Double Trouble: Differentiating Identical Twins by Face Recognition

Jeffrey R. Paone, Patrick J. Flynn, *Fellow, IEEE*, P. Jonathon Philips, *Fellow, IEEE*,
Kevin W. Bowyer, *Fellow, IEEE*, Richard W. Vorder Bruegge, Patrick J. Grother,
George W. Quinn, Matthew T. Pruitt, and Jason M. Grant

***Abstract*—Facial recognition algorithms should be able to operate even when similar-looking individuals are encountered, or even in the extreme case of identical twins. An experimental data set comprised of 17486 images from 126 pairs of identical twins (252 subjects) collected on the same day and 6864 images from 120 pairs of identical twins (240 subjects) with images taken a year later was used to measure the performance on seven different face recognition algorithms. Performance is reported for variations in illumination, expression, gender, and age for both the same day and cross-year image sets. Regardless of the conditions of image acquisition, distinguishing identical twins are significantly harder than distinguishing subjects who are not identical twins for all algorithms.**

***Index Terms*—Face and gesture recognition.**

I. Introduction

BIOMETRICS and facial recognition are based on the assumption that every individual has a unique identity that is distinguishable from that of others. Algorithms are designed to differentiate an image of one person from an image of another person or to confirm if the two images are of the same person. Identical twins present a challenging scenario since their facial features are very similar.

The primary focus of this paper is to assess the performance of current face recognition algorithms on a dataset containing face images of identical twins. While other modalities may be used to differentiate between identical twins; such as fingerprint or iris; face recognition is non-obtrusive, may be acquired from a distance, and does not require a fully cooperative subject.

The use of face recognition in forensic applications is becoming more and more common, especially because when other biometric modalities may not be available. Law enforcement and security agencies around the world are using face recognition to detect fraud and to identify unknown individuals depicted in the act of committing crimes, even when fingerprints or DNA may not be left behind. Similarly civil programs, such as driving licensing and passport issuance, use face recognition to detect duplicate applicants because the face has long had social acceptance in identity credentials and because capture equipment is so widely available. When utilizing such biometric tools, however, it is important that mis-identifications be avoided to minimize - or eliminate - the chance of inadvertently implicating an innocent person. The problem of mis-identification of twins with existing algorithms is so great, in fact, that some agencies issuing driver licenses have implemented special procedures to flag potential matches against twins.

Further, identical twins represent the worst case scenario for face recognition where two separate subjects have a very similar appearance. Subjects may have very similar appearance if one subject is trying to pose as another subject. It is important to test existing face algorithms on the hardest recognition cases. If the algorithms can perform sufficiently well on the hardest problems, then they will be able to solve the simpler problems as well.

In this paper, seven different anonymous face recognition algorithms are tested in various conditions. Performance is measured with respect to four covariates: (i) illumination, (ii) expression, (iii) gender, and (iv) age. There were two acquisition sessions that took place one year apart, under similar conditions. The effect of the four covariates can then be applied to images taken on the same day and one year apart to measure the effect of elapsed time on the recognition of twins. Our results have shown that distinguishing identical twins is a challenging problem and current face recognition algorithms have great difficulty in accurately differentiating between a pair of identical twins. As expected, images of twins taken one year apart in any condition have the worst

Manuscript received May 24, 2013; revised August 29, 2013 and November 19, 2013; accepted December 3, 2013. Date of publication January 2, 2014; date of current version January 14, 2014. This work was supported in part by the Federal Bureau of Investigation and Technical Support Working Group through Army Research Laboratory under Grant W91CRB-08-C-0093, in part by the Biometric Center of Excellence through the West Virginia University, in part by the University of Notre Dame through the Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity, and in part by the ARL Cooperative Agreement under Grant W911NF-10- 2-0067. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of FBI, TSWG, IARPA, the ODNI, the Army Research Laboratory, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. The identification of any commercial product or trade name does not imply endorsement or recommendation by the authors or the Federal Bureau of Investigation, National Institute of Standards and Technology, or the University of Notre Dame. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kar-Ann Toh.

J. R. Paone, P. J. Flynn, K. W. Bowyer, M. T. Pruitt, and J. M. Grant are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: jpaone@nd.edu; flynn@nd.edu; kwb@nd.edu; mpruitt@nd.edu; jgrant3@nd.edu).

