unfamiliar" to the algorithms. This kind of sequestering is likely to be comparable to the humans tested, who have general experience with faces, but no experience with the faces used as test stimuli in the experiments. Moreover, the unfamiliar face matching task is comparable for machines and humans operating in situations typical of security applications, where face recognition for previously unfamiliar people is required.

The main difference between measuring performance of humans and machines is the number of face pairs that can be compared. In the NIST competitions, machines compare millions of face pairs. Because it was impossible for human subjects to rate millions of face pairs, the human–machine comparison focused on a subset of the face pairs compared in machine experiments. The maximum number of face pairs that a subject can rate in an experiment is about 250. One of the factors differentiating the experiments is the method for selecting the face pairs in an experiment.

### 2.2. Still frontal face images

Over the last twenty years the most active research area in automatic face recognition has been developing algorithms to recognize faces from frontal still images. In the last ten years, one emphasis of the NIST competitions has been recognition from frontal face images acquired with a digital single lens reflex camera. The majority of these images are considered high quality to humans. The images were collected under two illumination conditions. One was in a studio environment with controlled lighting. The other was under ambient lighting indoors and outdoors.

Progress on recognizing images under these conditions has been measured through a series of US Government sponsored competitions [9]. Three recent competitions included comparing of human and machine performance: the Face Recognition Grand Challenge (FRGC) [10], the Face Recognition Vendor Test (FRVT) 2006 [3], and the Good, Bad, & Ugly face challenge problem (GBU) [11]. The FRGC study reports algorithm results from 2005. The GBU algorithm challenge has been ongoing since 2011; however, the best reported results are from the FRVT 2006.

Human and machines were compared for two categories of experiments. In the first category, one image in a face pair was taken in studio lighting and the other was taken in ambient lighting. In the second category, both images were images taken under ambient illumination.

The demographics and size of the face varied by experiment. The face size is measured by the number of pixels between the centers of

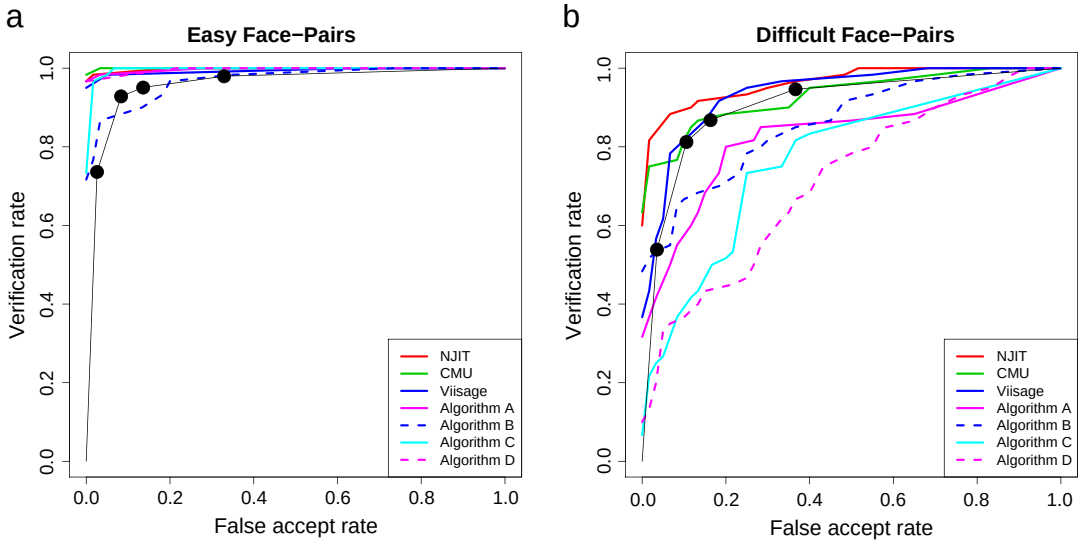


**Fig. 2.** Human and machine performance on the FRGC data set. (a) Performance on the easy face pairs and (b) performance on the difficult face pairs.

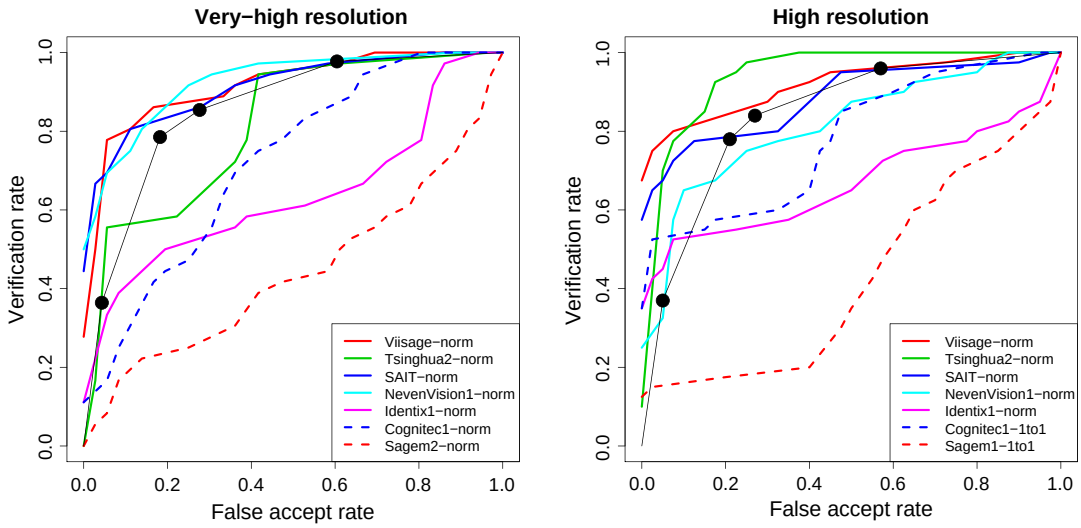


**Fig. 3.** Human and machine performance on the Notre Dame and Sandia data sets.

the eyes. Table 1 specifies face size by experiment, data set, and imaging condition. The experiment is labeled by the associated competition.

The data for the still face experiments came from two sources: University of Notre Dame and Sandia National Laboratory. For the data collected at Notre Dame, the overall demographic composition was 59% male and 41% female; 71% Caucasian and 10% East Asian; and 92% were 18 to 29 years old. The demographics varied slightly by experiment with precise numbers provided in the references. For data collected at Sandia the demographic composition was 55% female and 45% male; 64% Caucasian and 21% Hispanic; and 35% were 50 to 59 years old, the age range for the remaining 65% was roughly spread evenly over the 18 to 59 and 60 to 65 age ranges.

*2.2.1. Benchmark on matching studio against ambient illumination images*

In the first two studies comparing human and machine performance, one image was acquired in a studio environment and the second was captured in ambient indoor lighting, see Fig. 1.

In the first study we review, results were reported for two categories of face pairs: easy and difficult. The two categories allowed for the comparison of human and machine results at both ends of the performance range, see O'Toole et al. [1] for details. This baseline is an implementation of an algorithm based on principal components analysis (PCA) [12–14]. This algorithm achieved good recognition performance for the easy category and poor recognition performance for the difficult face pairs. The seven algorithms in the experiment were participants in the FRGC. Human and algorithm performance is reported for the same set of image pairs. The human and machine ROC curves for the difficult face pairs (Fig. 2(b)) show that three algorithms were more accurate than humans [15–17] and four algorithms were less accurate. For the easy face pairs (Fig. 2(a)), the algorithms and machines were highly accurate, with all but one algorithm performing more accurately than humans.

