FRVT 2006 and ICE 2006 Large-Scale Experimental Results

P. Jonathon Phillips, *Senior Member, IEEE,* W. Todd Scruggs, Alice J. O'Toole, Patrick J. Flynn, *Senior Member, IEEE,* Kevin W. Bowyer, *Fellow, IEEE,* Cathy L. Schott, and Matthew Sharpe

Abstract— **This paper describes the large-scale experimental results from the Face Recognition Vendor Test (FRVT) 2006 and the Iris Challenge Evaluation (ICE) 2006. The FRVT 2006 looks at recognition from high-resolution still frontal face images and three-dimensional (3D) face images, and measures performance for still frontal face images taken under controlled and uncontrolled illumination. The ICE 2006 evaluation reported verification performance for both left and right irises. The images in the ICE 2006 intentionally represent a broader range of quality than the ICE 2006 sensor would normally acquire. This includes images that did not pass the quality control software embedded in the sensor. The FRVT 2006 results from controlled still and 3D images document at least an order-of-magnitude improvement in recognition performance over the FRVT 2002. The FRVT 2006 and the ICE 2006 compared recognition performance from high-resolution still frontal face images, 3D face images, and the single-iris images. On the FRVT 2006 and the ICE 2006 datasets, recognition performance was comparable for high-resolution frontal face, 3D face, and the iris images. In an experiment comparing human and algorithms on matching face identity across changes in illumination on frontal face images, the best-performing algorithms were more accurate than humans on unfamiliar faces.**

Index Terms— **Biometrics, Face Recognition, Iris Recognition, Evaluations, Human Performance.**

I. Introduction

FACE recognition is a vibrant multi-disciplinary field with active research and commercial efforts [1]. The Face Recognition Vendor Test (FRVT) 2006 is the latest in a series of evaluations for face recognition that began in 1993 with the Face Recognition Technology (FERET) program [2][3]. With the expiration of the Flom and Safir [4] iris recognition patent in 2005, iris recognition algorithm development has become an active research topic [5]. The Iris Challenge Evaluation (ICE) 2006 is the first independent technology evaluation of iris recognition algorithms. Since face and iris are competitive and complementary biometric technologies, conducting a simultaneous technology evaluation allowed for assessments of each biometric and comparison of their capabilities.

The FRVT 2006 and the ICE 2006 are independent technology evaluations of face and iris recognition technology, respectively. An independent evaluation is conducted by an institution with no formal ties to those being evaluated and that does not benefit from the results. The purpose of a technology evaluation is to evaluate the performance of the underlying technology [6]. A technology evaluation is different from a scenario evaluation, which measures overall system performance for a prototype scenario that models an application domain. Both the FRVT 2006 and ICE 2006 share the same protocol and they report results on biometric samples from the FRVT 2006 and ICE 2006 multi-biometric dataset. Together, these evaluations constitute the first multi-biometric technology evaluation that measures performance of iris, still face, and three-dimensional (3D) face recognition techniques.

The FRVT 2006 and the ICE 2006 were designed to measure performance of state-of-the-art algorithms on the FRVT 2006 and the ICE 2006 datasets. To obtain unbiased measures of performance, algorithms were tested on sequestered data. These two evaluations were not designed to measure performance of operational face or iris recognition systems. The FRVT 2006 measures performance on three datasets. Two of the datasets collected frontal face images with multi-megapixel commercial cameras. These two datasets measure the art-of-the-possible (what is possible with state-of-the-art algorithms and data collection protocols). The third dataset was an operational dataset collected by the U.S. Department of State. Performance results on this dataset were reported in the previous FRVT 2002 and allow for a direct comparison between the FRVT 2002 and the FRVT 2006.

The key novel accomplishments of the FRVT 2006 and the ICE 2006 are:

- The FRVT 2006 established the first independent performance benchmark for 3D face recognition technology and provides an update of face recognition performance from still frontal images collected under controlled and uncontrolled illumination.
- The ICE 2006 established the first independent performance benchmark for iris recognition matching technology. The ICE 2006 is different than the Independent Test of Iris Recognition Technology (ITIRT) and Iris '06 that evaluated cross-sensor performance using the same matching algorithm [7],[8].
- The FRVT 2006 and the ICE 2006 are the first technology evaluations that compared iris recognition, high-resolution still frontal face recognition, and 3D face recognition performance.
- The Face Recognition Grand Challenge (FRGC) was a face recognition technology development effort with the goal of decreasing the error rate of face recognition algorithms by an order of magnitude over performance reported in the FRVT 2002 [9] [10][11]. The FRGC goal of an order of magnitude decrease in error rates was to be obtained on frontal still face images taken under controlled illumination conditions. One of the key goals of the FRVT 2006 was to establish if the goals of the FRGC were met. The FRVT 2006 documented

Manuscript received February 28, 2008.

P. J. Phillips is with the National Institute of Standards and Technology, Gaithersburg, MD 20899; W. T. Scruggs and M. Sharpe are with SAIC, Arlington, VA 22203; A. J. O'Toole is with School of Behavioral and Brain Sciences, The U. of Texas at Dallas, Richardson, TX 75083-0688; P. J. Flynn and K. W. Bowyer are with Computer Science & Engineering Department, U. of Notre Dame, Notre Dame, IN 46556; C. L. Schott is with Schafer Corp., Arlington, VA 22203. Please direct correspondence to P. J. Phillips at jonathon@nist.gov.

a decrease in the error rate by at least an order of magnitude over what was observed in the FRVT 2002 when matching frontal faces taken under controlled illumination conditions.

- The FRVT 2006 documented significant progress in face recognition when frontal faces are matched across different lighting conditions.
- For the first time in a biometric evaluation, the FRVT 2006 directly compared human and machine face recognition performance.

The FRVT 2006 and the ICE 2006 results in this report support the claims above. The report is organized as follows. Sections II provides background material for the two evaluations. Section III presents the ICE 2006 results, and Section IV presents the FRVT 2006 results. In Section IV, the still portion of the FRVT 2006, including human performance, is discussed first, followed by the 3D face recognition benchmark. The multi-biometric aspects of the ICE 2006 and the FRVT 2006 are discussed in section V and overall conclusions are presented and discussed in section VI.

II. ICE 2006 and FRVT 2006 Overview

The FRVT 2006 and the ICE 2006 protocols were built on the FRVT 2002 and FERET evaluation protocols [3], [11]. The primary modification to these protocols is that testing was conducted on executables delivered by participants and run on the National Institute of Standards and Technology's (NIST) servers. For the FRVT 2006, performance is reported on multiple sequestered datasets. All data was sequestered at the subject level; e.g., biometric samples from subjects in the FRGC or the ICE 2005 challenge problems were not included in the FRVT 2006 and the ICE 2006.

A. Protocol

Both the FRVT 2006 and the ICE 2006 were algorithm evaluations in which participants had to deliver algorithms to NIST to be evaluated. The FRVT 2006 executables had to be received by NIST by 30 January 2006 and by 15 June 2006 for the ICE 2006. The FRVT 2006 and the ICE 2006 were open to academia, industry, and research laboratories. Participants could submit multiple algorithms.

The format for submissions was binary executables that could be run independently on the test server. All submitted executables had to run using a specified set of command line arguments. The command line arguments included an experiment parameter file, files that contained the sets of biometric samples to be matched, and name of the output similarity file.

There were a number of options for submissions to the FRVT 2006. Participants could submit both fully automatic or partially automatic algorithms. Partially automatic algorithms were provided with the coordinates of the centers of the eyes; fully automatic algorithms were not provided with any information about the location of the face in the images. All participants were required to submit algorithms that performed one-to-one matching of face images with an option for submitting algorithms that performed normalized matching. Subsection II-C describes one-to-one and normalized matching. The FRVT 2006 had an optional face image quality task. For the quality task, executables gave a quality score for each face image. The quality score had to be an integer in the range between 0 and 100, with 100 being the highest quality. A quality score is a number that rates an image's utility to a recognition system and should be predictive of performance [12]. All submissions were required to be able to generate a complete similarity matrix of matching scores for all pairs of images in a 16,028 image set in 72 CPU-hours or less on the NIST servers.

The test system hardware for the FRVT 2006 and the ICE 2006 was a Dell PowerEdge 850 server with a single Intel Pentium 4 3.6GHz 660 processor, 2MB of 800Mhz cache, and 4GB of 533MHz DDR2 RAM. At no time did the test system have access to the Internet. The FRVT 2006 and the ICE 2006 allowed executables that would run under Windows Server 2003 (standard edition) and Linux Fedora Core 3 operating systems.

The FRVT 2006 results in this article are limited to the fully automatic algorithms. Table II lists the FRVT 2006 and the ICE 2006 algorithms whose results are presented in the body of this article. Algorithmic details are only available for the U of Houston and Viisage submissions [13][14]. The FRVT 2006 results are presented in three categories: controlled illumination, 3D face, and uncontrolled illumination. In the main body of the article, performance results are only presented for the better-performing algorithms and generally results are only given for one algorithm from each participating group. Results for all algorithms are in the online supplemental material.

The ICE 2006 was restricted to fully automatic algorithms and one-to-one matching. There were no time limits on the ICE 2006 submissions and there was an optional quality task available. The results presented in this article are limited to algorithms that completed the ICE 2006 experiments in 30 days or less. For each of the three groups that had algorithms that completed the experiments in the time limit, results for only one algorithm are presented in the body of the paper. Results for all algorithms are in the online supplemental material. Flynn and Phillips [15] report results of analyzing the quality scores.

B. Data

Results for the FRVT 2006 and the ICE 2006 are reported on three datasets: the FRVT 2006 and the ICE 2006 multi-biometric collected at the University of Notre Dame, the Sandia high-resolution frontal face images, and the Dept. of State low-resolution frontal images. The multi-biometric dataset consists of still and 3D face images, and iris images. The multi-biometric dataset makes it possible to measure performance on still face, 3D face, and iris on the same set of subjects.

