Preliminary Face Recognition Grand Challenge Results

P. Jonathon Phillips[1], Patrick J. Flynn[2], Todd Scruggs[3]
Kevin W. Bowyer[2], William Worek[3]

[1]National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899
[2]Computer Science & Engineering Depart., U. of Notre Dame, Notre Dame, IN 46556
[3]SAIC, 4001 N. Fairfax Dr., Arlington, VA 22203

Abstract

The goal of the Face Recognition Grand Challenge (FRGC) is to improve the performance of face recognition algorithms by an order of magnitude over the best results in Face Recognition Vendor Test (FRVT) 2002. The FRGC is designed to achieve this performance goal by presenting to researchers a six-experiment challenge problem along with a data corpus of 50,000 images. The data consists of 3D scans and high resolution still imagery taken under controlled and uncontrolled conditions. This paper presents preliminary results of the FRGC for all six experiments. The preliminary results indicate that significant progress has been made towards achieving the stated goals.

1. Introduction

In the past few years, a number of new face recognition techniques have been proposed. The new techniques include recognition from three-dimensional (3D) scans, recognition from high resolution still images, recognition from multiple still images, multi-modal face recognition, multi-algorithm, and preprocessing algorithms to correct for illumination and pose variations. These techniques hold the potential to improve performance of automatic face recognition significantly over the results in the Face Recognition Vendor Test (FRVT) 2002 [1].

The Face Recognition Grand Challenge (FRGC) is designed to achieve an order of magnitude increase in performance over the best results in FRVT 2002 by encouraging the development of algorithms for all of the above proposed methods. To facilitate the development of new algorithms, a data corpus consisting of 50,000 recordings divided into training and validation partitions was provided to researchers.

The starting point for measuring the increase in performance is the high computational intensity test (HCInt) of the FRVT 2002. The images in the HCInt corpus were taken indoors under controlled lighting. The performance point selected as the reference is a verification rate of 80% (error rate of 20%) at a false accept rate (FAR) of 0.1%. This is the performance level of the top three FRVT 2002 participants. An order of magnitude increase in performance is therefore defined as a verification rate of 98% (2% error rate) at the same fixed FAR of 0.1%.

Participants in FRGC submitted a set of raw similarity scores to the FRGC organizers on 14 January 2005. This paper provides a summary of performance from these submitted scores. A more detailed description of the FRGC challenge problem, data, and experiments is given in Phillips et al [2].

2. FRGCv2 Data

Data for the FRGC was collected at the University of Notre Dame. The FRGC data corpus is part of an ongoing multi-modal biometric data collection.

A *subject session* is the set of all images of a person taken each time a person's biometric data is collected. The FRGC data for a subject session consists of four controlled still images, two uncontrolled still images, and one three-dimensional image. The controlled images were taken in a studio setting, are full frontal facial images taken under two lighting conditions (two or three studio lights) and with two facial expressions (smiling and neutral). The uncontrolled images were taken in varying illumination conditions; e.g., hallways, atria, or outdoors. Each set of uncontrolled images contains two expressions, smiling and neutral. The 3D images were taken under controlled illumination conditions appropriate for the Vivid 900/910 sen-

Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)
0-7695-2503-2/06 $20.00 © 2006 IEEE

sor. In the FRGC, 3D images consist of both range and texture channels. The still images were taken with a 4 megapixel Canon PowerShot G2[1]. Figure 1 shows all the images from one subject session.

The data for the FRGC experiments was divided into training and validation partitions. The data in the training partition was collected in the 2002-2003 academic year. From the training partition, two training sets were distributed. The first is the *large still training set*, which consists of 12,776 images from 222 subjects, with 6,388 controlled still images and 6,388 uncontrolled still images. Images in the validation partition were collected during the 2003-2004 academic year. The validation set contains images from 466 subjects collected in 4,007 subject sessions.

3. Description of Experiments

The experiments in FRGC ver2.0 are designed to advance face recognition in general with emphasis on 3D and high resolution still imagery. Ver2.0 consists of six experiments. Table 1 gives the size of each experiment in terms of target and query set, and number of similarity scores.