The previous experiment examined performance for easy and difficult face pairs; the second experiment measures performance on face pairs of average difficulty. In this study, humans were compared to algorithms submitted to the FRVT 2006 [2,3]. Average difficulty is defined relative to algorithms in the FRVT 2006 competition. A face pair had average difficulty if approximately half of the algorithms performed correctly (i.e., in a face pair with images of the same person, then approximately half of the algorithms reported that the images were of the same person). Experiments were performed on images from two data sets, one collected at the U. of Notre Dame and the second at the Sandia National Laboratory.

ROCs for both data sets are presented in Fig. 3. There are two key conclusions. First, the results on both data sets are consistent with the difficult portion of the FRGC (previous experiment) comparison; machine performance is in the range of human performance, with the best algorithms surpassing humans. Second, human performance is stable across the two data sets.

*2.2.2. Benchmark on matching ambient illumination images only*

The next experiment relaxed the photometric constraints. Both faces in a pair were acquired in ambient lighting conditions. The images were taken outdoors or indoors in atriums and hallways. To better understand the range of performance under general illumination conditions, three partitions were created based on difficulty of matching.[1] To arrive at the performance-based partitions, three top-performing face recognition algorithms from the FRVT 2006 test were fused to produce a single algorithm. Based on performance of the fusion algorithm, images were divided into three partitions with high (the Good), challenging (the Bad), and very challenging (the Ugly) accuracy, hence the name Good, Bad, and Ugly (GBU) Face Challenge Problem. On the Good partition, the base verification rate (VR) was set to 0.98 at a false accept rate (FAR) of 0.001. For the challenging partition, the VR was set to 0.80 at a FAR of 0.001, and on the very challenging partition the VR was set to 0.15 at a FAR of 0.001.

In the GBU, the effects of natural variations in a person's day-to-day appearance (hair, facial expression, etc.) and variations in illumination across both indoor and outdoor settings were considered. All of these images were nominally frontal. Because all images were collected between August 2004 and May 2005, aging cannot be a factor. There is the same number of images of each person in all three partitions. Thus, only the images, not the individual identities, changed across the three partitions. This provides an assurance that the accuracy differences were due to factors other than the particular set of face identities tested. Human performance on the GBU is reported in O'Toole et al. [4]. Fig. 4 shows three face pairs of the same person, sampled from the good (left column), challenging (middle column), and very challenging (right column) performance conditions. This figure illustrates the wide variation in the appearance of a person across frontal images. It also highlights the difficulties that may occur in matching identity in faces that are taken in different settings and which include variations in expression and appearance-based features such as hairstyle. These factors become even more salient in combination (cf., Fig. 4 right column).

[1] An overview of the creation of the GBU partitions is presented in this Section, details are given in Phillips et al. [11].


**Fig. 4.** Examples of face pairs of the same person from each of the GBU partitions: (a) good, (b) challenging, and (c) very challenging.

ROCs comparing human and machine performance are presented in Fig. 5. For humans, performance on the Good partition is superior to the challenging and very challenging partitions, see Fig. 5(a). The difference in human performance on the challenging and very challenging partitions is not statistically significant (cf. O'Toole et al. [4]). For all three partitions, performance on the fusion algorithm is superior to humans, see Fig. 5(b, c, d).

To gain better understanding of the relative strengths of human performance, Rice et al. [8] examined human performance when algorithms completely fail. From the very-challenging partition in the GBU, 50 same-identity face pairs and 50 different-identity face pairs were selected so that the similarity score for all same-identity pairs was lower than all different-identity pairs. A higher similarity score implies a greater likelihood that the face pairs consist of two images of the same face. Thus, performance of the FRVT 2006 fusion algorithm was 100% incorrect. Thus, we refer to these as extremely-difficult face pairs.

To understand the reason for algorithm failure, Rice et al. [8] measured the contribution of face and body, face only, and body only to recognition by humans. To measure the contribution of these three conditions, three versions of the face images were created, see Fig. 6. In the first experiment, human observers were presented with the original images, see Fig. 6(a). In the second experiment, humans were presented with images where the face was masked, see Fig. 6(b). In the third experiment, the images consisted of only the face, see Fig. 6(c). The ROCs for all three human viewing conditions and the fusion algorithm are shown in Fig. 7.

Performance between the body only and original images was indistinguishable. Performance on the face only images was remarkably inaccurate, but greater than chance. The results indicate that the body, rather than the face, accounts for human accuracy at identifying people in the original unedited images.

### 2.3. Video challenge

In our daily lives, faces are recognized as we interact with people. This allows the incorporation of motion and non-face identity cues into the recognition process. The equivalent model for algorithms is recognition from video.

O'Toole et al. [7] extensively studied human performance on video sequences. The data set in the study consisted of two categories of video sequence [18]. The videos were captured in standard-definition progressive-scan format by a digital video camera. In the first, a person walked towards the camera; in the second, a person was engaged in a conversation, see Fig. 8. O'Toole et al. [7] measured effect of face, body, and motion on performance. The analysis in this paper is restricted to two key cases. The first was recognition from the entire video sequence. The second was recognition when only the head and face were visible in the video sequence; the background and the person's body were masked, see Fig. 8(c). The original video sequence case reports performance when information about the head, face, body, and motion is available. The videos with only the head and face were designed to measure the performance when information about the body was not present.

The video sequences in the above study were included in the Video Challenge of the Face and Ocular Challenge Series (FOCS).[2,3] The current paradigm in automated video recognition is to first detect frontal faces and then feed the frontal faces into a recognition algorithm. In video algorithms, a key challenge is recognizing people in sequences that do not contain frontal faces. In the Video Challenge, this challenge is represented by the conversation sequences. The video dictionary algorithm of Chen et al. [20] reports performance on the conversation video sequences in Video Challenge. The video dictionary algorithm extracts the face from each frame. The extracted faces are then grouped by pose. From each group a dictionary is learned. Non-frontal faces are recognized by comparing similar pose groups. Since features are only extracted from the face, the video dictionary algorithm does not incorporate body information in the recognition process.

[2] Information on obtaining the FOCS can be found at http://face.nist.gov.

[3] The video challenge was originally included in the Multiple Biometrics Grand Challenge (MBGC) [19].

Human and machine performance is shown in Fig. 9. Results are reported for comparing walking-vs-walking videos, conversation-vs-conversation videos, and conversation-vs-walking videos. For humans, performance is reported for both original and face only sequences. For humans, comparing walking-vs-walking videos is superior to the other two cases, Fig. 9(a). On the original video sequences, humans are superior to the algorithm in all three cases, Fig. 9(b, c, d). On the face only sequences, the algorithm is roughly equivalent to humans on the walk-vs-walking and conversation-vs-conversation cases, Fig. 9(b, d). Since the algorithm only encodes identity from the face, comparing human performance on the face only sequences is meaningful. In the cross pose case, conversation-vs-walking, humans are better, Fig. 9(c).

## 3. Cross experiment comparison

The next step is to take the experiments reviewed in the previous section and to analyze them as a group. The analysis is performed by using the cross-modal performance analysis (CMPA) framework, which we introduce.