P. J. Philips, P. J. Grother, and G. W. Quinn are with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: jonathon@nist.gov; patrick.grother@nist.gov; gw@nist.gov).

R. W. V. Bruegge is with the Federal Bureau of Investigation, Science and Technology Branch, Digital Evidence Laboratory, Quantico, VA 22135 USA (e-mail: richard.vorderbruegge@ic.fbi.gov).

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

Digital Object Identifier 10.1109/TIFS.2013.2296373

performance. However, even when images of twins are taken on the same day minutes apart, the recognition performance is signficantly worse than the baseline scenario when twins are not present. Current face recognition algorithms do not perform well enough on the hardest problems and additional improvement is needed before algorithms can handle the toughest problems.

This paper is organized as follows. Section II discusses the related work in biometric recognition between twins. Section III relates a population without twins to a population of only twins. Section IV outlines the datasets and algorithms used, and describes how performance will be analyzed. Section V presents the experiments and results for images acquired on the same day. Section VI presents the experiments and results for images taken a year apart. Last, section VII discusses the results and presents concluding remarks.

II. Background

There is particular interest in using biometrics to distinguish identical twins. While several algorithms and recognition systems capable of differentiating between a single set of twin siblings have been introduced [3], [7], [9], [15], the only other significant study of biometric recognition of twins is by Sun et al. [12]. They conducted matching experiments using the face, iris, and fingerprint modes as well as a fusion of these modes. The dataset contained images of 134 subjects (64 pairs of twins and two sets of triplets) collected at the Annual Festival of Beijing Twins Day. They determined it was easier to distinguish identical twins using iris or fingerprint biometrics than using face biometrics. They also concluded for face biometrics the identical twin impostor distribution (i.e. the set of scores for a pair of images of identical twin siblings) was more similar to the match distribution than a general impostor distribution (i.e. the set of scores for pairs of images not containing any identical twin siblings). There has also been work on distinguishing identical twins based on other biometrics including palmprint [8], fingerprint [6], iris [5], and speech recognition [1].

The Twins Days dataset used here was first introduced by Phillips et al. [10], also later discussed by Pruitt et al. [11], and is available online [13]. In introducing the Twins Days dataset, Phillips et al. examined a set of covariates and studied the effects of illumination and expression on differentiating identical twins. They also compared the identical twin impostor distribution to the general impostor distribution. Pruitt et al. later applied a set of algorithms to the dataset and studied the performance of each algorithm. This paper expands the prior work of Phillips et al. and Pruitt et al. by considering all values of a covariate across several algorithms. The effect of elapsed time between acquisitions on each covariate is also examined.

III. Twins Compared to Non-Twins

As originally noted by Sun et al. [12] and confirmed by Phillips et al. [10], differentiating between identical twins is harder than differentiating within a general population due to the overlap in the match and identical twin impostor distributions. The general impostor distribution has little overlap with the match distribution, while the identical twin impostor distribution has significant overlap with the match distribution. An example comparing a general impostor distribution against an identical twins impostor distribution can be seen in Fig. 1. In this figure, a similarity score of 1.0 is a perfect match. It is clear from these distributions that the identical twins impostor distribution has significantly more non-match pairs with a higher similarity score than the general impostor distribution. These non-match pairs with a high similarity score will overlap with the match pairs and make identification difficult for a face recognition system.

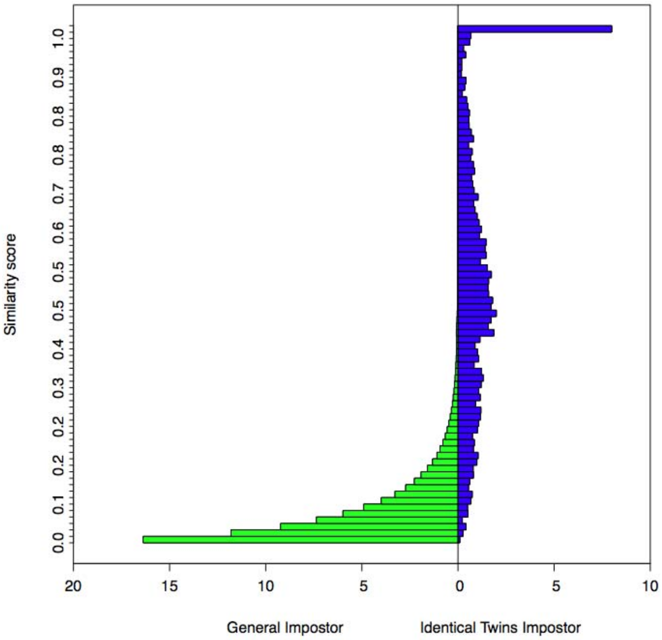


Fig. 1. Distribution of nonmatch scores for a population with no twins on the left and for a population of only identical twins on the right.