Experiment 1 measures performance on the classic face recognition problem, namely the recognition from frontal facial images taken under controlled illumination. To encourage the development of algorithms that exploit potential additional information in high resolution images, all controlled still images are high resolution. In experiment 1, the biometric samples in the target and query sets consist of a single controlled still image.

Experiment 2 is designed to the examine the effect of multiple still images on performance. In this experiment, each biometric sample consists of the four controlled images of a person taken in a subject session. The biometric samples in the target and query sets are composed of the four controlled images of each person from a subject. Experiment 4 is designed to measure progress on recognition from uncontrolled frontal still images. In experiment 4, the target set consists of single controlled still images, and the query set consists of a single uncontrolled still image.

Experiments 3, 5, and 6 look at different potential implementations of 3D face recognition. Experiment 3 measures performance when both the enrolled and query images are 3D. In experiment 3, the target and query sets consist of 3D facial images (both the shape and texture channels). Experiment 3s is the same as Experiment 3 except the data consists solely of the shape channel. Similarity, Experiment 3t consists of the Experiment 3 data restricted to the texture channel. Experiments 3, 3s, and 3t allow for an assessment of the contribution of the shape and textures to the performance of 3D facial imagery.

[1]The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, the University of Notre Dame, or SAIC.

Table 1. Size of ver2.0 experiments. For each experiment the size of the target and query set is given. The number of similarity scores in each experiment's similarity matrix is provided.

Exp.	Target set size	Query set size	No. sim scores (million)
1	16,028	16,028	257
2	4,007	4,007	16
3	4,007	4,007	16
3s	4,007	4,007	16
3t	4,007	4,007	16
4	16,028	8,014	128
5	4,007	16,028	64
6	4,007	8,014	32

One potential scenario for 3D face recognition is that the enrolled images are 3D and the target images are still 2D images. Experiments 5 looks at this scenario when the query images are controlled and experiment 6 looks at the case when the query images are uncontrolled. In both experiments, the target set consists of 3D images.

4. FRGCv2 Protocol

The complete FRGCv2 data and challenge problem were made available to participants on 27 September 2004. The data consisted of both the training and validation partitions for all six experiments. For a FRGC participant's results to be include in the initial analysis in this paper, complete similarity matrices needed to be submitted to the first author by 14 January 2005. The initial analysis was presented at the Third Face Recognition Grand Challenge Organization Workshop held on 16 February 2005, To be included in the analysis, participants were required to submit complete similarity matrices. Participants could submit results for any subset of the six experiments, and the results could be either fully automatic or partially automatic algorithms. Participants could submit results for multiple algorithms for an experiment. In order to provide the face recognition research community with an unbiased assessment of the performance of algorithms participating in FRGC, results of the analysis in this paper are not