### 3.1. Analysis

Traditionally, the performance of humans and machines is compared by plotting their ROCs on a single plot. ROCs are a standard method for reporting performance on a small number of experiments. However, they do not allow for a concise summary across a large number of experiments. Comparing performance across experiments is accomplished by placing ROCs side-by-side; e.g., Figs. 2, 3, 5, and 9. Ideally, to compare results across experiments, each ROC needs to be summarized by a single number. For our analysis, we summarize a ROC by the area under the curve (AUC) [21,22]. The range of values for AUC is [0,1]. An AUC value of 1 is perfect performance and a value of 0.5 is random performance. An AUC value of 0 corresponds to no correct answers; e.g., algorithm performance on the extremely hard face pairs.

In the CMPA framework, the relative performance of humans and algorithms is characterized by their AUC statistics on the same experiment. In the experiments reviewed in this paper, the AUCs are computed from responses to stimuli from the same data set. This technique is extensible to analysis were the two underlying ROCs do not

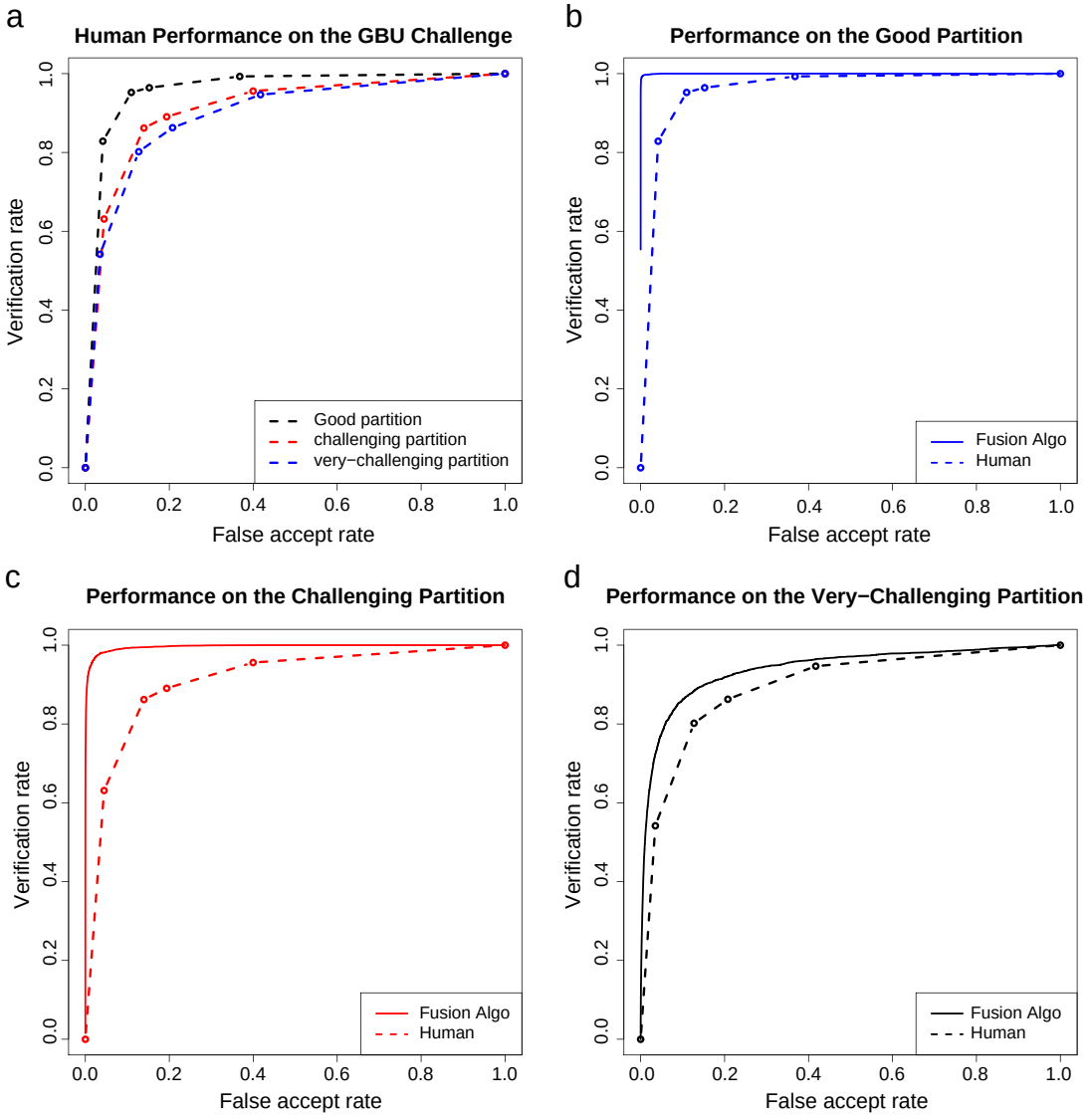


**Fig. 5.** Human and machine performance on the GBU partitions. (a) Human performance on all three partitions. Comparison of human and machine performance by partition: (b) Good, (c) challenging, and (d) very challenging.

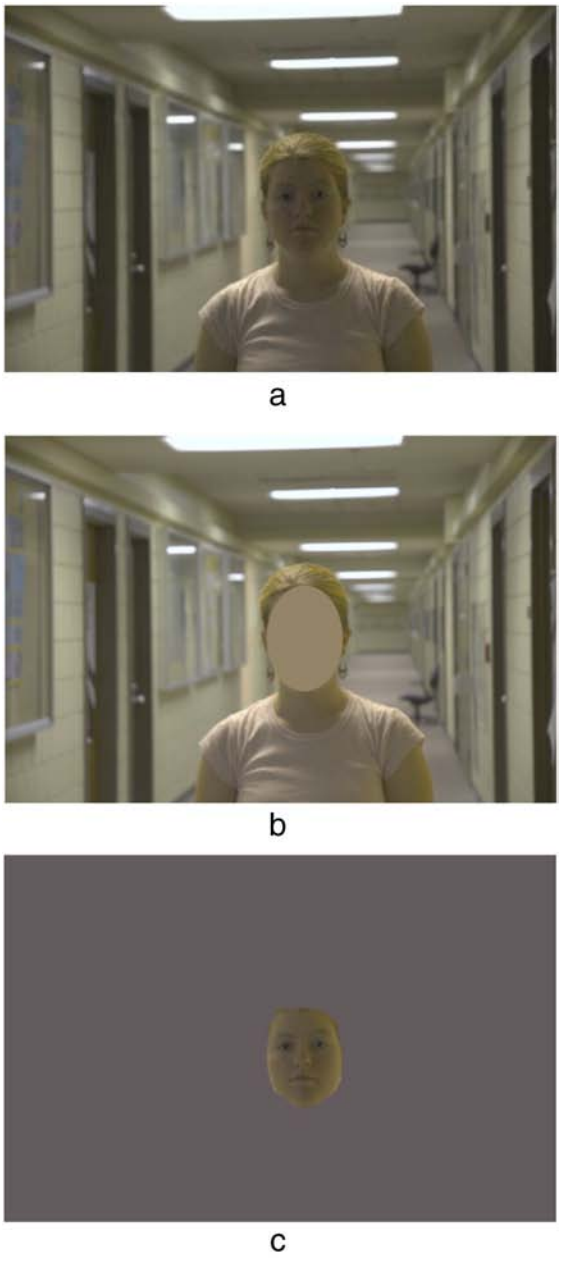


**Fig. 6.** Example of one set of images from the extremely hard face pair study. (a) Original image. (b) Image with the face masked. (c) Image with only the face visible, where the background, body, and hair have been masked.