The first dataset is *the FRVT 2006 and the ICE 2006 multi-biometric dataset*, which consists of high-resolution still frontal facial images (referred to as the *Notre Dame* dataset), frontal 3D facial scans (referred to as the *3D* dataset), and iris images, see Figure 1. The Notre Dame high-resolution images were taken with a 6 Mega-pixel Nikon D70 camera, the 3D images with a Minolta Vivid 900/910 sensor, and the iris images with a LG EOU 2200. All the sensors chosen were state-of-the-art in the Fall of 2003 and the Winter 2004.

The ICE 2006 images were collected with the LG EOU 2200 and intentionally represent a broader range of quality than the sensor would normally acquire. This includes iris images that did not pass the quality control software embedded in the LG EOU 2200. The LG EOU 2200 is a complete acquisition system and has automatic image quality control checks.

The image quality software embedded in the LG EOU 2200 is one of numerous iris quality measures. Flynn and Phillips [15] showed that in the ICE 2006, quality measures are paired with matching algorithms; different quality measures are not correlated; and none of the iris quality measures generalize to all algorithms in the ICE 2006. This implies that evaluations risk being biased against submissions if the iris images are screened by a quality measure. Prior to the start of the multi-biometric data collection, an arrangement was made to minimize the effect of the LG EOU 2200 quality screening software on the data collection. The subsequent analysis of the effect of quality scores on performance shows that this decision was appropriate.

By agreement between U. of Notre Dame and Iridian, a modified version of the acquisition software was provided. The modified software allowed all images from the sensor to be saved under certain conditions, as explained below.

The iris images are 480x640 in resolution. For most "good" iris images, the diameter of the iris in the image exceeds 200 pixels. The images are stored with 8 bits of intensity, but every third intensity level is unused. This is the result of a contrast stretching automatically applied within the LG EOU 2200 system. The iris images were digitized from NTSC video and hence the iris images may have interlace artifacts due to motion of the subject.

In our acquisitions, the subject was seated in front of the system. The system provides recorded voice prompts to aid the subject to position their eye at the appropriate distance from the sensor. The system takes images in "shots" of three, with each image corresponding to illumination of one of the three near-infrared light-emitting diodes (LEDs) used to illuminate the iris.

For a given subject at a given iris acquistion session, two "shots" of three images each are taken for each eye, for a total of 12 images. The system provides a feedback sound when an acceptable shot of images is taken. An acceptable shot has one or more images that pass the LG EOU 2200's built-in quality checks, but all three images are saved. If none of the three images pass the built-in quality checks, then none of the three images are saved. At least one third of the iris images do pass the Iridian quality control checks, and up to two thirds do not pass. A manual quality control step was performed at Notre Dame to remove images in which, for example, the eye was not visible at all due to the subject having turned their head.

In the multi-biometric dataset, biometric samples for all three biometrics were collected from the same subject pool. The Notre Dame high-resolution face still images in the multi-biometric data set were collected under controlled and uncontrolled illumination conditions. The average face size for the controlled images was 400 pixels between the centers of the eyes and 190 pixels for the uncontrolled images. Table I provides a summary of the face size in the still images, broken out by dataset and illumination condition. The 3D and iris data were collected by sensors that contain an active illumination component as an integral part of the sensor. For the 3D sensor, the active illumination is a laser light stripe that sweeps the scene, and this enables triangulation-based calculation of the 3D shape.

The second dataset is the *Sandia dataset*, which consisted of high-resolution frontal facial images taken under both controlled and uncontrolled illumination. The Sandia dataset was collected at the Sandia National Laboratory. The Sandia images were taken with a 4 Megapixel Canon PowerShot G2. The average face size for the controlled images was 350 pixels between the centers of the eyes and 110 pixels for the uncontrolled images.

TABLE I
FOR THE STILL FACES, FACE SIZE IN PIXELS BETWEEN THE CENTERS OF THE EYES IS SUMMARIZED. AVERAGE FACE SIZE IS GIVEN BROKEN OUT BY DATASET AND ILLUMINATION CONDITION.

Dataset	Illumination	Face size
Notre Dame	controlled	400
Sandia	controlled	350
Dept. of State	controlled	75
Notre Dame	uncontrolled	190
Sandia	uncontrolled	110

The third is the *Dept. of State dataset*, consisting of low-resolution frontal facial images taken under controlled illumination conditions, see Figure 2. The images in the Dept. of State dataset were provided by the Visa Services Directorate, Bureau of Consular Affairs of the U.S. Department of State. Consequently, results on the Dept. of State dataset provide a performance benchmark for operational low-resolution highly compressed imagery. The Dept. of State dataset is the same dataset used in the HCInt portion of the FRVT 2002. The Dept. of State images were JPEG compressed to a size of approximately 10,000 bytes. They have an average face size of 75 pixels between the centers of the eyes.

The maximum time lapse between samples of subjects was 8 months for the Notre Dame still and 3D face images used in the FRVT2006, and 17 months for the iris images used in the ICE2006. For the Sandia dataset, the maximum time lapse between samples of subjects was 20 months. The authors are not aware of any peer reviewed papers or scientific technical reports that measure performance of iris and 3D face for greater time lapses [1],[16],[5]. The Dept. of State face image dataset used in the FRVT2006 includes pairs of samples from subjects where the elapsed time is up to three years. Demographic information for each dataset is provided in Table III. Demographic information is given for sex, race, and age.

C. Performance statistics

The performance statistics in the FRVT 2006 and the ICE 2006 are based on those in the FRVT 2002 [17]. For the FRVT 2006 and the ICE 2006, performance is reported for verification. Verification performance is measured by false reject rate (FRR) and false accept rate (FAR), see Appendix I for a review of FRR and FAR.

Algorithms were required to compare two biometric samples and return a scalar similarity score. In the FRVT 2006, biometrics samples are limited to still and 3D face images and in the ICE 2006 samples to still iris images. A similarity score is a measure of the sameness of identity of the individuals appearing in two biometric samples. A large similarity score implies that the identifies are more likely to be the same. Algorithms could report either a similarity score or distance measure. Distance measures, where a small value indicates sameness of identity, have their values negated before any processing.

The FRVT 2006 and the ICE 2006 analyses were structured around similarity matrices. In the evaluations, an algorithm is required to compute a similarity score between all pairs of samples in a *query* set, $\mathcal{Q}$, with all samples in a *target* set $\mathcal{T}$. The result is a similarity matrix whose ij-th element is the similarity score s_{ij} between the i-th sample of $\mathcal{T}$ and the j-th sample of $\mathcal{Q}$. A target set represents the set of biometric samples known to a system, and a query is a sample presented to a system for verification. A similarity score s_{ij} is a *match* if the i-th sample of $\mathcal{T}$ and the j-th sample of $\mathcal{Q}$ are of the same person, and a *non-match* if they are samples of different people. A sample of a subject in a query set is a *true impostor* if that subject is not in the corresponding target set. True impostors are important for measuring performance in normalized matching.

In FRVT 2006, performance is computed for both one-to-one matching and normalized matching. In one-to-one matching, a similarity score s_{ij} is only a function of target sample t_i and query sample q_j, and is independent of the other samples in either the target or query set. One-to-one matching makes it possible to have multiple samples in a target set because the multiple samples do not affect the computation of s_{ij}.

Normalized matching allows for algorithms to adjust their representation based on the subjects in a target set. For normalized matching, the target set contains only one sample per person. This type of target set is referred to as a *gallery*. In normalized matching, a similarity score s_{ij} is a function of a gallery sample t_i, a query sample q_j, and the gallery $\mathcal{G}$ that contains t_i. If the contents of a gallery change, then similarity score s_{ij} could change. The similarity score s_{ij} is independent of the other samples in the query set.

The performance of a biometric system will vary with different sets of biometric samples. This is true even when biometric samples are taken under the same conditions; e.g., in face recognition, matching images taken under controlled illumination. It is important to measure both the overall performance of a biometric system and the scale of the variability of the performance statistic. Measuring variability quantifies statistical uncertainty. In the FRVT 2006 and the ICE 2006, performance variability is measured by partitioning a target set into a set of smaller target sets. Performance is then computed on each of the partitions. For each partition, the FRR at a FAR = 0.001 is computed, where the FAR is computed for each partition. If there are n partitions, there are n FRRs, and the n FRRs are summarized by a boxplot. See Appendix I for a review of boxplots. Table IV lists the number of images, subjects, and partitions for each FRVT 2006 and ICE 2006 experiment.

For example, the Dept. of State dataset was partitioned into twelve small target sets of 3,000 images. These twelve small target sets were the same as the twelve small galleries in the HCInt portion of the FRVT 2002 and allow for a direct comparison of results between the FRVT 2002 and the FRVT 2006. Each of the twelve targets consisted of one image of each of 3,000 individuals, and the twelve target sets were disjoint. There were twelve corresponding query sets which consisted of 12,000 images each. The query set consisted of two images of each of the 3,000 people in the target set and two images of each of 3,000 people not in the target set. For each algorithm, the FRR at a FAR of 0.001 was computed independently for each partition. Performance for each algorithm at a FAR of 0.001 was characterized by twelve FRRs which were summarized by a boxplot.

The Notre Dame still face and 3D face data were collected over two academic semesters: Fall 2004 and Spring 2005. The target set and its partitions consisted of images taken in the Fall 2004 semester and the query set consisted of images collected in the Spring 2005 semester. Only the target was partitioned and each of the partitions was matched to the query set. For each partition, the FAR was computed from true impostors. Because of the requirements for normalized matching, the target set was partitioned into a set of galleries.

The face images in the Sandia data were collected over a 20 month period. The Sandia target sets consisted of images collected in the first five months of data collection and the query sets consisted of images collected in the subsequent 15 month period. Because of the requirements for normalized matching, the target set was partitioned into a set of galleries and the true impostor criteria was imposed for computing FAR.

The images for the ICE 2006 were collected over three academic semeseters: Spring 2004, Fall 2004 and Spring 2005. In computing performance, all similarity scores are cross semesters; i.e., iris images taken in the same semester were not compared. The iris image from the earlier semester was always in the target set. For the ICE 2006, performance is broken out by left and right iris. There were 30 partitions for the left eye and 30 partitions for the right eye. Since the ICE 2005 only reported one-to-one match performance, there were multiple iris images per subject in the partition target sets and the true impostor criteria was not imposed in computing performance.