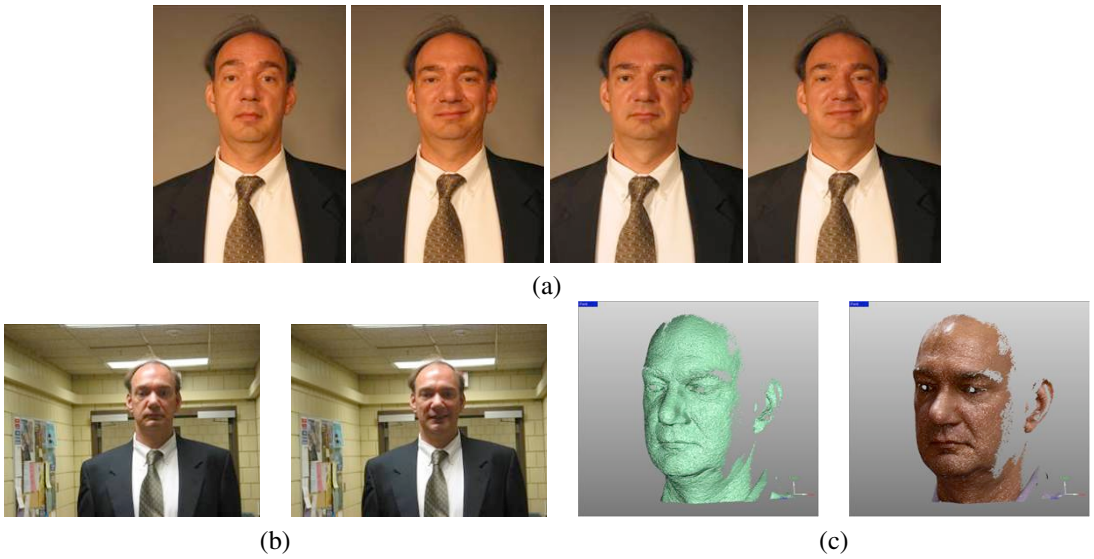

Figure 1. Images from one subject session. (a) Four controlled stills, (b) two uncontrolled stills, and (c) 3D shape channel and texture channel pasted on 3D shape channel.

labeled by participating group.

5. Summary of FRGCv2 Results–January 2005

By the 14 January 2005 deadline, 63 similarity matrices were received from 19 groups. The 19 groups consisted of 10 companies and 9 universities from 6 countries. Table 2 reports the number of similarity matrices analyzed for each experiment. Figure 2 summarizes performance for each experiment as a bar graph. Performance is summarized by the verification rate at a FAR of 0.001, the vertical axis. For each experiment, three statistics are reported. The first is the performance of the baseline algorithm (blue or left bar). The best performance over all submitted similarity matrices for an experiment is reported on the orange (right) bar. The green (center) bar reports the median performance over submitted results for each experiment. For Experiments 5 and 6, a baseline algorithm was not provided and only one result was submitted, which is reported.

The still images had only two expressions, neutral and smile. The 3D images had a variety of expressions. Figure 3 breaks out Experiment 3 performance by effect of expression. For the expression analysis, 3D scans are divided into two categories, neutral and non-neutral expression. We break out performance for neutral versus neutral expressions, and neutral versus non-neutral expressions. Figure 3 breaks out performance for all ten Experiment 3 algorithms. For each algorithm performance is reported for all 3D images, neutral versus neutral, and neutral versus non-neutral. Discussion of these results is postponed to the Conclusion.

Table 2. Number of results submitted for each experiment.

Experiment No.	No. results
1	17
2	11
3	10
3t	4
3s	5
4	12
5	1
6	1

6. Analysis of Results

The size and structure of the FRGC corpus allows researchers to investigate questions on a scale not previously examined. The first novel structure of the FRGC corpus is the large number of repeated acquisition from each person. The large number of images per person makes it possible to investigate the variation in recognizability of a population, e.g., how much harder are some people to recognize than others. The second novel feature of the FRGC corpus is the large training set of 12,776 images. The large training set makes it possible to examine the effect of training set size on algorithm

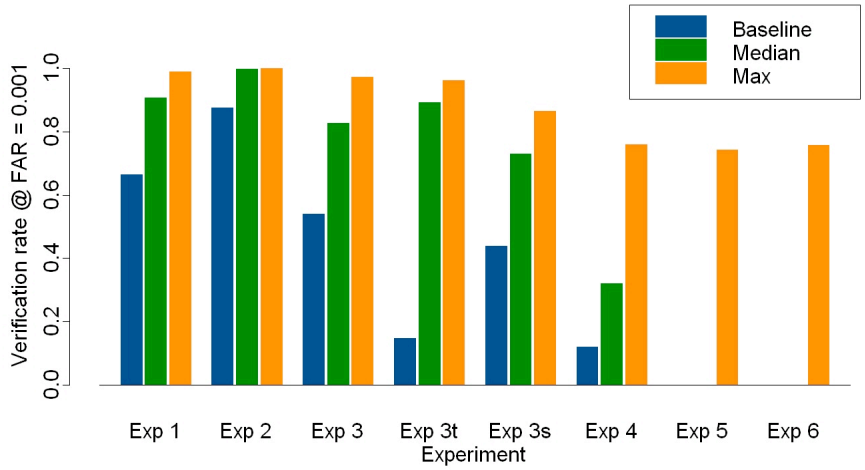

Figure 2. Summary of performance for Experiments 1, 2, 3, 4, 5, and 6

performance. We do not provide answers to these question, rather we show they are interesting questions that merit further attention.