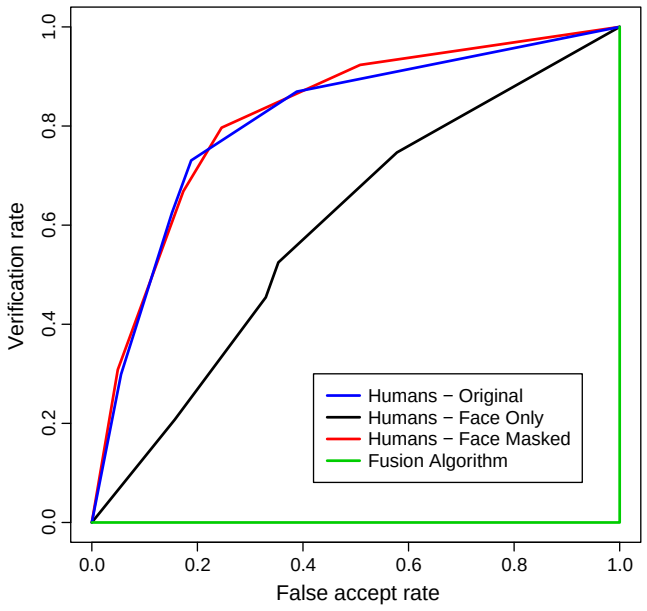


**Fig. 7.** Performance on the extremely hard face pairs. Because algorithm performance is 100% incorrect, its ROC, green line, hugs the bottom and right axes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

need to be responses to stimuli from the same data set. We graphically show this relationship on a scatter plot, see Fig. 10. The *x*-axis is AUC for human performance and the *y*-axis is AUC for algorithm performance. If the AUC for both human and machine performance are approximately equal, then performance for humans and machines is comparable. In Fig. 10 this is the diagonal line. Points in the region above the diagonal line correspond to experiments where machines perform better than humans (as measured by AUCs). Likewise, points in the region below the diagonal line correspond to experiments where humans are better. For algorithm developers, in the ideal case, all points would be on horizontal line with machine AUC equal to 1.0. For those modeling the human face recognition system, in the ideal case, all points will lie along the diagonal line.

Our CMPA analysis builds on the work on DiCarlo [5], where human and machine performance on object recognition is characterized by the statistic $d'$ [21,22]. A related technique is representational similarity analysis for comparing different brain imaging modalities with human perceptual similarity judgments [6].

To assess the relative capabilities of humans and machines, we apply our CMPA analysis to all the experiments reviewed in Section 2. Fig. 11 shows the results of this analysis. In Fig. 11, experiments are grouped into three categories.

The first category consists of the still image experiments with the exception of the difficult-face pairs. For the first two experiments reviewed, see Section 2.2, performance is reported for multiple algorithms. For the analysis in this section, machine performance is represented by one of the top performers. This is because of multiple experiments in each competition, there is not a clear top performer. For the two FRGC experiments, machine performance was reported for the New Jersey Institute of Technology (NJIT) algorithm [15]. Likewise, for the FRVT 2006 results, performance is reported for the Viisage-norm submission to the FRVT 2006. For the GBU experiments results are for the FRVT 2006 fusion algorithm in Phillips et al. [11]. For these seven experiments, a regression line has been plotted in Fig. 11.

In this category, the images were acquired with digital single lens reflex cameras and the majority of the images are considered "high quality" by the face recognition community. The face-image pairs are selected by different criteria and cover a range of imaging conditions. For these experiments, performance of machines is superior to humans and the regression line suggests that there is a linear relationship between the difficulty of the experiments for humans and algorithms.

On the video challenge, performance is compared on six experiments. The experiments are organized into three experimental conditions that characterize the pairs of video that are compared. For each condition, there are two viewing conditions: the original sequences and face only video sequences. The machine results on the video experiments is for a video-dictionary based algorithm [20].

For the extremely difficult face pairs, AUCs for three experiments, face only, face masked, and original image, are presented. For all three experiments, machine performance is the FRVT 2006 fusion algorithm. By the design of the selection process, the AUC for the fusion algorithm is 0.

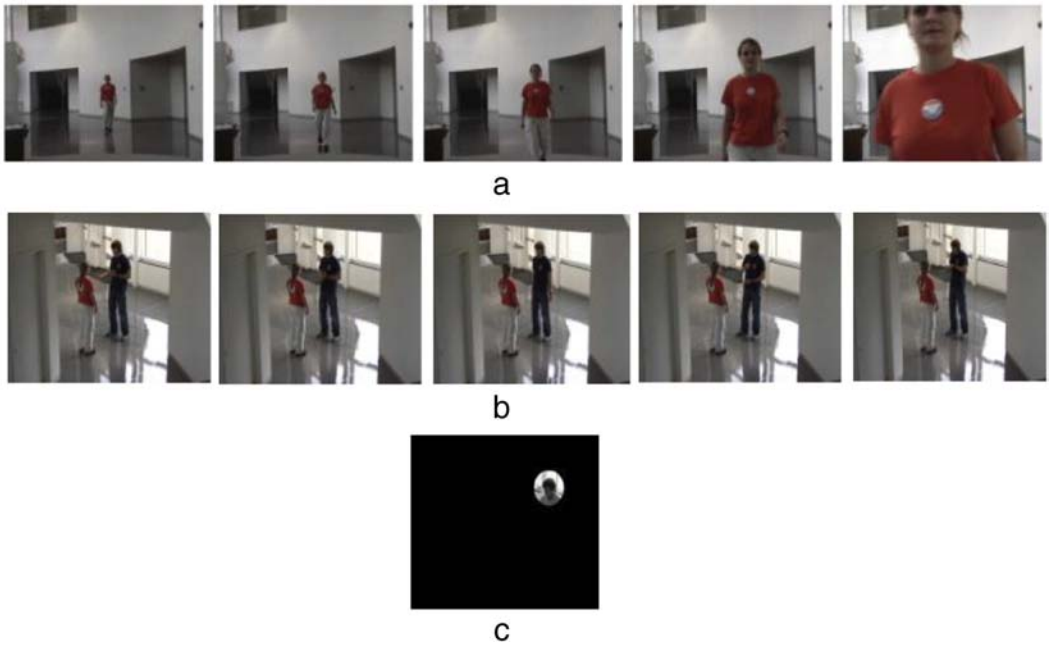


**Fig. 8.** Example imagery from the video experiments. (a) Frames from a sequence of a person walking towards the camera. (b) Frames from a sequence of two people talking. The goal is to recognize the person facing the camera. (c) One frame from a floating head sequence that only contains the head and face of a person.

### *3.2. Conclusions*

The most studied area of face recognition is recognition from high quality still frontal face images. In our studies these are represented by images taken with a digital single lens reflex camera. For the seven experiments described in Section 2.2, our analysis shows that machines are superior to normal humans. For these experiments, the faces contain significant identity cues.