In order to validly compare the results of these tests, we must choose relevant point(s) on the receiver operating characteristic (ROC) curve for these tests. The critical issue for determining valid comparison points is test size. Mansfield and Wayman [18] state, "The number of people tested is more significant than the total number of attempts in determining test accuracy."

The number of subjects in the FRVT 2006 and the ICE 2006 ranged from 240 to 12,000, see Table IV; and the number of non-match comparisons ranged from about 700,000 to more than 250 million. The experiments using the Dept. of State dataset had 216 million comparisons and the ICE 2006 experiments had over 250 million for each eye. An analysis of FRRs at an FAR of one in a 100,000 means that for the smaller experiments (about 300 subjects and 750,000 non-match comparisons) the expected number of false matches is only seven; at an FAR of one in 10,000, seventy false matches are expected. This number of errors is too small to definitively compare these tests. Further, this data is highly correlated because of the re-use of same subjects' data, and it is correlated on the score level. Within the non-match distribution of each experiment, each person contributes on average anywhere from 2800 non-match similarity scores to 2.3 million non-match similarity scores. Given the number of subjects and comparisons in these studies, we chose to report and compare the FRRs at an FAR of 1 out of 1000. For completeness, the supplemental material presents FRR at FARs of 0.01, 0.001, and 0.0001. The primary difference between the FRRs at the three FARs is that FRR decreases as FAR decreases. If there is a significant change in the relative FRRs among the algorithms, it is noted in the text of this article.

III. ICE 2006

The ICE 2006 establishes the first independent performance benchmark for iris recognition algorithms. Performance for the ICE 2006 benchmark is presented in Figure 3 for algorithms from three groups: Sagem-Iridian (SG-2), Iritech (IRTCH-2), and Cambridge (CAM-2), see Figure 10 for an explanation of boxplots. The interquartile range for all three algorithms overlaps, with the largest amount of overlap between Iritech (IRTCH-2), and Cambridge (CAM-2). Over all three algorithms, the smallest

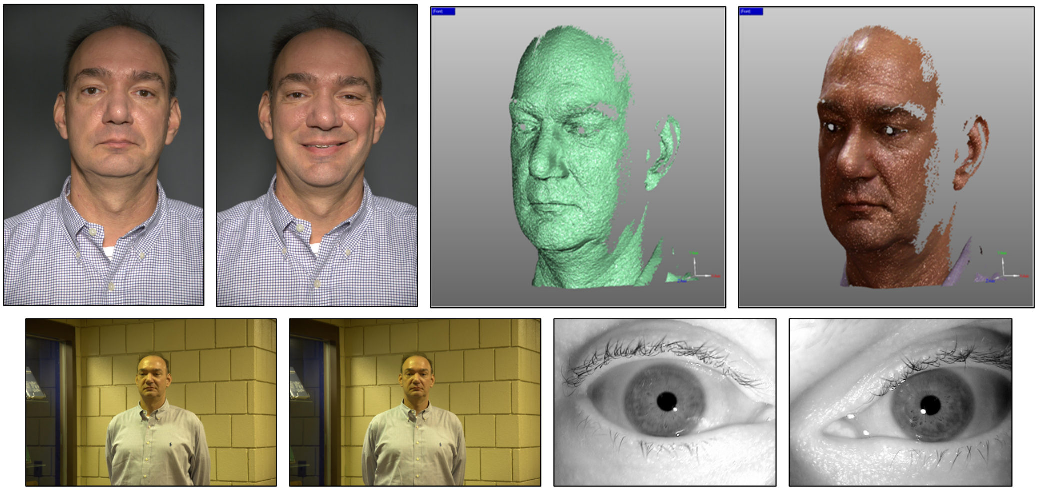

Fig. 1. An example of the types of images used in the FRVT 2006 and the ICE 2006. The two left frontal images in the top row were taken under controlled illumination with neutral and smiling expressions. The two left images in the bottom row were taken under uncontrolled illumination with neutral and smiling expressions. The two right images in the top row show the shape channel and texture channel pasted on the shape channel for a 3D facial image. The two right images in the bottom row show right and left iris images. All samples are from the multi-biometric dataset.

TABLE II

The list of algorithms covered in the large scale analysis. Column headings identify each participant group and five biometric matching tasks in the FRVT 2006 and the ICE 2006. The organization that submitted an algorithm is listed in the group column. The abbreviations used in the figures are presented in the table. A blank cell in a column for a group means they did not submit an algorithm for the task in that column.

Group	Iris	Still 1to1	Still norm	3D 1to1	3D norm	Shape
U. of Cambridge	Cam-2					
Cognitec		Cog1-1to1	Cog1-norm	Cog1-3D	Cog1-3D-n	
Geometrix						Geo-Sh
U. of Houston						Ho3-Sh
Identix		Idx4-1to1	Idx1-norm			
Iritech	Irtch-2					
Neven Vision		NV1-1to1	NV1-norm			
Rafael		Ra-1to1	Ra-norm			
Sagem		SG2-1to1	SG2-norm			
Sagem-Iridian	SI-2					
SAIT		ST-1to1	ST-norm			
Toshiba		To2-1to1	To1-norm			
Tsinghua U.		Ts2-1to1	Ts2-norm	Ts1-3D		
Viisage		V-1to1	V-norm	V-3D	V-3D-n	

TABLE III

Demographic breakout is given for sex, race, and age. Breakout values within a demographic category are by percent. If the number of subjects in a given category is less than 2.5%, then the cell is left blank. For the Dept. of State dataset, less than 0.5% of the subjects have country of birth other than Mexico. For the Dept. of State dataset, the age categories are 18-27, 28-37, 38-47, 48-57, and 58+.

	Sex		Race			Age				
Dataset	Female	Male	Caucasian	East Asian	Hispanic	18-29	30-39	40-49	50-59	60+
Notre Dame	62	38	76	13		92	7			
Sandia	55	45	64		21	15	11	23	35	18
Dept. of State	50	50			$\sim$100	38	25	15	11	10

TABLE IV

SUMMARY OF EXPERIMENTS IN THE FRVT 2006 AND THE ICE 2006. THE FIRST COLUMN LISTS THE EXPERIMENTS AND THE SECOND COLUMN THE DATASET. THE TARGET SET (QUERY SET) COLUMN LISTS THE TYPE OF IMAGES THE TARGET (QUERY) SET. THE NUMBER OF IMAGES AND THE NUMBER SUBJECTS IN AN EXPERIMENT ARE GIVEN. THE LAST COLUMN STATES THE NUMBER OF PARTITIONS USED TO COMPUTE PERFORMANCE.

Experiment	Dataset	Target set	Query set	No. subjects	No. images	No. Partitions
Controlled-face	Notre Dame	controlled still face	controlled still face	336	7496	26
Controlled-face	Sandia	controlled still face	controlled still face	263	14,365	20
Controlled-face	Dept. of State	controlled still face	controlled still face	36,000	108,000	12
Uncontrolled-face	Notre Dame	controlled still face	uncontrolled still face	335	5402	26
Uncontrolled-face	Sandia	controlled still face	uncontrolled still face	257	7192	20
3D-face	3D	3D face	3D face	330	3589	13
Iris right-eye	iris	iris right-eye	iris right-eye	240	29,056	30
Iris left-eye	iris	iris left-eye	iris left-eye	240	30,502	30

Fig. 2. Reasonable representations of images in the Dept. of State dataset. Because of privacy consideration, actual images could not be shown.

interquartile is a FRR of 0.009 at a FAR of 0.001 and the largest interquartile is a FRR of 0.026 at a FAR of 0.001.

The results in the ICE 2005, a technology development effort, showed that for the top four groups, recognition performance on the right eye was better than the left eye. In the ICE 2006, the median FRR for the left eye is always smaller than the median FRR for the right eye; however, the range of the boxplots is similar. The results of the ICE 2006 show the same relative performance level. This is seen in Figure 3 by the range of the boxplots for all three algorithms. Hence, the difference in performance observed in ICE 2005 was not confirmed by the results in the ICE 2006. The difference between the ICE 2005 and the ICE 2006 conclusions may be because of the smaller number of samples in the ICE 2005 than the ICE 2006 (2953 versus 59,558) and because the ICE 2005 characterized performance for each eye by one partition versus 30 partitions for each eye in the ICE 2006.

The execution time varied significantly between the Cambridge submission and the Sagem-Iridian and Iritech submissions. The Cambridge algorithm (CAM-2) took 6 hours to complete the ICE 2006 large scale experiments and the Sagem-Iridian (SI-2) algorithm and Iritech (IRTCH-2) algorithm took approximately 300 hours.

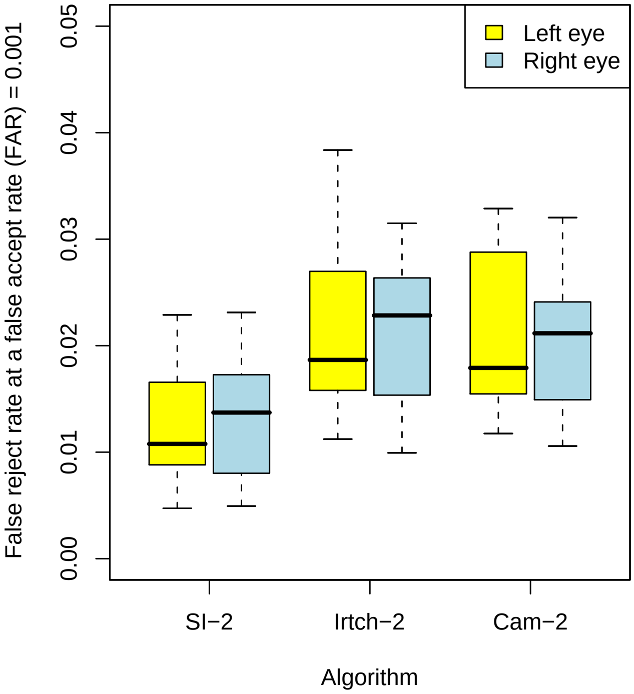

Fig. 3. Summary performance of the ICE 2006. Results are presented for three groups: Cambridge (Cam-2), Iritech (IrTch-2) and Sagem-Iridian (SI-2). Performance is broken out by right and left eyes. The false reject rate (FRR) at a false accept rate (FAR) of 0.001 is reported. Performance is reported for 29,056 right and 30,502 left iris images from 240 subjects with 30 partitions for each eye.