One of the open questions in face recognition and biometrics in general is: are some people harder to recognize than others? In the context of speaker recognition, Doddington et al. [3] first looked at this question.

The first is to compute an estimate of recognizibility that allows for a comparison among algorithms.To allow for comparisons between algorithms, all the match scores are normalized to have mean zero and standard deviation one. A *match score* is a similarity score between two images of the same face. From the normalized match scores, the mean (μ_j^a)and standard deviation (σ_j^a) of the match scores for each person j were computed for each algorithm a. The recognizability of person j is estimated by the mean match score μ_j^a for that person.

Figure 4 contains scatter plots of mean versus standard deviation for four algorithms. Figure 4(a) and 4(b) are for Experiment 1 and 4(c) and 4(d) are from Experiment 3s. Each point (μ_j^a, σ_j^a) in a scatter plot is the mean μ_j^a and standard deviation σ_j^a for the match scores for person j. Experiment 1 was chosen because the images were taken with the same controlled illumination and so variation in match scores due to changes in illumination would be minimized. Experiment 3s (3D shape channel) was selected to look at similarity scores for a different mode. To avoid the situation were a difference in scatter plots reflects different levels of performance, better performing algorithms were selected. For the all four algorithms in Figure 4, the similarity scores were distance measures, e.g., a smaller similarity scores means that two faces are more a like. Figure 4 clearly shows that some variation in the mean match score implies that some people people are harder and some are easier to recognize.

It is not understood what causes recognition rates to vary among members of a population. In our analysis we will look at two competing hypotheses. The first cause is that some people are intrinsically harder or easier to recognize. If this were the main cause of variability, then the standard deviation should be independent of the mean. After accounting for outliers, this seems to be the case for Algorithms B and D in Figure 4(b) and Figure 4(d). A second possible explanation is that the primary source of different recognizability is that some people's facial images naturally vary more than others. At one end of the spectrum are the "stoics," whose images are very similar regardless of expression priming or underlying emotional state. At the other end of the spectrum are the "expressives," whose images display high variability. If a person's natural variability was the primary explanation for a person recognizability, then the standard deviation of a person's match score should increase with that person's mean match score. Figure 4(a) and Figure 4(c) support the hypothesis that a person's variability is the main explanation for varying recognizability among subjects.

Unfortunately, the scatter plots in Figure 4 provide evidence to support both competing hypotheses. Both competing hypotheses are supported in both texture and shape modes. Since all four algorithms were better performers in the experiments, the removes an obvious potential explanation that shape of the scatter plots is predicted by performance.

The previous analysis examined recognizability of people for a single algorithm. The next step is to exam-

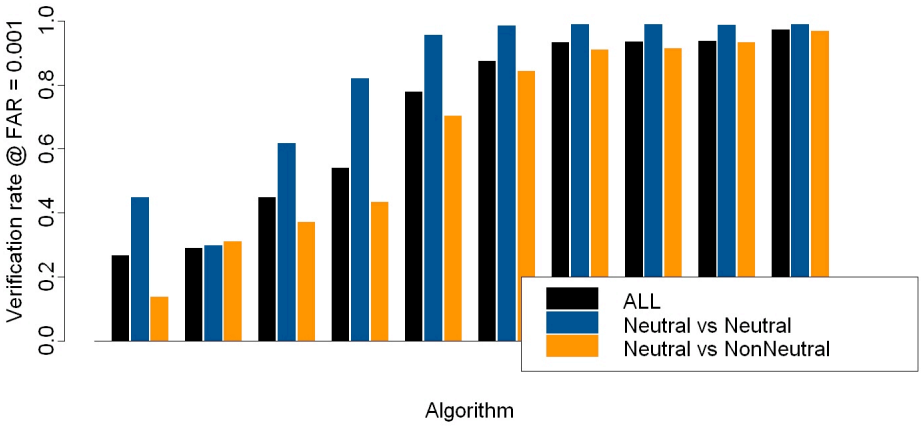

Figure 3. Effect of expression on 3D face recognition performance.