In order to distinguish between a pair of twins, the two subjects from a pair of twins will be labeled as Twin A and Twin B.

IV. Methods and Materials

A. Data

The dataset was collected at the Twins Days festival [14] in Twinsburg, Ohio in August 2009 and August 2010. The Twins Days festival is a weekend-long event that draws between 1,500 and 2,000 twin sibling pairs. Attendees range in age from infants to elderly and represent a variety of ethnic groups and races. All subjects in the dataset are over the age of eighteen and Caucasians are the largest single racial group. All participants in the acquisition sessions were self-identified identical twins. Example images can be seen in Fig. 3 where Fig. 3(a) and (b) is Twin A and Fig. 3(a) and (d) is Twin B.

The 2009 acquisition setup included four cameras as shown in Fig. 2(a). The subjects were instructed to look at the front camera, then rotate to look at the side camera. Before rotating, the subject had a neutral expression; after facing the other camera, they would express a smile. Both cameras would acquire images at the same time from different angles causing differences in the sensor and lighting on the face. The same

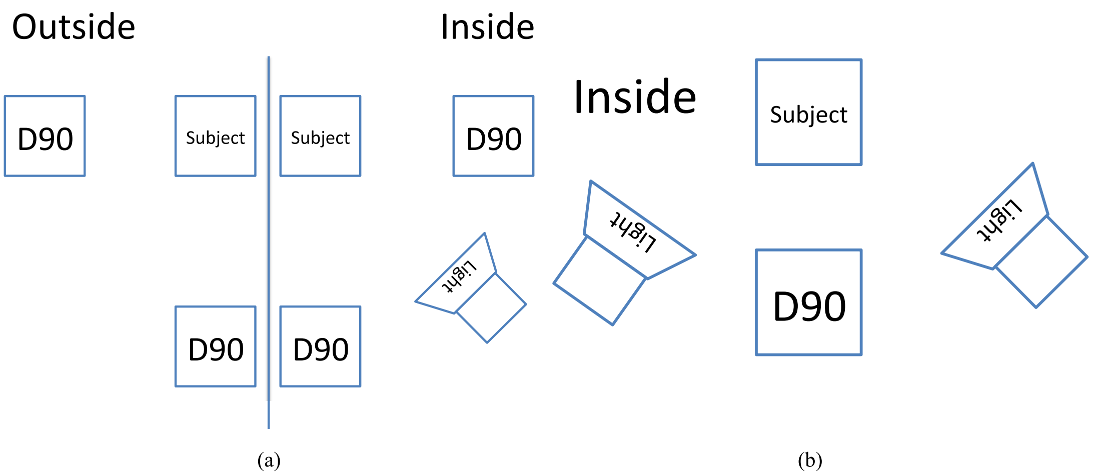


Fig. 2. Acquisition setups for 2009 and 2010. (a) 2009 Setup. Frontal images were taken both indoors and outdoors using all four cameras. (b) 2010 Setup. Frontal images were taken only indoors using a single camera.

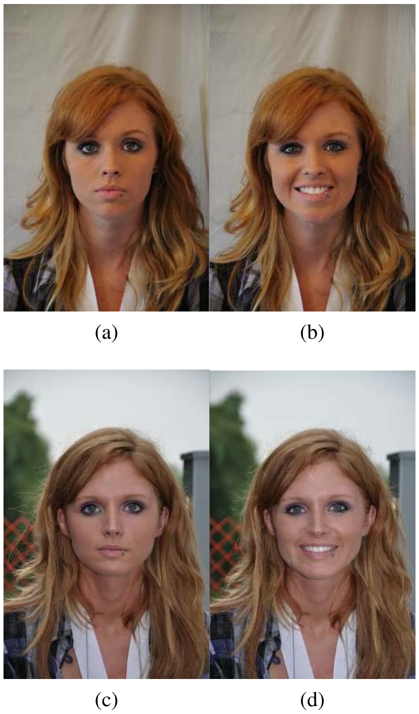


Fig. 3. Example frontal images. Images (a) and (b) are of Twin A while images (c) and (d) are of Twin B.

setup was used both in a tent with controlled lighting and outside under ambient light. The 2009 collection acquired 17,486 face stills from 252 subjects (126 pairs of identical twins).