The results on the video and extremely-difficult face pairs show conditions where human performance is superior to machines. In the video experiment, human and machines are at near parity when pose is the same in both video sequences and only face identity cues are considered. When there was a change in pose, human performance was superior to machine performance. On both the video and extremely-difficult face pair experiments, human performance is superior when non-face identity cues are dominant. These results suggest that humans effectively integrate non-face identity cues into the recognition process and that humans take advantage of the head and body in identifying someone. Because the automatic face recognition community has developed algorithms that compensate for changes in pose, future experiments should directly compare human and machine performance on changes in pose.

The CMPA provides a high level summary across multiple experiments. One direction for future analysis is developing statistical models of both human and machine performance. One example is generalized linear models that allow for analysis that explicitly models the effect of covariates [23,24]. This class of analysis has been restricted to algorithm performance on a single data set. To extend this technique to the problem described in this paper, the models need to be able to analyze results on multiple data sets and incorporate human matching results.

## 4. Insights from structural comparisons

The analysis in Section 3 directly compared human and machine performance. There is more to learn from the interplay of machines and humans than what can be learned from relative performance comparisons. We will examine this interplay in the context of three topics. The first is the other-race effect, where algorithms have contributed to understanding the human face processing systems and human face processing has contributed to understanding machine performance. Second, it has been possible to improve techniques for the analysis of machine performance based on the design of human experiments. Finally, the effect of fusing machine and humans is reviewed and can reveal strategy differences in the way humans and machines perform the tasks.

The *other-race effect* is a classic property of human face recognition. Our ability to recognize the identity of faces from our own race is better than our ability to recognize the identity of faces of other-races. The other-race effect for face recognition has been established in numerous human memory studies [25] and in meta-analyses of these studies [26–28].

Phillips et al. [29] looked for an other-race effect in algorithms submitted to the FRVT 2006. They compared the performance of a fusion of East Asian algorithms and a fusion of Western algorithms matching identity in pairs of Caucasian and East Asian faces. The East Asian algorithm was a fusion of five algorithms submitted from East Asian countries, and the Western algorithm was a fusion of eight algorithms from Western countries.

The study showed an other-race effect for the algorithms. Specifically, performance of face recognition algorithms varies as a function of the demographic origin of the algorithm and the demographic contents of the test population. The mechanisms underlying the other-race effect for humans are reasonably well understood and are based in early experience with faces of different races. Because the algorithms tested were black boxes, conclusions about the mechanisms underlying the algorithm effects are tentative, but the effects reported were not. The results point to an important performance variable combination that has not received much attention. The results also suggest a need to understand how the ethnic composition of a training set impacts the robustness of algorithm performance. From a practical perspective, algorithms need to be evaluated on face sets whose demographics match those at the location(s) where they will be used.

Furl et al. [30] used computational models to investigate two competing hypotheses to account for the other-race effect in humans. First, the generic contact hypothesis links the magnitude of the other-race effect in individuals to the relative amount of contact they have with own

versus other race faces. Thus, people who see many individuals of another race on a daily basis should have a smaller other-race effect than those who rarely see other-race individuals. This effect is modeled using a principal components analysis (PAC)-based face recognition algorithm, where the proportion of faces in the training set from two races was varied [31,14,13]. Second, the developmental contact hypothesis assumes also that experience is the cause of the other-race effect, but that experience early in life (up to about 5 years of age) with other-race faces is the critical factor. The rationale for this hypothesis is that the neural system early in life tunes itself to the statistical structure of the environment as it selects a feature set that will optimally represent the stimuli encountered most frequently. Consistent with the predictions of a *developmental* contact hypothesis, experience-based models demonstrated an other-race effect *only* when the representational system was developed through experience that warped the perceptual space in a way that was sensitive to the overall structure of the model's experience with faces of different races. These models were based on combinations of PCA and Fisher discriminant analysis (FDA) applied as the system acquired features for representing faces [32–34]. The results from this study supported a developmental contact hypothesis for the formation of the other-race effect in humans.

Traditionally, attempts to improve algorithm performance have emphasized methods that increase the degree of similarity between two images of the same person; i.e., by modeling the effects of changes in illumination between two images of the same face. Less consideration has been given to the effects of the composition of the different-identity distributions in producing stable estimates of algorithm performance. In human experiments, the faces in different-identity pairs always have the same sex, race, and approximate age. The reason for this condition is that humans rarely confuse faces of different sexes, races, or age groups. However, in the majority of face recognition algorithm competitions, different-identity pairs are cross demographic. In fact, in many cases, the majority of different-identity pairs are cross-demographic. Inspired by the design of human experiments, O'Toole et al. [35,36] looked at the effect of algorithm performance as cross-demographic face pairs were limited. Experiments were performed on the GBU face challenge. Performance was measured in four cases. First, performance was measured when there was no pruning of the