ine recognizability across algorithms and modes, e.g., is a person easy or hard for different algorithms? This analysis is performed by plotting the subject match score means of two experiments on a scatter plot. Figure 5(a) is a scatter plot that compares the mean similarity score of subjects of Algorithms A and B on Experiment 1. The x-axis for all three scatter plots in Figure 5(a) is the mean score for Algorithm A on experiment 1, and allows for an easier comparison across all three panels. Figure 5(b) shows the correlation between Algorithm A on Experiments 1 and 3t (texture channel of the 3D image). A comparison between Figures 5(a) and (b) shows the correlation between algorithms on the same data is tighter than between the same algorithm on two different datasets. Because of the structure of the FRGC image collection protocol, the corresponding images in Experiments 1 and 3t were taken within five minutes. Figure 5(c) shows the correlation across modes by comparing Experiments 1 and 3s (shape channel of the 3D image). The results in Figure 5 show that there is greater correlation in subject recognizability between two different algorithms on the same data set than on the same algorithm on two different data sets.

The majority of face recognition algorithms have a learning component to them. However, the effect of training set size on performance has not been well studied. Prior to FRGC, the largest standard training set in the literature is 501 images in the FERET Sep96 protocol.

Phillips et al [2] looked at the effect of the training set size on the FRGC PCA-based baseline algorithm. Figure 6 reports performance on Experiment 1 for training sets of size 512, 1,024, 2,048, 4,096, and 8,192. Verification performance at a FAR of 0.1% is reported (vertical axis). The horizontal axis is the number of eigenfeatures in the representation. The eigenfeatures selected are those with the largest n variances as estimated from the training set. The training set of size 512 approximates the size of the training set in the FERET Sep96 protocol. This curve approximates what was observed in Moon and Phillips, where performance increases, peaks, and then decreases slightly [4]. Performance peaks for training sets of size 2,048 and 4,096 and then starts to decrease for the training set of size 8,192. For training sets of size 2,048 and 4,096, there is a large region where performance is stable. The training sets of size 2,048, 4,096, and 8,192 have tails where performance degrades to near zero.

Liu [5] looked at the effect of training set size on Experiments 1 and 3. His results are summarized in Table 3. The results show a steady increase overall three training sets and that performance has not saturated as a function of training set size.

Table 3. Effect of training set size for Liu [5]. Results for the verification rate at a FAR of 0.001. Results are reported for both Experiment 1 and 4.

Training set size	Exp. 1	Exp. 4
1623	0.87	0.64
3194	0.90	0.70
6388	0.92	0.76

There are three immediate observations from these two examples: first, the size of the training set is important; second, training on large sets of data can improve performance; and third, the performance of the PCA-