The 2010 acquisition setup was a single camera positioned only indoors as shown in Fig. 2(b). An initial image with a neutral expression was taken then the subject was instructed to stand up, turn around, and sit back down. A second photo was then taken with a smiling expression. The 2010 collection acquired 6,863 face stills from 240 subjects (120 pairs of identical twins). There were 48 subjects (24 pairs of identical twins) who participated in both the 2009 and 2010 acquisitions.

From the complete dataset, only the subjects without glasses in a front-facing pose were used in these experiments.

B. Algorithms

Seven algorithms will be compared. Three of the algorithms are from the top submissions to the Multiple Biometric Evaluation (MBE) 2010 Still Face Track [4]. The other four are commercially available algorithms. In order to stress the abilities of algorithms to distinguish between identical twins and to de-emphasize the methods of the individual algorithms, the algorithms will be labeled ‘A’ through ‘G’.

All algorithms were run with their default settings. Therefore, it is important to note that Algorithm C has a built-in threshold value. If a match score would fall below this threshold, the similarity score is automatically returned as zero.

C. Reporting Performance

The primary goal of this paper is to determine the ability of algorithms to differentiate between pairs of identical twins. The metric used allows for an equal comparison between all of the algorithms. A genuine match pair is two images of the same person. For this paper, the genuine match pair consists of two images of either Twin A and Twin A or Twin B and Twin B. From the genuine match pairs the false reject rate (FRR) and verification rate (VR) can be computed. The false reject rate is the percentage of genuine match pairs that are incorrectly regarded as being two different subjects at a given threshold while the verification rate is the percentage of genuine match pairs that are correctly identified as being of the same subject for a given threshold. Unless stated otherwise, the non-match pair consists of one image from each subject in a pair of identical twins; one image is taken from Twin A and the second image is taken from Twin B. From the non-match pairs, the false accept rate (FAR) can be computed. The false accept rate corresponds to the percentage of non-match pairs that are incorrectly regarded as being the same person at a given threshold. Since the non-match pairs are of identical twins, our analysis measures the ability of an algorithm to differentiate between a pair of identical twins.

The primary statistic for reporting the performance of an algorithm will be the equal error rate (EER). The EER is the point where the FRR and the FAR are equal. By plotting the

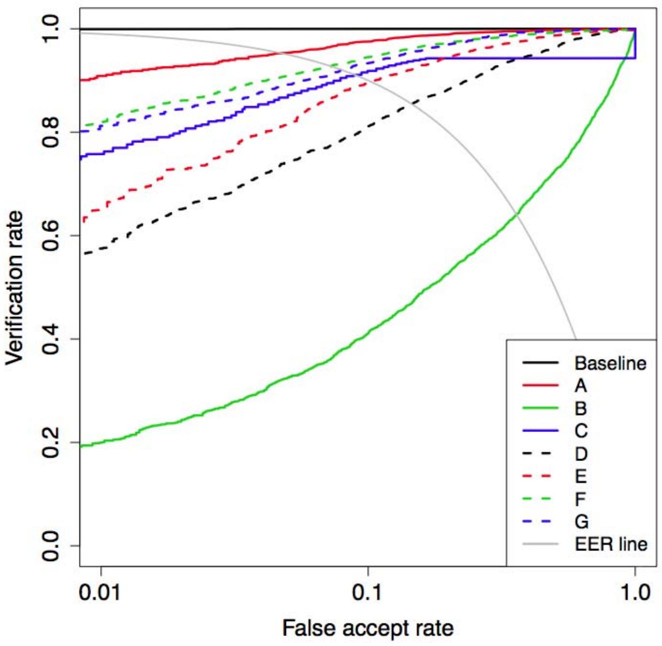


Fig. 4. ROC Curve for Same Day Controlled-Controlled Illumination.