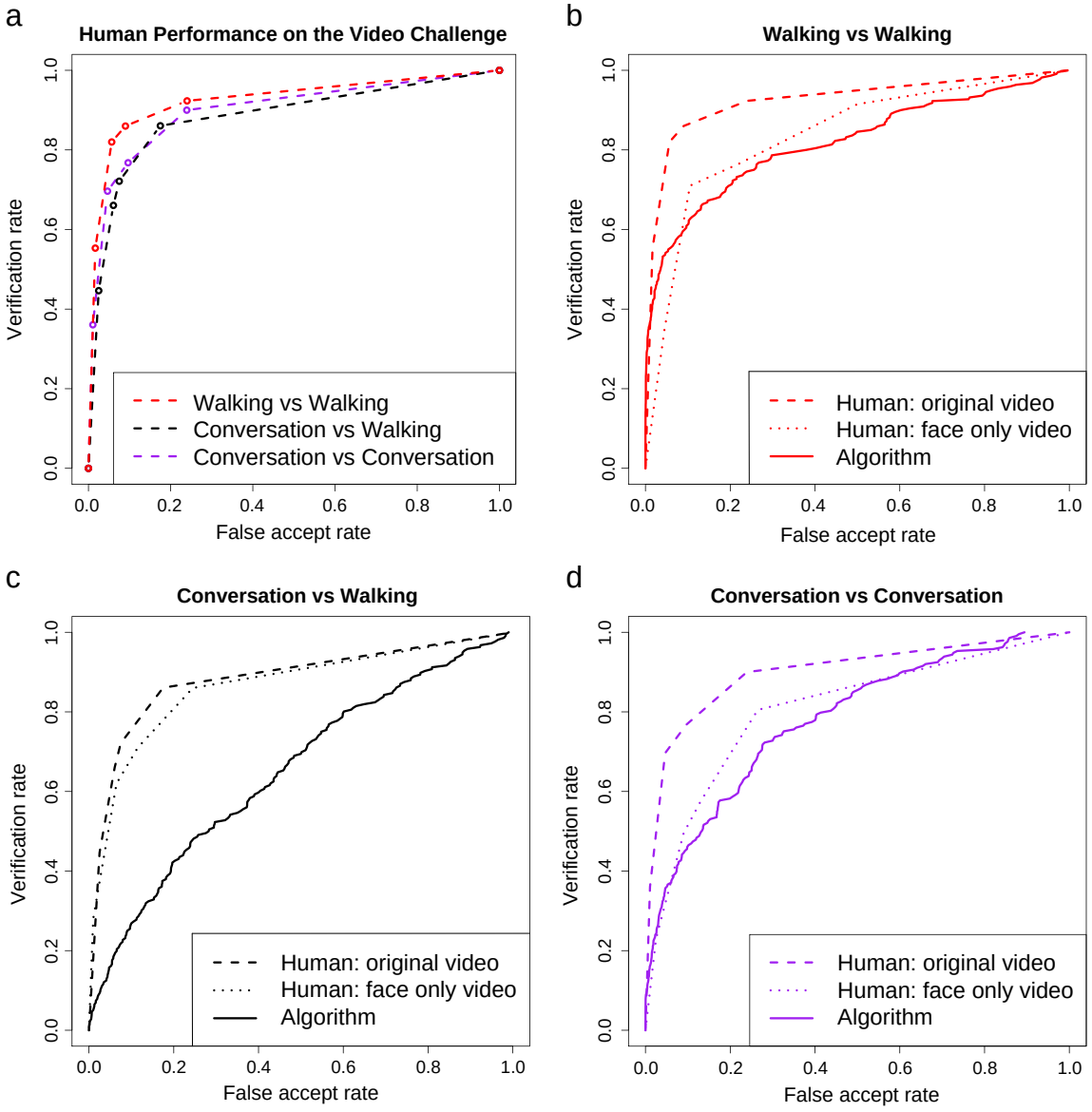


**Fig. 9.** Human and machine performance on the video challenge. (a) Human performance on all three conditions on the original video sequences. Comparison of human and machine performance by condition: (b) walking vs walking, (c) conversation vs walking, and (d) conversation vs conversation. Human performance is reported for both the original and face only video sequences.

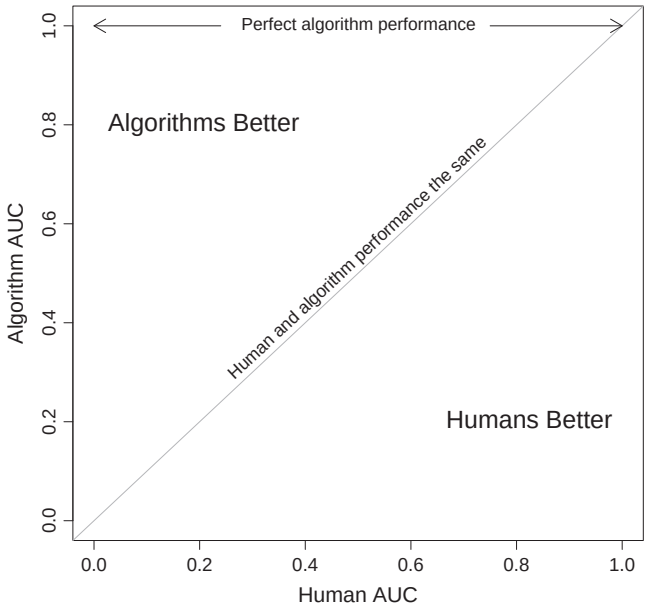


**Fig. 10.** Properties of CMPA scatterplot for comparing human and machine performance.

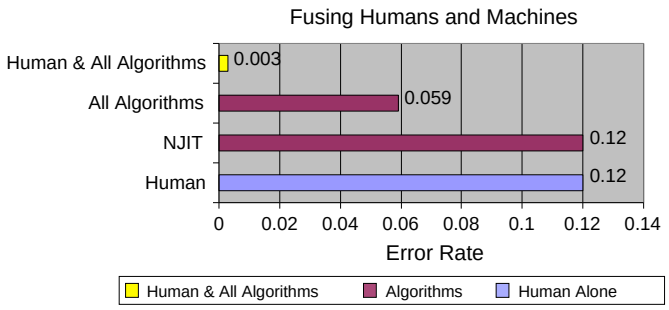


**Fig. 12.** Performance on fusion experiments. Total error rate is reported for humans, NJIT, all algorithms fused, and all algorithms and humans fused.

different-identity face pairs. This is the control case. Second, the faces in all different-identity pairs were of the same sex. Third, the faces in all different-identity pairs were of the same race. Fourth, the faces in all different-identity pairs were of both the same race and sex. On the Bad partition, at a false accept rate of 1 in 1000, the verification rate was 0.79 for the control case, 0.74 and 0.74 for demographic matching on sex or race, and 0.69 for different-identity pairs with the same race and sex. In these experiments, performance was measured for the fusion algorithm [11]. A similar reduction in performance, measured as verification rate, was observed for face pairs in the Ugly partition.

The simulations over the four cases showed that differences in the demographic composition of the different-identity distribution can significantly alter the estimates of algorithm performance. These estimates are important for predicting how algorithms will perform in real-world environments. Furthermore, these results pose a new and pressing challenge for the biometric community to find a method for tuning algorithm performance to the constantly changing demographic environments in which systems must operate reliably.

We end this review with two related questions. "Is an algorithm a reasonable model for human face recognition?" and "Does fusing human and machines improve performance?" Fusion is an effective tool for improving performance when the models encode complementary features. In other words, there is qualitative diversity in the way the models (human and machine) encode and recognize faces. If an algorithm is a good model for human face recognition, then there will be similar approaches to recognizing people. Thus, fusing them will not significantly improve performance. O'Toole et al. [37] looked at fusing human and machine results. Experiments were performed on the difficulty set of images in the FRGC experiments, see Section 2.2.1 and Fig. 2(b). The fusion algorithm was based on least partial squares regression [38,39]. Fig. 12 shows the key results from the study and the total error rate that was reported. The first two cases are controls, performance on humans and the FRGC submission from NJIT algorithm. The third case is fusing all seven algorithms from the FRGC. The best results were achieved by fusing humans and all seven algorithms in the study, with an error rate of 0.003.

That study demonstrated that fusing algorithms and humans can substantially reduce the error rate. For the data set and algorithms in the study, the fused error rate was almost zero. Because of the large reduction in the error rate, the results support two significant conclusions. First, designing systems to effectively fuse humans and machines can significantly improve overall performance. Second, the mechanisms underlying the recognition process in machines and humans are qualitatively different.

The three structural studies reviewed in this section demonstrate the potential for a broad interaction between human and machine studies. This has led to improvement in experiment design, deeper understanding of the principles of face processing, and the ability to effectively combine humans and machines.

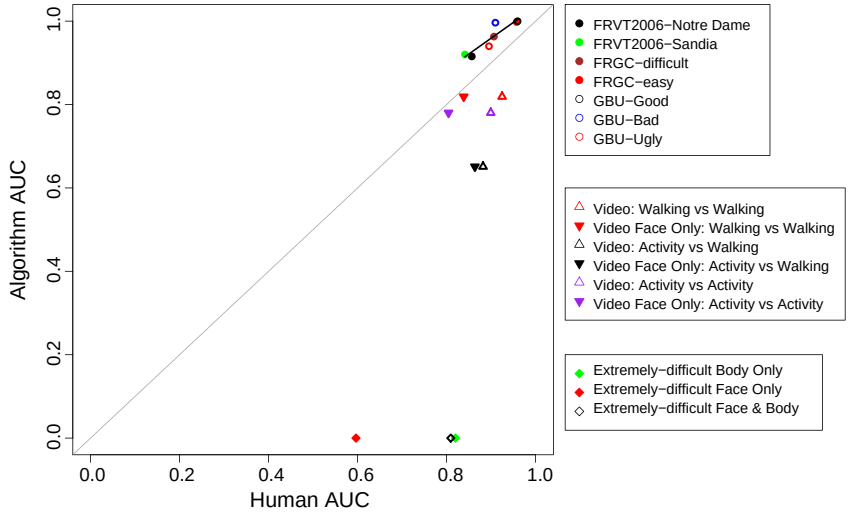


**Fig. 11.** CMPA analysis for experiments reviewed in Section 2.

## 5. Future directions

The cross-modal performance analysis framework was designed to compare human and machine performance across a series of experiments. Although this framework is useful in its own right, we apply this technique to establish goals for advancing face recognition technology.