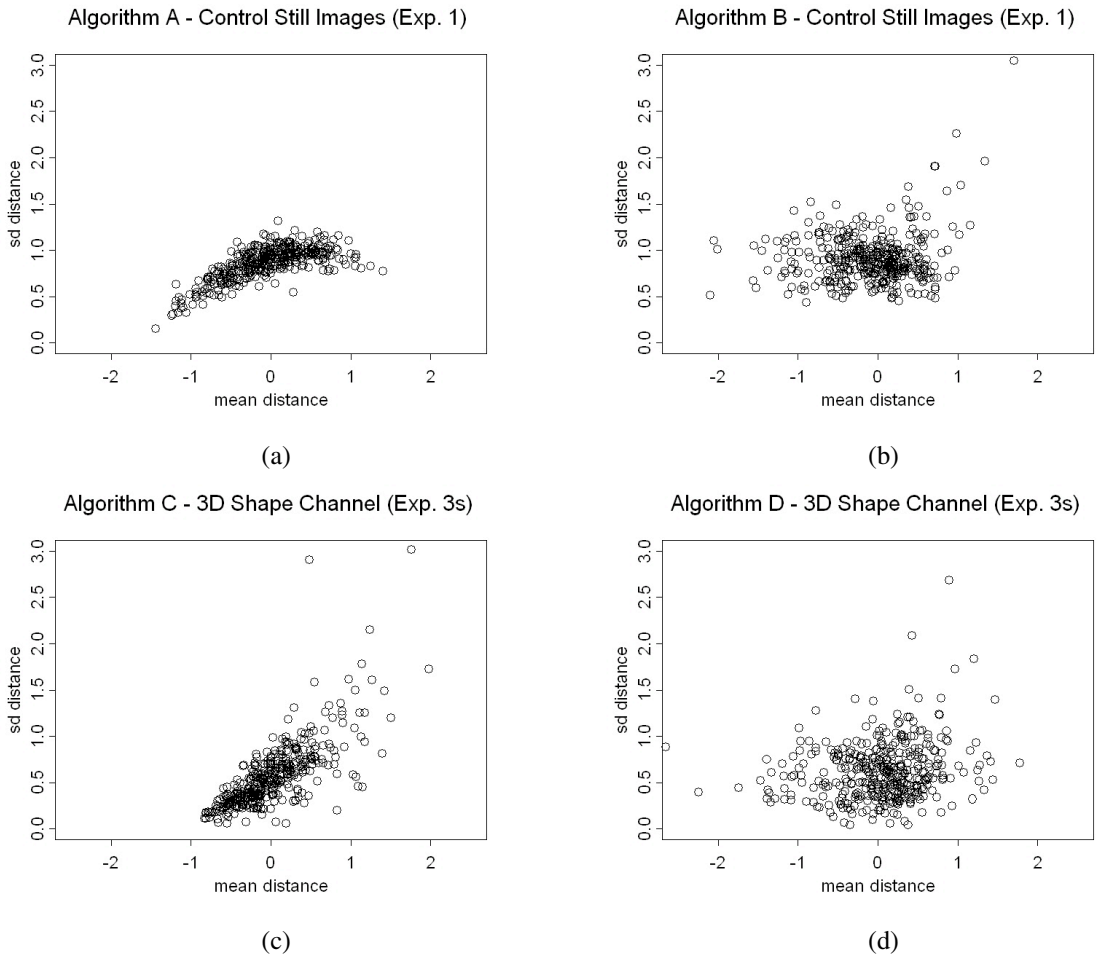

Figure 4. Scatter plots of subject mean versus standard deviation. (a) and (b) plot Algorithms A and B on Experiment 1, and (c) and (d) plot Algorithms C and D on Experiment 3s (shape channel).

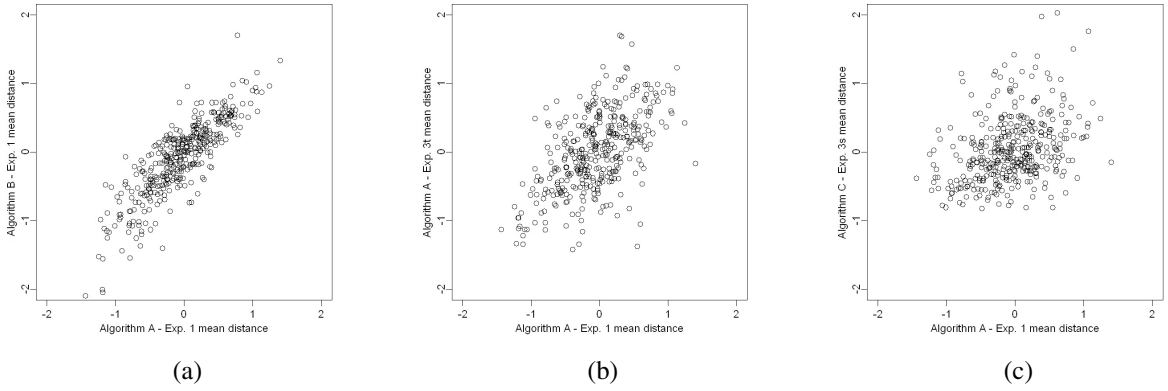

Figure 5. Comparison of recognizability between (a) different algorithms on the same data, (b) same algorithm on different data sets, and (c) different algorithms on different modes (still versus 3D shape).