IV. FRVT 2006

The FRVT 2006 large-scale experiments documented progress in face recognition in four areas. First, the FRGC goal of improving performance by an order of magnitude over FRVT 2002 was achieved. Second, the FRVT 2006 established the first 3D face recognition benchmark. Third, the FRVT 2006 showed significant progress has been made in matching faces across changes in lighting. Fourth, on the task of matching face identity across changes in illumination on the Sandia dataset, using a comparison based on an identical set of frontal face image pairs, the best-performing algorithms performed more accurately than humans on unfamiliar faces.

A. Controlled Illumination

The goal of the FRGC was to improve face recognition performance to achieve a FRR of 0.02 at a FAR of 0.001 for matching face images taken under controlled illumination. This goal was exceeded on the FRVT 2006 Notre Dame still face dataset with algorithms achieving a FRR of 0.01.

Figure 4 summarizes performance of face recognition for still images under controlled illumination for three datasets: Notre Dame, Sandia, and Dept. of State. On the Notre Dame dataset, four algorithms met or exceeded the FRGC goal of a FRR of 0.02. These algorithms are from Neven Vision (NV1-NORM and NV1-1TO1[1]), Viisage (V-NORM) and Cognitec (COG1-NORM). On the Sandia dataset, the Neven Vision (NV1-NORM) algorithm with a FRR interquartile range of 0.021 to 0.023 came close to meeting the FRGC goal.

On the Notre Dame dataset, three algorithms had performance that crossed the FRR of 0.01 at a FAR of 0.001 threshold. The FRR interquartile range for the three algorithms are 0.006 to 0.015 for NV1-NORM, 0.008 to 0.016 for NV1-1TO1, and 0.010 to 0.017 for V-NORM.

The best performer on the Dept. of State dataset at FAR=0.001 was Toshiba (TO1-NORM) with an interquartile FRR range of 0.024 to 0.027. Four algorithms, Neven Vision (NV1-NORM), Viisage (V-NORM), Cognitec (COG1-NORM), and Sagem (SG2-NORM) had performance in the neighborhood of FRR = 0.05 at a FAR of 0.001. The lowest quartile from this grouping was a FRR of 0.043 and the highest was a FRR of 0.053. While Toshiba performed extremely well on the Dept. of State data set at FAR=0.01 and FAR=0.001, their performance was not consistent across all the still datasets.

For the four algorithms Neven Vision (NV1-NORM), Viisage (V-NORM), Cognitec (COG1-NORM), and SAIT (ST-NORM), there is a clear ranking of the difficulty of the three datasets, with the Dept. of State being the most difficult and the Notre Dame dataset being easiest; i.e., having the best performance. The primary difference between the three datasets is the size of the faces and consistency of the lighting.

B. 3D Face Recognition

The FRVT 2006 provides the first benchmarks of 3D face recognition technology. Benchmarks are provided for one-to-one and normalization approaches that use both shape and texture, and for one-to-one shape-only techniques. Performance for 3D face recognition is summarized in Figure 5. All results are from the 3D portion of the multi-biometric dataset.

Performance on the 3D dataset meets the FRGC goal of an order of magnitude improvement in performance. The best performers for 3D have a FRR interquartile range of 0.005 to 0.015 at a FAR of 0.001 for the Viisage normalization (V-3D-N) algorithm and a FRR interquartile range of 0.016 to 0.031 at a FAR of 0.001 for the Viisage 3D one-to-one (V-3D) algorithm. Both algorithms met the FRGC performance goal. The shape only benchmark was set by the Geometrix (GEO-SH) and the U. of Houston (HO3-SH) submissions.

On the FRVT 3D dataset, the normalized algorithms performed better than the one-to-one algorithms. This is seen by comparing the results for the Cognitec and Viisage 3D normalized algorithms (COG1-3D-N and V-3D-N) to their counterpart one-to-one algorithms (V-3D and V-3D).

C. Uncontrolled Illumination

When compared with the FRGC results, the FRVT 2006 shows a significant improvement in recognition when matching faces across changes in lighting. In these experiments, the enrolled images are frontal facial images taken under *controlled* illumination and the probe images are frontal facial images taken under *uncontrolled* illumination, see Figure 1 for sample images. These experiments will be referred to as *uncontrolled* experiments.

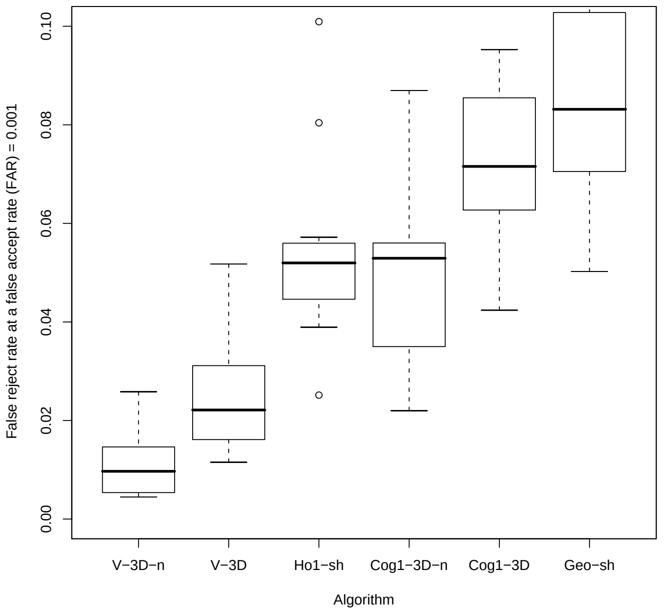

Fig. 5. Summary of performance for 3D face recognition algorithms.

Performance on controlled versus uncontrolled experiments was measured on the Notre Dame and Sandia datasets. Figure 6 summarizes the results of the uncontrolled experiments.

In January 2005, the three best self-reported results in the FRGC uncontrolled illumination experiments were FRRs of 0.24, 0.39, and 0.56 at a FAR of 0.001 [10][2]. In FRVT 2006, four algorithms, Cognitec (COG), Neven Vision (NV1-NORM), SAIT (ST-NORM), and Viisage (V-NORM) had performance on both the Notre Dame and Sandia datasets that was better than the best FRGC results. On the Notre Dame dataset, SAIT (ST-NORM) had a FRR interquartile range of 0.103 to 0.130 at a FAR of 0.001. On the Sandia dataset Viisage (V-NORM) had a FRR interquartile range of 0.119 to 0.146 at a FAR of 0.001.

In terms of difficulty level, the results in Figure 6 show that there is no clear ranking of the two datasets in terms of difficulty since three algorithms have better performance on the Sandia dataset; two algorithms had better performance on the Notre Dame datasets; and two algorithms had equivalent performance for both datasets. Restricting our attention to the best results, we see comparable performance for SAIT (ST-NORM) on the Notre Dame dataset and Viisage (V-NORM) on both datasets.

D. Human Performance

FRVT 2006 integrated human face recognition performance into an evaluation for the first time. This inclusion allowed a direct

[1] The algorithm NV1-1TO1 was not plotted on Figure 4.

[2] These results are on ROC III for Experiment 4 on the FRGC v2 challenge problem.

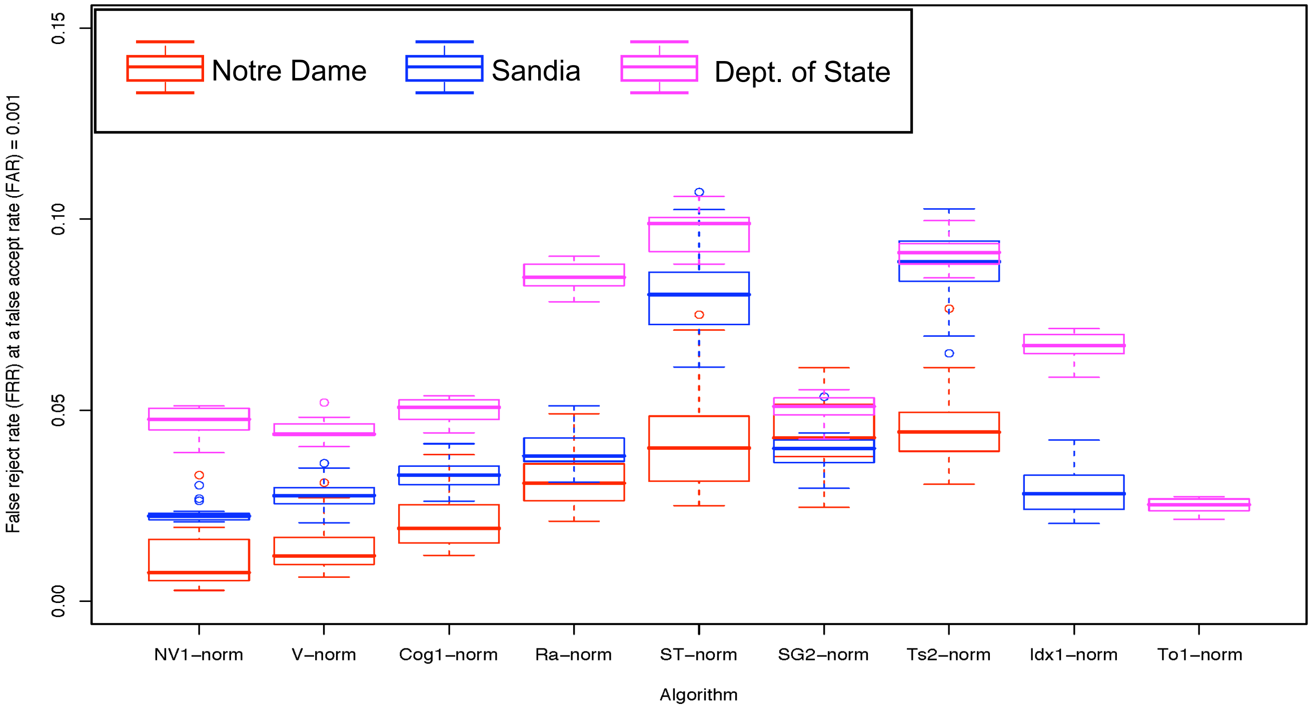

Fig. 4. Summary of still face recognition performance on the Notre Dame, Sandia, and Dept. of State datasets. Each column in the graph reports performance for one algorithm with results provided for up to three data sets. For each algorithm, the performance results on a data set are reported by a different color boxplot. For a Sagem (SG2-NORM) algorithm, the body of the boxplots overlap for all three datasets. For a Tsinghua (TS2-NORM) algorithm, the body of the boxplots overlaps the Sandia and Dept. of State datasets. For Identix (IDX1-NORM) and Toshiba (TO1-NORM), performance was outside the range of this graph for at least one dataset.