Over the last two decades, phenomenal progress has been made in automated face recognition from frontal images taken in mobile studio or mugshot environments. Results from the MBE 2010 report a false reject rates of 27 in 10,000 at a false accept rate of 1 in a 1000; and an identification rate of 0.93 from a gallery of 1.6 million faces [40]. Between 1993 and 2010, the false reject rate at a false accept rate of 1 in a 1000 has decreased by a factor of two every two years [9].

Clearly, this level of performance cannot be achieved for faces acquired under all conditions. The question then becomes: What is a reasonable performance bound or goal? We propose a goal based on human performance. Establishing a goal based on one set of images will not adequately characterize a problem. To provide the needed benchmarking data, performance should be characterized by a set of experiments, where each experiment focuses on a different aspect of the challenge. For example, the analysis in Section 3 is conducted on a set of 16 experiments. We formalize this concept as a *Face Performance Index*. The goal in designing a face performance index is to select a set of experiments that adequately characterize performance under the range of conditions that are relevant to a problem.

How does the CMPA framework and a face performance index establish a performance benchmark relative to humans? The first step is to create a face performance index that consists of a set of experiments that adequately characterizes human performance. Initially, this face performance index could be a first order approximation. Later iterations of this index would provide better approximations.

For each experiment, human and machine performance is measured, the appropriate performance statistic is computed, and plotted on a CMPA scatterplot. In our analysis the performance statistic was the AUC. The goal of combining CMPA and a face performance index is to spur progress. This is illustrated in Fig. 13, which is adapted from Fig. 11. The green region in the upper left region of the figure is the 'goal box.' The goal box is the region of the plot where algorithms perform better than humans and algorithm performance is better than random. When all of the experiments are in the goal box, then within the CMPA framework, the performance of an algorithm on a face performance index is better than humans. The fraction of experiments that are in the goal box is a measure of the success of an algorithm.

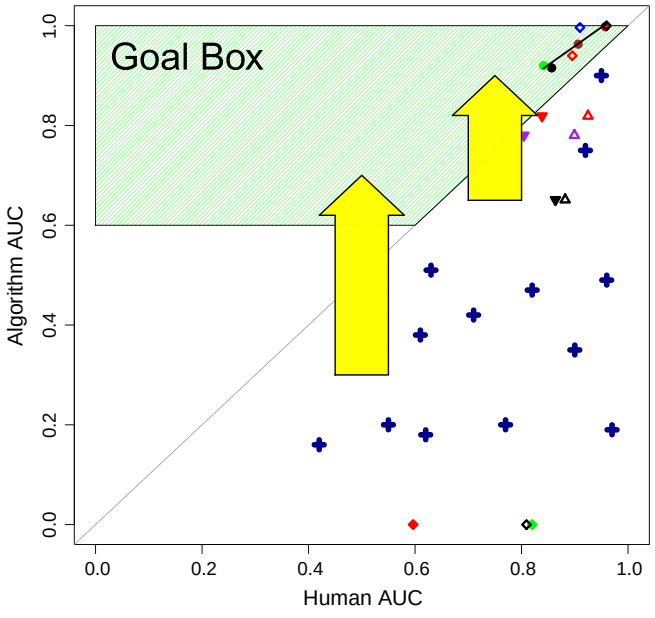


**Fig. 13.** The CMPA framework and a face performance index for setting goals for machine performance.

To illustrate the process, we create a notational challenge problem. Here a face performance index is created by augmenting the 16 experiments from Section 3 with additional notational experiments, which are annotated with blue + s. The position of the + s represents performance at the start of a challenge problem, which assumes that performance of humans is superior. The goal of the challenge is to improve algorithm performance so the experiments represented by the + s are in the goal box, which is illustrated by the yellow arrows.

In the analysis in this paper, performance is measured for average-recognizers who have not received training. One common assumption is that trained law enforcement officers and forensic examiners are better face recognizers. Russell et al. [41] reported the existence of four super-recognizers. However, there are few published papers that report the performance of law enforcement officers or forensic examiners and their results are mixed [42–44].

The goal could be updated to have machines match the abilities of super-recognizers, trained law enforcement professionals, or forensic face examiners. The goal of the notational challenge is absolute, better than humans. By modifying the goal-box region, the goal can be relative, with machine performance better than humans by a fixed factor. These two modifications illustrate the flexibility of the CMPA and face performance index for formulating challenge problems.

The discussions in this paper have focused on the recognition accuracy of machines and humans on a verification task. This ignores many other aspects of performance of a face recognition system. Face recognition systems can search millions of mugshots, adjust to changing watch lists on demand, and process face imagery work 24 h a day. Humans are substantially better at recognizing familiar faces than unfamiliar faces. The algorithm development community has focused on recognition of unfamiliar faces. A challenge for the algorithm community is developing techniques that achieve human-level performance for familiar faces. This will most likely include developing an understanding of when this accuracy can be achieved.

Accepted conventional wisdom in the face recognition community is that humans are the most robust face recognition system. Humans perform face recognition across numerous imaging conditions; i.e., changes in natural illumination, pose, expression, imaging artifacts, etc. Like algorithms, human performance varies greatly under natural imaging conditions [45]. Also, human recognition of a person improves as a face transition from unfamiliar to familiar, and humans intuitively integrate all identify cues during recognition. The CMPA and face performance index has the potential to assist in developing algorithms that have these human capabilities. In turn, algorithms have the potential to serve as computational models that assist in explaining human face processing.

## Acknowledgments

PJP was supported by the Federal Bureau of Investigation and AJO was supported by the Department of Defense. The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST or U of Texas at Dallas.

## References

[1] A.J. O'Toole, P.J. Phillips, F. Jiang, J. Ayyad, N. Pénard, H. Abdi, Face recognition algorithms surpass humans matching faces across changes in illumination, IEEE Trans. PAMI 29 (1642–1646) (2007) 1642–1646.
[2] A.J. O'Toole, P.J. Phillips, A. Narvekar, Humans versus algorithms: comparisons from the FRVT, Eighth International Conference on Automatic Face and Gesture Recognition, 2006.
[3] P.J. Phillips, W.T. Scruggs, A.J. O'Toole, P.J. Flynn, K.W. Bowyer, C.L. Schott, M. Sharpe, FRVT 2006 and ICE 2006 large-scale results, IEEE Trans. PAMI 32 (2010) 831–846.
[4] A.J. O'Toole, X. An, J. Dunlop, V. Natu, P.J. Phillips, Comparing face recognition algorithms to humans on challenging tasks, ACM Trans. Appl. Percept. 9 (2012).