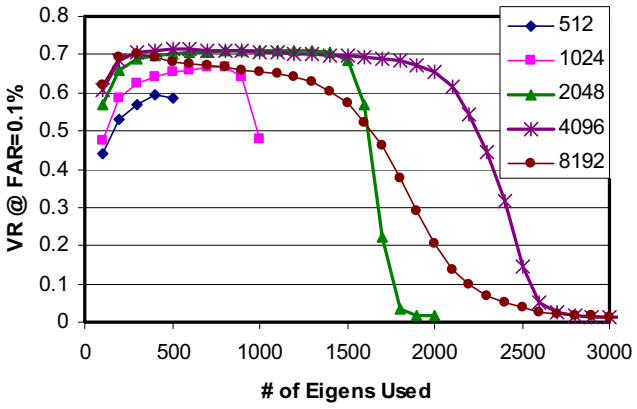

Figure 6. Performance as a function of the number of eigenfeatures for different size training sets. Verification performance at a FAR of 0.1% is reported. The numbers in the legend are the size of the training set.

based algorithm on a training set of size 512 is a warning about drawing conclusions from too small a sample.

7. Conclusions

The primary goal of the FRGC is to encourage the development of face recognition algorithms that have performance an order of magnitude better than observed in FRVT 2002. The specific target performance is a verification rate of 98% at a FAR of 0.1% as measured in the Face Recognition Vendor Test (FRVT) 2006[2] [6].

FRGC pursues three possible avenues for meeting its performance goals: high resolution still imagery, multiple still images, and 3D facial imagery.

The maximum score for Experiment 1 was 99% and a median of 91%. The comparable scores for Experiment 2 are 99.9% and 99.9%. Since FRGC is a challenge problem and the results are based on raw similarity scores submitted by participating groups, these results are not conclusive that the performance goals of FRGC have been meet. However, they do provide evidence that the goals are likely to be met. The difference in performance between the results for Experiments 1 and 2, especially for median score, indicate that having multiple still images of a person has the potential to increase performance.

FRGC is the first challenge problem with a large set of 3D facial imagery. The maximum score for of 97% for Experiment 3 shows the potential of 3D facial imagery. The results in this paper for Experiment 3 are three months after the first release of a large 3D data set. By comparison, the results on still images are based on over a decade of intensive research after the first large still image datasets were released.

The impact of the size of the training set has important implications for face recognition algorithm development. The first is that large data sets need to be collected and assembled for training algorithms. Second, researchers will need to develop methods for training face recognition on very large sets. For example, Liu's method requires computing eigenvectors from a matrix of the size of the number of training images.

Not only has FRGC spurred the development of new face recognition algorithms and techniques, but it has allowed for investigation of scientific questions that could not of previously been addressed. As of 30 January 2006, 138 groups have been given access to FRGC. The algorithm development portion of FRGC has lead to a new generation of algorithms [7]. Research of the new scientific questions will provide a deeper understanding of the principles of face recognition.

[2]The start date for FRVT 2006 was 30 January 2006. FRVT 2006 was open to academia, research institutions, and companies.

References

[1] 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", Tech. Rep. NISTIR 6965, National Institute of Standards and Technology, 2003, http://www.frvt.org.

[2] 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.

[3] G. Doddington, W. Ligget, A. Martin, M. Przybocki, and D. Reynolds, "Sheeps, goats, lambs, and wolves: A statistical analysis of speaker performance in the NIST 1998 recognition evaluation", in *Proceedings ICSLP '98*, 1998.

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

[5] C. Liu, "Capitalize on dimensionality increasing techniques for improving face recognition performance", *IEEE Trans. PAMI*, (in press 2006).

[6] "Face recognition vendor test 2006", http://face.nist.gov/.

[7] P. J. Phillips and K. W. Bowyer, Eds., *Proceedings of The IEEE Workshop on Face Recognition Grand Challenge Experiments*, 2005.