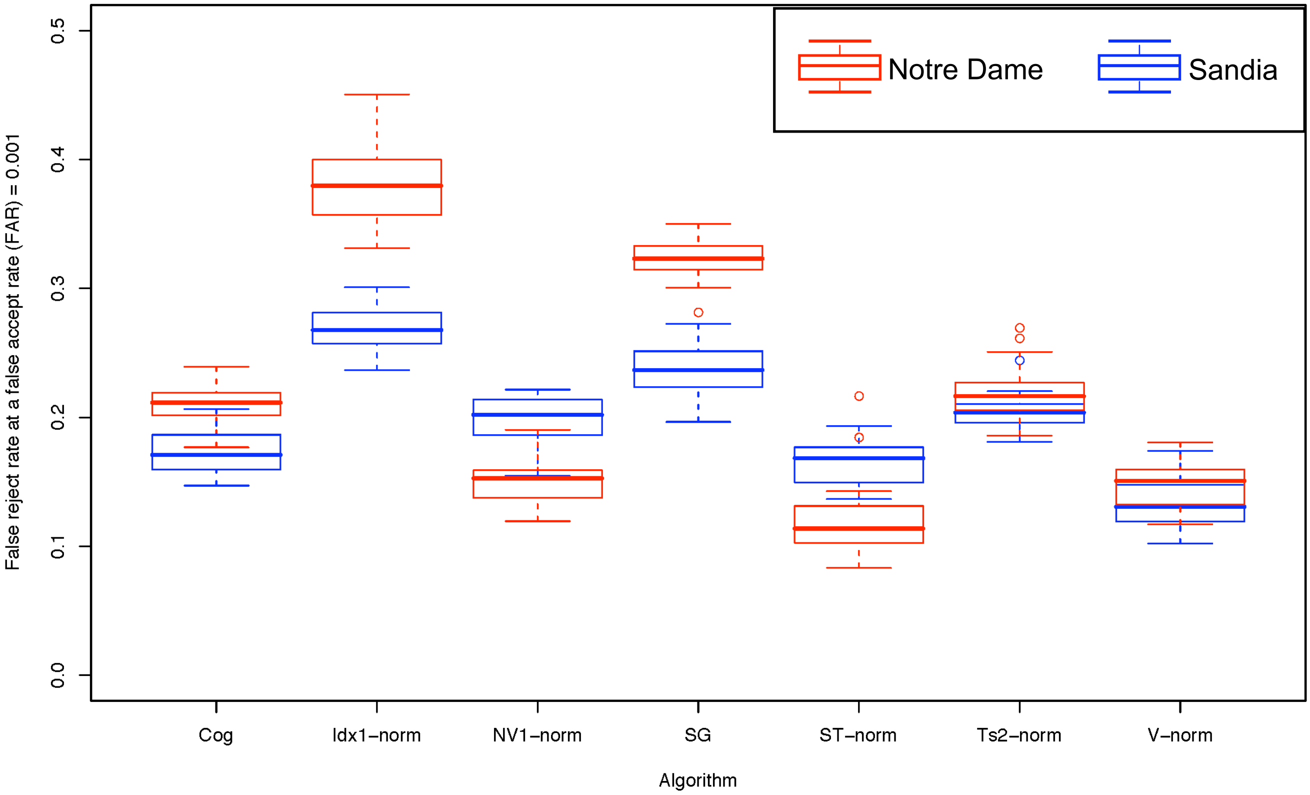

Fig. 6. Summary of still face recognition performance across illumination changes on the Notre Dame and Sandia datasets. For Cognitec and Sagem, results for the COG1-NORM and SG2-NORM algorithms are reported on the Notre Dame dataset, and results for the Cog1-1to1 and SG1-1to1 algorithms are reported on the Sandia dataset.

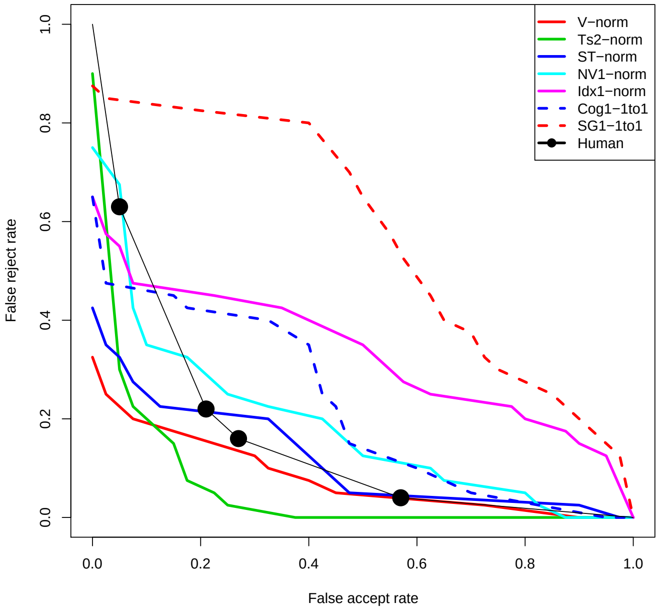

Fig. 7. ROC of human and computer performance on matching faces across illumination changes. ROCs for algorithms in Figure 6 are plotted. The ROC plots FAR against FRR. Perfect performance would be the lower left hand corner (FAR=FRR=0).

comparison between humans and state-of-the-art computer algorithms. In this study, we focused on recognition across changes in illumination. Specifically, humans matched faces taken under controlled illumination against faces taken under uncontrolled illumination on images from the Sandia dataset.

The human experiments were set up as a face identity match task to be comparable to the protocol used in the FRVT 2006. Although some algorithms may have had a training phase, the faces tested in the FRVT 2006 were sequestered and it was not possible for the algorithms to training using the faces to be matched in this evaluation. This kind of training is likely to be comparable to the humans we tested, who have general experience with faces, but do not have previous experience with the faces they were asked to match in this experiment. Moreover, we tested humans with an unfamiliar face matching task to ensure a fair comparison between machines and humans operating in situations typical for security applications, where face recognition for previously unfamiliar people is required. In the human performance experiments, individuals were asked to judge the similarity of 80 pairs of faces. To directly compare performance with face recognition algorithms, performance was computed for seven algorithms for the same 80 face pairs. This experimental design allowed for a direct comparison of humans and algorithms, and followed the design in O'Toole et al. [19]. The only difference is the method for selecting face image pairs.

Because humans can only rate a limited number of pairs of faces, 80 face pairs were selected from the approximately 10 million face pairs that the algorithms compared in the uncontrolled illumination experiments. To gain insight into the relative performance of humans and a set of algorithms, moderately difficult face pairs were selected for this experiment. A face pair is moderately difficult if approximately half of the algorithms performed correctly (e.g., if a face pair were images of the same person, then approximately half of the algorithms reported that the images were of the same person).

The sampling of face pairs was done as follows. All face pairs in the uncontrolled illumination experiment on the Sandia dataset, see Section IV-C, were given a difficulty score. The difficulty score was based on the number of algorithms that incorrectly estimated the match status of the face pairs at a FAR of 0.001. For face pairs of the same person, the difficulty score was the number of algorithms that failed to report the face pair as being the same person. Similarly for face pairs of different people, the difficulty score was the number of algorithms that failed to report the face pair as being different people. The difficulty score was computed based on the results of eight one-to-one algorithms. The easiest face pairs were assigned the minimum difficulty score of zero because all eight algorithms assigned the correct match status. The most difficult face pairs were assigned the maximum score of eight, because none of the algorithms assigned the face pair the correct match status. Moderately difficult face pairs with a difficulty score of between 3 and 5 were selected for this experiment. From these pairs, we selected 40 pairs of male and 40 pairs of female faces for the human performance experiments. Half of these pairs were match pairs (images of the same person) and half were non-match pairs (images of different people). Face pairs were presented side by side on the computer screen for two seconds. The presentation time of two seconds was chosen based on our previous study showing that human accuracy at this task was stable between 2 seconds and unlimited time [19]. After each pair of faces was presented, subjects rated the similarity of the two faces on a scale of 1 to 5. Subjects responded, using labeled keys on the keyboard as follows: 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. A total of 26 undergraduates at the University of Texas at Dallas participated in the experiment.

The results are as follows. On the FRVT 2006 human benchmark, Tsinghua (TS2-NORM) performed better than humans, and Viisage (V-NORM) and SAIT (ST-NORM) were comparable at all operating points. Figure 7 compares human and computer performance for the algorithms in Figure 6. Results in Figure 7 are reported on a receiver operating characteristic (ROC) to show the change in relative performance of humans and computers over a range of operating points. Human performance is reported at four operating points (the black dots in Figure 7). The lowest FAR of the four is 0.05. At a FAR of 0.05, six of seven algorithms have the same or better performance than humans. The FRVT 2006 human and machine experiments are in agreement with the results of O'Toole et al. [19] on "difficult" image pairs. Combined, the data suggest that for the uncontrolled illumination case, algorithm and human performance are comparable on unfamiliar faces.