[5] J.J. DiCarlo, Untangling object recognition:which neuronal population codes can ex-7Q1125 plain human object recognition performance? in: Neural Computation: Population 713 Coding of High-Level Representations, 2011.
[6] N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis—connecting the branches of systems neuroscience, Front. Syst. Neurosci. 2 (2008).
[7] A.J. O'Toole, P.J. Phillips, S. Weimer, D.A. Roark, J. Ayyad, R. Barwick, J. Dunlop, Recognizing people from dynamic and stable faces and bodies: dissecting identity with a fusion approach, Vis. Res. 51 (2011) 74–83.
[8] A. Rice, P.J. Phillips, V. Natu, X. An, A.J. O'Toole, Unaware person recognition from the body when face identification fails, Psychol. Sci. 24 (2013) 2235–2243.
[9] P.J. Phillips, Improving face recognition technology, IEEE Comput. (2011) 96–98.
[10] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Ed.), 2005, pp. 947–954.
[11] P.J. Phillips, J.R. Beveridge, B.A. Draper, G. Givens, A.J. O'Toole, D.S. Bolme, J. Dunlop, Y.M. Lui, H. Sahibzada, S. Weimer, An introduction to the good, the bad, and the ugly face recognition challenge problem, in: Proceedings Ninth IEEE International Conference on Automatic Face and Gesture Recognition, 2011.
[12] J.R. Beveridge, D. Bolme, B.A. Draper, M. Teixera, The CSU face identification evaluation system, Mach. Vis. Appl. 16 (2005) 128–138.
[13] H. Moon, P.J. Phillips, Computational and performance aspects of PCA-based face-recognition algorithms, Perception 30 (2001) 303–321.
[14] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1991) 71–86.
[15] C. Liu, Capitalize on dimensionality increasing techniques for improving face recognition performance, IEEE Trans. PAMI 28 (2006) 725–737.
[16] M. Husken, B. Brauckmann, S. Gehlen, C. von der Malsburg, Strategies and benefits of fusion of 2D and 3D face recognition, in: IEEE Workshop on Face Recognition Grand Challenge Experiments (Ed.) Computer Society Digital Library, 2005.
[17] C. Xie, M. Savvides, V. Kumar, Kernel correlation filter based redundant class-dependence feature analysis (KCFA) on FRGC2.0 data, in: IEEE International Workshop Analysis and Modeling Faces and Gestures (Ed.), 2005, pp. 32–43.
[18] A.J. O'Toole, J. Harms, S.L. Snow, D.R. Hurst, M.R. Pappas, J.H. Ayyad, H. Abdi, A video database of moving faces and people, IEEE Trans. PAMI 27 (2005) 812–816.
[19] P.J. Phillips, P.J. Flynn, J.R. Beveridge, W.T. Scruggs, A.J. O'Toole, D. Bolme, K.W. Bowyer, B.A. Draper, G.H. Givens, Y.M. Lui, H. Zahibsada, J.A. Scallan III, S. Weimer, Overview of the multiple biometrics grand challenge, in: Proceedings Third IAPR International Conference on Biometrics (Ed.), 2009.
[20] Y.-C. Chen, V.M. Patel, P.J. Phillips, R. Chellappa, Dictionary-based face recognition from video, in: Proceedings of the European Conference on Computer Vision (Ed.), 2012.
[21] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press, 1975.
[22] N.A. Macmillan, C.D. Creelman, Detection Theory: A User's Guide, Cambridge University Press, Cambridge, 1991.
[23] G.H. Givens, J.R. Beveridge, P.J. Phillips, B.A. Draper, Y.M. Lui, D.S. Bolme, Introduction to face recognition and evaluation of algorithm performance, Comput. Stat. Data Anal. 67 (2013) 236–247.
[24] J.R. Beveridge, G.H. Givens, P.J. Phillips, B.A. Draper, Factors that influence algorithm performance in the Face Recognition Grand Challenge, Comput. Vis. Image Underst. 113 (2009) 750–762.
[25] R.S. Malpass, J. Kravitz, Recognition for faces of own and other race faces, J. Pers. Soc. Psychol. 13 (1969) 330–334.
[26] C.A. Meissner, J.C. Brigham, Thirty years of investigating the own-race bias in memory for faces: a meta-analytic review, Psychol. Public Policy Law 7 (2001) 3–35.
[27] R.K. Bothwell, J.C. Brigham, R.S. Malpass, Cross-racial identification, Personal. Soc. Psychol. Bull. 15 (1989) 19–25.
[28] P.N. Shapiro, S.D. Penrod, Meta-analysis of face identification studies, Psychol. Bull. 100 (1986) 139–156.
[29] P.J. Phillips, F. Jiang, A. Narvekar, A.J. O'Toole, An other-race effect for face recognition algorithms, ACM Trans. Appl. Percept. 8 (2011).
[30] N. Furl, P.J. Phillips, A.J. O'Toole, Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis, Cogn. Sci. 26 (2002) 797–815.
[31] M. Kirby, L. Sirovich, Application of the karhunen-loeve procedure for the characterization of human faces, IEEE Trans. PAMI 12 (1990) 103–108.
[32] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, J. Weng, Discriminant analysis of principal components for face recognition, in: H. Wechsler, P.J. Phillips, V. Bruce, F. Fogelman Soulie, T.S. Huang (Eds.), Face Recognition: From Theory to Applications, Springer-Verlag, Berlin, 1998, pp. 73–85.
[33] K. Etemad, R. Chellappa, Discriminant analysis for recognition of human face images, J. Opt. Soc. Am. A 14 (1997) 1724–1733.
[34] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs fisher faces: recognition using class specific linear projection, IEEE Trans. PAMI 19 (1997) 711–720.
[35] A.J. O'Toole, P.J. Phillips, X. An, J. Dunlop, Demographic effects on estimates of automatic face recognition performance, in: Proceedings, Ninth International Conference on Automatic Face and Gesture Recognition (Ed.), 2011.
[36] A.J. O'Toole, P.J. Phillips, X. An, J. Dunlop, Demographic effects on estimates of automatic face recognition performance, Image Vis. Comput. 30 (2012) 169–176.
[37] A. O'Toole, H. Abdi, F. Jiang, P.J. Phillips, Fusing face recognition algorithms and humans, IEEE Trans. Syst. Man Cybern. B 37 (2007) 1149–1155.
[38] I.S. Helland, Partial least squares regression and statistical models, Scand. J. Stat. 17 (1990) 97–114.
[39] R. Rosipal, N. Kramer, Overview and recent advances in partial least squares, in: C. Saunders, M. Grobelnik, S. Gunn, J. Shawe-Taylor (Eds.), Subspace, Latent Structure and Feature Selection Techniques, Springer, 2006, pp. 34–51.
[40] P.J. Grother, G.W. Quinn, P.J. Phillips, MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms, NISTIR 7709, National Institute of Standards and Technology, 2010.
[41] R. Russell, B. Duchaine, K. Nakayama, Super-recognizers: people with extraordinary face recognition ability, Psychon. Bull. Rev. 16 (2009) 252–257.
[42] A.M. Burton, S. Wilson, M. Cowan, V. Bruce, Face recognition in poor-quality video: evidence from security surveillance, Psychol. Sci. 10 (1999) 243–248.
[43] W.-J. Lee, C. Wilkinson, A. Memon, K. Houston, Matching unfamiliar faces from poor quality closed-circuit television (CCTV) footage, Axis Online J. CAHId 1 (2009) 19–28.
[44] C. Wilkinson, R. Evans, Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? Sci. Justice 49 (2009) 191–196.
[45] R. Jenkins, D. White, X. Van Montfort, A.M. Burton, Variability in photos of the same face, Cognition 121 (2011) 313–323.