V. COMPARISON OF BIOMETRIC MODALITIES

FRVT 2006 and ICE 2006 are the first technology evaluations that allowed iris recognition, still face recognition, and 3D face recognition performance to be compared. The comparison is performed on the multi-biometric dataset; to maintain consistency, still face and iris recognition are compared on one-to-one matching and still face and 3D face are compared on normalized

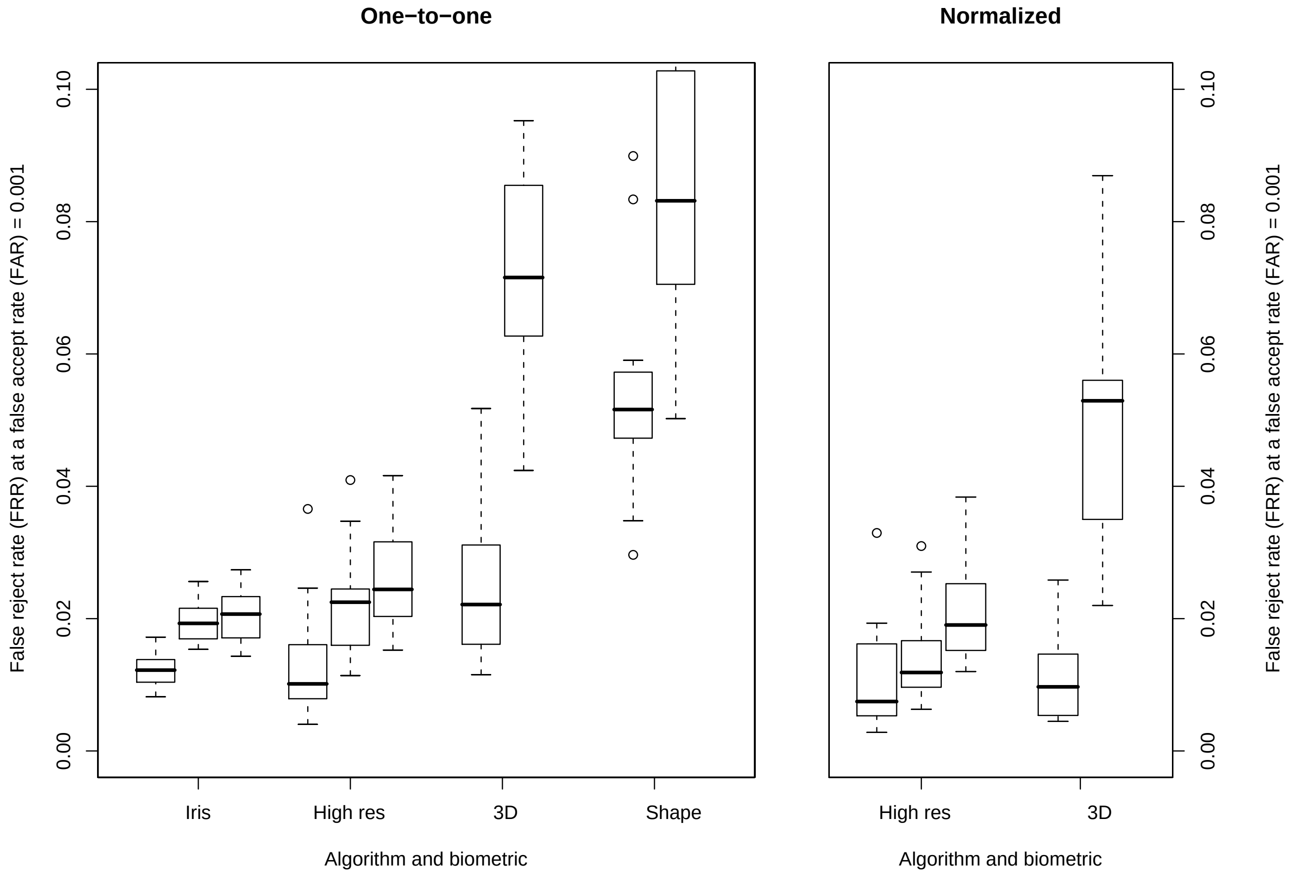

Fig. 8. A comparison of three biometrics on the Notre Dame multi-biometric dataset: iris, high-resolution still face, and 3D face. The left-hand panel reports performance for one-to-one algorithms and the right-hand panel reports performance for normalized algorithms. Each group on the horizontal axis corresponds to a biometric. For each biometric, the best two or three results are presented. The results for an algorithm are summarized on a boxplot. The false reject rate (FRR) at a false accept rate (FAR) of 0.001 is reported. The algorithms reported are Sagem-Iridian, Cambridge, and Iritech for iris; Neven Vision, Viisage, and Cognitec for still face; Viisage and Cognitec for 3D face; and Houston and Geometrix for shape. The right panel reports performance for normalized algorithms. In the right panel the algorithms reported are Neven Vision, Viisage, and Cognitec for still face; and Viisage and Cognitec for 3D face.

matching. Figure 8 compares the top performers on each of the three biometrics.

The multi-biometric dataset is an appropriate dataset for comparing performance across the different biometrics because the dataset controls for population, illumination, and time frame. In this dataset:

- Biometric samples were collected from the same population.
- Biometric samples were collected in the same laboratory during the same time period.
- The samples for all three biometrics were collected under controlled conditions appropriate for each of the modalities.
  - The iris sensor and 3D sensor have active illumination sources.
  - The still face images were collected under a constant controlled illumination source following the recommendations on the NIST mugshot best practices [20].

While the comparison among biometrics in the FRVT 2006 and ICE 2006 evaluation does control for the factors list above, there are other factors that are not controlled. These include maturity of the sensor technology, acquisition time for a biometric sample, cooperation required from a subject, and resolution of the sensor. In general, sensors for 3D biometric imaging of faces are less mature than cameras for iris and face imaging [16]. The 3D sensor used to collect data for the FRVT 2006 has a longer image acquisition time than the iris sensor or digital camera. The iris sensor requires a greater degree of user interaction and cooperation than the 3D sensor; and the 3D sensor requires a greater degree of user interaction and cooperation than the digital camera. Sensors for iris imaging and 3D imaging have fewer sample points than the number of pixels in a normal high-resolution camera image. For iris and 3D face, the sensor contains an active illumination source and for still face the data was collected under static controlled lighting. However, the sensors selected for the multi-biometric dataset collection were representative of the state-of-the-art commercial sensors available at the start of the collection effort. In terms of cost, the 3D sensor was the most expensive and the still camera was the cheapest.

To be consistent, we compared iris and still face recognition on only one-to-one matching because all the ICE 2006 submissions were one-to-one matching algorithms. The performance of the Sagem-Iridian (SI-2) iris algorithm with a FRR interquartile range of 0.011 to 0.014 at FAR of 0.001 and Neven Vision (NV1-1TO1) still face with a FRR interquartile range of 0.008 to 0.016 at a FAR of 0.001 are comparable. Figure 8 compares the top performers on each of the three biometrics.

To see if the relative performance of face and iris is stable across different false accept rates, we also examined the relative performance at a false accept rate of 0.0001 (one in ten thousand). Considering the number of subjects and biometric samples available, this is the limit of performance that can be measured for face recognition on the multi-modal dataset. At a false accept rate of 0.0001, the relative performance of the NevenVision face submission and the iris submissions is the same. The one-to-one Cognitec and one-to-one Viisage submissions are not comparable with the iris submissions. However, the performance of their normalization submissions is comparable to the one-to-one iris submissions.

We compared normalized still and 3D face recognition algorithms because performance with normalized face recognition algorithms was superior to the performance of one-to-one matchers. The performance of the Viisage (V-3D-N) 3D algorithm with a FRR interquartile range of 0.005 to 0.015 at FAR of 0.001 and Neven Vision (NV1-1TO1) still face with a FRR interquartile range of 0.006 to 0.015 at a FAR of 0.001 are comparable.

The results for the Viisage still and 3D submissions show the potential of fusing shape and texture information to improve performance over still imagery alone. For the Viisage still algorithm (V-NORM), the FRR interquartile range was 0.010 to 0.017 at a FAR of 0.001 on the Notre Dame high-resolution still face data. The Viisage (V-3D-N) 3D algorithm has a FRR interquartile range of 0.005 to 0.015 at FAR of 0.001, where the 3D consists of both shape and texture channels.

VI. Discussion and Conclusion

A. Iris recognition

The ICE 2006 established the first independent performance benchmark for iris recognition algorithms. The ICE 2006 performance is presented for algorithms from three groups: Sagem-Iridian, Iritech, and U. of Cambridge. The median FRR at a FAR of 0.001 for these algorithms is 0.012 for Sagem-Iridian, 0.019 for U. of Cambridge, and 0.021 for Iritech.

To better to understand the state-of-the-art in iris recognition, Newton and Phillips [21] performed a meta-analysis on the ICE 2006, the ITIRT, and the Iris 06 [7],[8]. While the ICE 2006 measured performance of algorithms on the same iris images, IRIRT and Iris 06 measured performance of one algorithm on data from different sensors. In the meta-analysis, to be able to compare performance across evaluations, performance statistics were selected that controlled for evaluation type, failure to enroll and failure to acquire, sensor quality software, and subject variability. Based on the selection criteria, across all three evaluations, reported FRR at a FAR of 0.001 ranged from 0.012 to 0.038. The lowest error rate observed was in the ICE 2006 for the Sagem-Iridian algorithm on data acquired on a LG EOU 2200; the highest error rate was observed in the ITIRT for an Iridian's KnoWho OEM SDK v3.0 on data acquired on a LG 3000. At an FAR of 0.001, the range of FRR for the best performers in each test was 0.012 to 0.015, with an average FRR of 0.014. Despite the differences in the testing protocols, sensors, image quality, subject variability and failures to enroll and acquire, the performance results from all three evaluations were comparable.

B. Controlled illumination still & 3D face recognition

The FRGC was a technology development effort that preceeded the FRVT 2006. The goal of the FRGC was to improve face recognition performance on frontal face images taken under controlled illumination by an order of magnitude over FRVT 2002. The baseline performance in FRVT 2002 was a FRR of 0.20 at a FAR of 0.001. Meeting the goal required that algorithms achieve a false reject rate (FRR) of 0.02 at a false accept rate (FAR) of 0.001 for matching frontal face images. This goal was exceeded on the FRVT 2006 Notre Dame dataset by four algorithms: Viisage, Cognitec, and two from Neven Vision. The median FRR at a FAR of 0.001 for these algorithms is 0.012 for Viisage, 0.019 for Cognitec, and 0.008 and 0.010 for Neven Vision. On the Dept. of State dataset the best median FRR at a FAR of 0.001 was 0.026. This performance was achieved by Toshiba on an algorithm designed to work on lower resolution facial images such as passport images.

Face recognition performance on still frontal images taken under controlled illumination has improved by at least a factor of 20 (greater than an order of magnitude) since the FRVT 2002. There are three primary components to the improvement in algorithm performance since the FRVT 2002: a) the recognition technology, b) higher resolution imagery, and c) improved quality due to greater consistency of lighting. Since performance was measured on the Dept. of State dataset in both the FRVT 2002 and the FRVT 2006, it is possible to estimate the improvement in performance due to algorithm design alone. The improvement in algorithm design resulted in an increase in performance by a factor of 7.7.

For the results on the Notre Dame and Sandia high-resolution datasets, the improvement in performance comes from a combination of algorithm design and image size and quality. Factors that influence quality include lighting, image compression, and ability to resolve details of the face. This is because new recognition techniques have been developed to take advantage of the larger high quality face images.

The performance on the Notre Dame high-resolution dataset shows one path for improving the performance of face recognition systems. The existence of the Notre Dame high-resolution dataset shows high quality data can be collected in large scale laboratory collection efforts. One of the challenges for the face recognition community is to develop acquisition techniques, protocols, and systems that allow for this quality of data to be collected in fielded applications.

The FRVT 2006 provides the first benchmarks of 3D face recognition technology. Performance on the 3D dataset met the FRGC goal of an order of magnitude improvement. The best performer was Viisage with a median FRR of 0.01 at a FAR of 0.001. The Viisage performance was achieved by processing both the texture and range channels in the 3D imagery. The U. of Houston achieved a median FRR of 0.052 at a FAR of 0.001 by processing on the range channel.

The Notre Dame multi-biometric component of ICE 2006 and FRVT 2006 allowed for comparisons among of iris, high-resolution still face, and 3D face recognition technology. On the ICE 2006 iris images and the Notre Dame high-resolution still frontal face images taken with controlled illumination, face and iris recognition performance on the one-to-one matching task is comparable. On the 3D images and Notre Dame high-resolution still frontal face images taken with controlled illumination, 3D and still frontal face recognition on the normalized matching task is comparable.

The images in the Dept. of State dataset were provided by the Visa Services Directorate, Bureau of Consular Affairs of the U.S. Department of State. Consequently, results on the Dept. of State dataset provide a performance benchmark for operational low-resolution highly compressed imagery. Performance on the Notre Dame and Sandia datasets provide an art-of-the-possible performance benchmark for acquisition systems that are specifically designed to maximize face recognition performance. Fingerprint, hand geometry, and iris sensors are designed specifically to capture biometric samples for recognition. Whereas, face capture systems have not been optimized for biometric recognition, but have been driven by the properties of commercial off-the-shelf cameras. One path for advancing face recognition is to design face recognition acquisition systems optimized for algorithm performance.

C. Uncontrolled illumination still & human face recognition

The ability of algorithms to recognize faces across illumination changes has improved significantly. The FRVT 2006 measured progress on this problem by matching images taken under uncontrolled illumination against images taken under controlled illumination. In January 2005, the three best self-reported results in the FRGC uncontrolled illumination experiments were FRRs of 0.24, 0.39, and 0.56 at a FAR of 0.001 [10]. In FRVT 2006, four algorithms, Cognitec, Neven Vision, SAIT, and Viisage had performance on both the Notre Dame and Sandia datasets that was better than the best FRGC results. On the Notre Dame dataset, SAIT had the best performance with a median FRR of 0.11 at a FAR of 0.001. On the Sandia dataset, Viisage had the best performance with a median FRR 0.13 at a FAR of 0.001. Thus, performance on sequestered uncontrolled images in FRVT 2006 was better than self-reported results in FRGC in January 2005.

The difference between the design of the controlled and uncontrolled illumination experiments in the FRVT 2006 was the probe images. In both experiments, the same set of controlled illumination images was used for the enrolled images. In the controlled experiments, the probe images were also taken under the same controlled light conditions; in the uncontrolled experiments, the probe images were taken under uncontrolled illumination conditions. The FRVT 2006 results show that relaxing the illumination condition has a dramatic effect on performance. For the controlled illumination experiment the best performance was a median FRR of 0.008 at a FAR of 0.001, whereas, for the uncontrolled illumination experiment the best performance had a median FRR of 0.11 at a FAR of 0.001. For the controlled illumination experiments, performance of the Notre Dame dataset was better than the Sandia dataset. By contrast, relaxing the illumination constraints on the probe images resulted in comparable performance on the Notre Dame and Sandia datasets.

The FRVT 2006, for the first time, integrated measuring human face recognition capability into an evaluation. The human visual system contains a very robust face recognition capability that is excellent at recognizing familiar faces [22]. However, human face recognition capabilities on unfamiliar faces falls far short of the capability for recognizing familiar faces. In FRVT 2006, performance of humans and computers was compared on the same set of images. The FRVT 2006 human and computer experiment measured the ability to recognize faces across illumination changes. This experiment found that on the Sandia dataset, algorithms are capable of human performance levels, and that at false accept rates in the range of 0.05, machines can out-perform humans.

D. Characterizing still face datasets

For still face recognition, the FRVT 2006 presents five sets of performance results. The results are from three data sets and two illumination conditions. One natural question is: how to characterize the difference between the five sets of performance results. One commonly proposed method is to report performance for a baseline algorithm for each condition. Following this approach we report recognition performance for a principal components analysis (PCA) based face recognition that was included on the FRGC distribution. The nearest neighbor classifier distance is the Malahanobis cosine distance, which is regarded as the current de facto standard for PCA-based algorithms [23]. The PCA algorithm was trained on images from the FRGC because these images were available to the FRVT 2006 participants.

Table V lists the FRR at a FAR = 0.001 for the FRVT 2006 still face experiments. To allow for a comparison with an establish dataset, we include performance on the FERET dataset on the Dup I probe set from the Sept 1996 evaluation [3]. Because the FERET dataset was taken with studio lighting it is categorized as a controlled illumination experiment. Because the original FERET images were used, this is categorized as low-resolution images. The baseline performance on the still face experiments falls into two categories. The first category consists of the controlled illumination experiments on the Notre Dame and Sandia dataset. In this category, the FRRs are 0.388 and 0.391 at a FAR = 0.001. The second category consists of the controlled illumination experiments on the Dept. of State and FERET datasets and the uncontrolled illumination experiments on the Notre Dame and Sandia dataset. In this category, the FRRs are 0.800, 0.870, 0.769, and 0.809 at a FAR = 0.001. At a coarse level, the PCA-baseline performance is able to categorize the FRVT 2006 high resolution datasets into controlled and uncontrolled illumination experiments. Also, for the FRVT 2006 controlled illumination experiments, baseline performance is able to categorize the datasets into high and low resolution.

The next step is to identify the factors in the imagery that account for the difference in performance among these experiments. This requires finding quantitative measures that characterize illumination and resolution. One step in this direction is Beveridge et al [24]. This study quantified the effect of image and subject factors on performance of the Notre Dame dataset. Factors included in the study are face size, a measure of focus, illumination environment, sex, and race. To be able to adequately understand the differences among datasets, the face recognition community needs to quantify and understand how the above factors effect algorithm performance.

E. Progress in frontal face recognition

The face recognition community has benefited from a series of U.S. Government funded technology development efforts and evaluation cycles, beginning with the FERET program in September 1993. One of the key contributions and legacies of these development efforts is the large data sets collected to support

TABLE V
Summarizes performance of the baseline PCA-based face recognition algorithm on the still face experiments. Performance is FRR at a FAR = 0.001.

Dataset	Illumination	Resolution	FRR @ FAR = 0.001
Notre Dame	controlled	high	0.388
Sandia	controlled	high	0.391
Dept. of State	controlled	low	0.800
Notre Dame	uncontrolled	high	0.769
Sandia	uncontrolled	high	0.809
FERET dup I	controlled	low	0.870

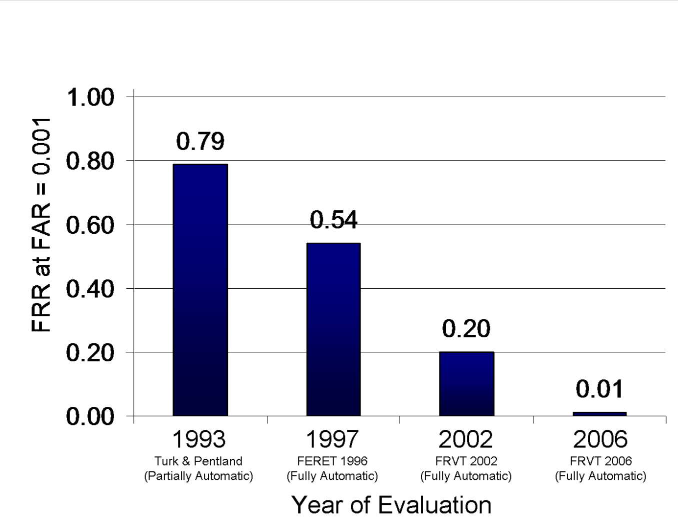

Fig. 9. The reduction in error rate for state-of-the-art face recognition algorithms as documented through the FERET, the FRVT 2002, and the FRVT 2006 evaluations.

these efforts. The large datasets have spurred the development of new algorithms. The independent evaluations have provided an unbiased assessment of the state-of-the-art in the technology and have identified the most promising approaches. In addition, the evaluations have documented two orders of magnitude improvement in performance from the start of the FERET program through the FRVT 2006.

Figure 9 quantifies this improvement at four key milestones. For each milestone, the false reject rate (FRR) at a false accept rate (FAR) of 0.001 (1 in 1000) is given for a representative state-of-the-art algorithm. The 1993 milestone is a retrospective implementation of Turk and Pentland's eigenface algorithm [25], which was partially automatic (it required that eye coordinates be provided). Performance is reported on the eigenface implementation of Moon and Phillips [26] with the FERET Sept96 protocol [3], in which images of a subject were taken on different days (dup I probe set). The 1997 milestone is for the Sept97 FERET evaluation, which was conducted at the conclusion of the FERET program. Performance is quoted on the U. of Southern California's fully automatic submission to the final FERET evaluation [27][28]. The 1993 and 1997 results are on the same test dataset and show improvement in algorithm technology under the FERET program. Technology improved from partially automatic to fully automatic algorithms, while error rate declined by approximately a third.

The 2002 benchmark is from the FRVT 2002. Verification performance is reported for the Cognitec, Eyematic, and Identix submissions on the Dept. of State facial image dataset. Because both the FERET and Dept. of State datasets are low-resolution and have similar performance on the baseline algorithm (see Table V), one can make the case that they are comparable and a significant portion of the decrease error rate was due to algorithm improvement.

The 2006 benchmark is from the FRVT 2006. Here, a FRR of 0.008 at a FAR of 0.001 was achieved by Neven Vision (NV1-norm algorithm) on the Notre Dame high-resolution controlled-illumination still images and Viisage (V-3D-n algorithm) achieved a FRR of 0.01 at a FAR of 0.001 on the 3D images. Both sets of images were from the Notre Dame multi-biometric dataset. Because of the difference between the 2002 and 2006 benchmark dataets, the improvement in algorithm performance between FRVT 2002 and FRVT 2006 is due to advancement in algorithm design, sensors, and understanding of the importance of correcting for varying illumination across images.

One key factor in the rapid reduction in the error rate over 13 years was the U.S Government sponsored evaluations and challenge problems. The FERET and the FRGC challenge problems focused the research community on large datasets and challenge problems designed to advanced face recognition technology. The FERET, the FRVT 2002 and the FRVT 2006 evaluations provided performance benchmarks, measured progress of, and assessed the state of the underlying technology with the goal of providing researchers with feedback on the efficacy of their approaches.

Appendix I
Performance Statistics

The FRVT 2006 and the ICE 2006 report verification performance. Verification models the situation were a person presents a biometric sample q_j to a system with a claimed identity. The system either accepts or rejects the claim. If t_i is the enrolled biometric sample of the person with the claimed identity, then the claim is accepted if the similarity score s_{ij} comparing the samples q_j and t_i is greater than or equal to a threshold t. The threshold t is the system's operating point. Two types of error can occur in this process: first a false accept in which an imposter claims an identity and is matched by the system above threshold; and secondly a false reject in which the system incorrectly matches the individual below threshold.

The Receiver Operating Characteristic (ROC) is computed to quantify verification performance. It shows the tradeoff between the two types of error by plotting estimates of the FRR against the FAR as a parametric function of an operating threshold, t. The FRR is the fraction of match similarity scores less than a threshold value t:

$$\mathrm{FRR}(t) = \frac{\left|\left\{s_{ij} < t, \quad \text{where } s_{ij} \in M\right\}\right|}{|M|}, \tag{1}$$

where M is the set of match similarity scores. The FAR is the fraction of non-match similarity scores greater than or equal to a threshold value t:

$$\mathrm{FAR}(t) = \frac{\left|\left\{s_{ij} \geq t, \quad \text{where } s_{ij} \in N\right\}\right|}{|N|}, \tag{2}$$

where N is the set of match similarity scores.

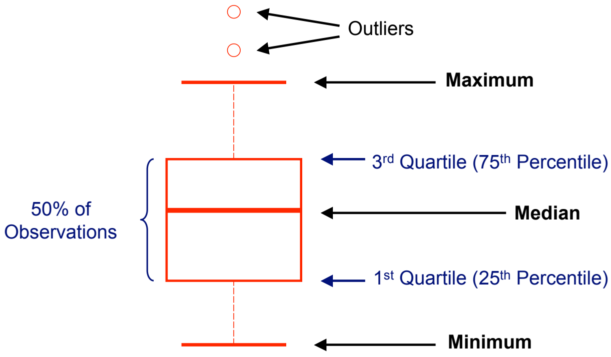
Outliers
Maximum
3rd Quartile (75th Percentile)
50% of Observations
Median
1st Quartile (25th Percentile)
Minimum

Fig. 10. An example of a boxplot with location of descriptive statistics labeled. The horizontal line through the middle of the box is the median of the performance range (50% of the observations are greater than the median and 50% are less than the median). The top and bottom of the box marks the 1*st* quartile (25*th* percentile) and 3*rd* quartile (75*th* percentile) values of the observations respectively. (At the 25*th* percentile point, 25% of the data has values less than this point.) Thus, 50% of the performance range is contained in the box. Above and below the box are vertical dashed lines, the "whiskers", that end with a short horizontal line. The ends of whiskers correspond to the minimum and maximum data value. The circles above or below the whiskers represent outliers. (To be technically accurate, the length of the whisker is the smaller of the maximum minus the 3*rd* quartile the (or the 1*st* quartile minus the minimum) and 1.5 times the vertical dimension of the box. Outliers are points whose values fall beyond the maximum extent of either whisker.)

A boxplot is a standard descriptive statistical technique for summarizing a dataset of scalar values [29]. The dataset is summarized by the minimum and maximum values, first and third quartiles, median, and outliners. Figure 10 shows a sample boxplot.

ACKNOWLEDGMENT

We acknowledge the support of Department of Homeland Security's Science and Technology Department and Transportation Security Administration (TSA), the Director of National Intelligence's Information Technology Innovation Center, the Federal Bureau of Investigation (FBI), the National Institute of Justice, and the Technical Support Working Group (TSWG). The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, SAIC, Schafer Corp., U. of Texas at Dallas or U. of Notre Dame.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, pp. 399–458, 2003.

[2] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing Journal*, vol. 16, no. 5, pp. 295–306, 1998.

[3] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. PAMI*, vol. 22, pp. 1090–1104, October 2000.

[4] L. Flom and A. Safir, "Iris recognition system," U.S. Patent 4,641,349, 1987.

[5] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn, "Image understanding for iris biometrics: a survey," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 281–307, 2008.

[6] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction to evaluating biometric systems," *Computer*, vol. 33, pp. 56–63, 2000.

[7] International Biometric Group, "Independent testing of iris recognition technology," International Biometric Group, Tech. Rep., May 2005. [Online]. Available: http://www.ibgweb.com/reports/public/ITIRT.html

[8] Authenti-Corp, "Iris recognition study 2006 (IRIS06)," Authenti-Corp, Tech. Rep. version 0.40, March 2007. [Online]. Available: http://www.authenti-corp.com/iris06/report/

[9] 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 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 947–954.

[10] P. J. Phillips, P. J. Flynn, W. T. Scruggs, K. W. Bowyer, and W. Worek, "Preliminary face recognition grand challenge results," in *Seventh International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 15–24.

[11] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone, "Face recognition vendor test 2002: Evaluation report," National Institute of Standards and Technology, Tech. Rep. NISTIR 6965, 2003, http://www.frvt.org.

[12] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Trans. PAMI*, vol. 29, pp. 531–543, 2007.

[13] G. Passalis, I. Kakadiaris, and T. Theoharis, "Intra-class retrieval of non-rigid 3D objects: Application to face recognition," *IEEE trans. PAMI*, vol. 29, no. 2, pp. 1 – 11, 2007.

[14] M. Husken, B. Brauckmann, S. Gehlen, and C. von der Malsburg, "Strategies and benefits of fusion of 2D and 3D face recognition," in *IEEE Workshop on Face Recognition Grand Challenge Experiments*, ser. Computer Society Digital Library, P. J. Phillips and K. W. Bowyer, Eds., 2005.

[15] P. J. Flynn and P. J. Phillips, "ICE mining: Quality and demographic investigation of ICE 2006 performance results," National Institute of Standards and Technology, Tech. Rep., 2008. [Online]. Available: http://iris.nist.gov

[16] K. W. Bowyer, K. Chang, and P. J. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[17] P. Grother, R. J. Micheals, and P. J. Phillips, "Face recognition vendor test 2002 performance metrics," in *Third Inter. Conf. on Audio- and video-based biometric person authentication*, J. Kittler and M. S. Nixon, Eds., vol. LNCS 2688. Springer, 2003, pp. 937–945.

[18] A. J. Mansfield and J. L. Wayman., "Best practices in testing and reporting performance of biometric devices. version 2.1," National Physical Laboratory, Tech. Rep., 2002. [Online]. Available: http://www.cesg.gov.uk/site/ast/biometrics/media/BestPractice.pdf

[19] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Pénard, and H. Abdi, "Face recognition algorithms surpass humans matching faces across changes in illumination," *IEEE Trans. PAMI*, vol. 29 1642-1646, pp. 1642–1646, 2007.

[20] R. M. McCabe, "Best practice recommendation for the capture of mugshots version 2.0," 1997, http://www.nist.gov/itl/div894/894.03/face/face.html.

[21] E. M. Newton and P. J. Phillips, "Meta-analysis of third-party evaluations of iris recognition," *IEEE trans. on SMC-A*, vol. 39, no. 1, pp. 4 – 11, 2009.

[22] P. J. B. Hancock, V. Bruce, and A. M. Burton, "Recognition of unfamiliar faces," *Trends in Cognitive Sciences*, vol. 4, pp. 330–337, 2000.

[23] J. R. Beveridge, D. Bolme, B. A. Draper, and M. Teixera, "The CSU face identification evaluation system," *Machine Vision and Applications*, vol. 16, no. 2, pp. 128–138, 2005.

[24] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui, "Focus on quality, predicting FRVT 2006 performance," in *Eighth International Conference on Automatic Face and Gesture Recognition*, 2008.

[25] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[26] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," *Perception*, vol. 30, pp. 303–321, 2001.

[27] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. PAMI*, vol. 17, no. 7, pp. 775–779, 1997.

[28] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg, "The Bochum/USC face recognition system," in *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, Eds. Berlin: Springer-Verlag, 1998, pp. 186–205.

[29] M. J. Crawley, *Statistical computing*. Wiley, 2002.