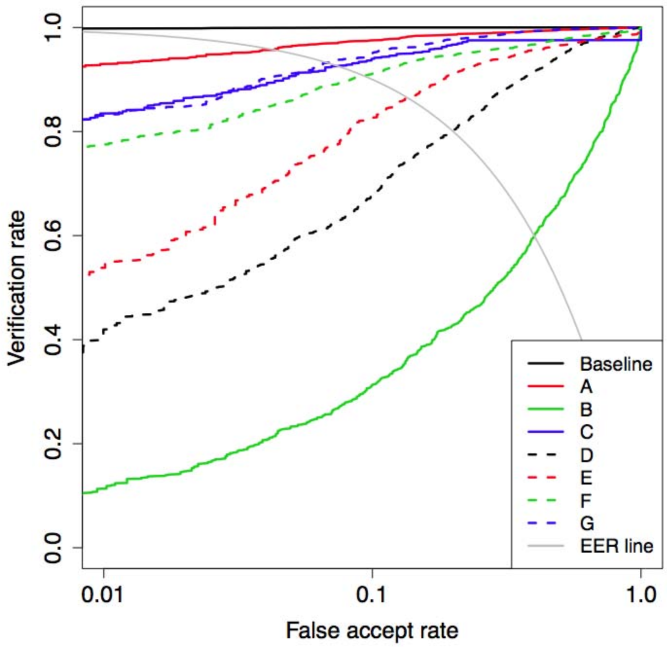


Fig. 5. ROC Curve for Same Day Smiling-Smiling Expression.

TABLE I

EER Results for Same Day Illumination

Alg.	Probe-Gallery Conditions Cont.-Cont.	Cont.-Uncont.	Uncont.-Uncont.
Baseline	0.2%	0.5%	1.1%
A	4.7%	5.9%	11.5%
B	35.9%	40.7%	41.4%
C	9.0%	34.1%	32.3%
D	14.5%	20.9%	26.5%
E	10.2%	13.8%	24.0%
F	7.3%	12.4%	19.4%
G	8.0%	7.8%	16.2%
$\mu \pm \sigma$	12.8 ± 10.6%	19.4 ± 13.8%	24.5 ± 10.1%

TABLE II

EER Results for Same Day Expression

Alg.	Probe-Gallery Conditions Neutral-Neutral	Neutral-Smiling	Smiling-Smiling
Baseline	0.1%	0.5%	0.3%
A	4.5%	7.0%	4.2%
B	39.4%	39.2%	40.0%
C	6.7%	37.6%	7.4%
D	22.2%	22.9%	19.9%
E	14.4%	13.5%	13.5%
F	9.4%	10.8%	9.3%
G	7.7%	8.8%	6.8%
$\mu \pm \sigma$	14.9 ± 12.3%	20.0 ± 13.6%	14.4 ± 12.4%

VR against the FAR, a receiver operating characteristic (ROC) curve can be drawn. For each experiment run, an ROC curve will be plotted and the EER will be used to compare algorithms under the corresponding conditions.

V. Same Day Experiments

The first set of experiments looks to differentiate between identical twins when images of subjects are taken on the same day. Since the images were taken on the same day within a short time interval, they simulate ideal conditions and can provide an approximate upper bound on performance.

All images in these experiments are from the 2009 and 2010 collections while the match and non-match pairs are two images from either 2009 or 2010. Performance was computed from images of 438 subjects (219 pairs of identical twins). Each experiment was run on the baseline algorithm as well as the seven experimental algorithms. The four conditions examined are the effects of illumination, expression, gender, and age.

A. Baseline

In each experimental scenario, the baseline algorithm is run under the same conditions as each algorithm under review. However instead of using the non-match pairs as defined in Section IV.C, the non-match pairs will consist of one image of a subject from a pair of twins, Twin A, and one image of a subject that is not Twin B. By omitting the non-match pairs of Twin A and Twin B, the impostor population is representative of the general impostor distribution. Match pairs for the baseline are the same set of match pairs used for the other algorithms in which Twin A is compared against Twin A. The baseline algorithm used is Algorithm A.

B. Illumination

When examining the effects of illumination, there are two possible conditions to test under, either controlled illumination or uncontrolled illumination. An image taken in controlled illumination was acquired from the tent setup during the 2009 and 2010 acquisitions. The images taken in uncontrolled,

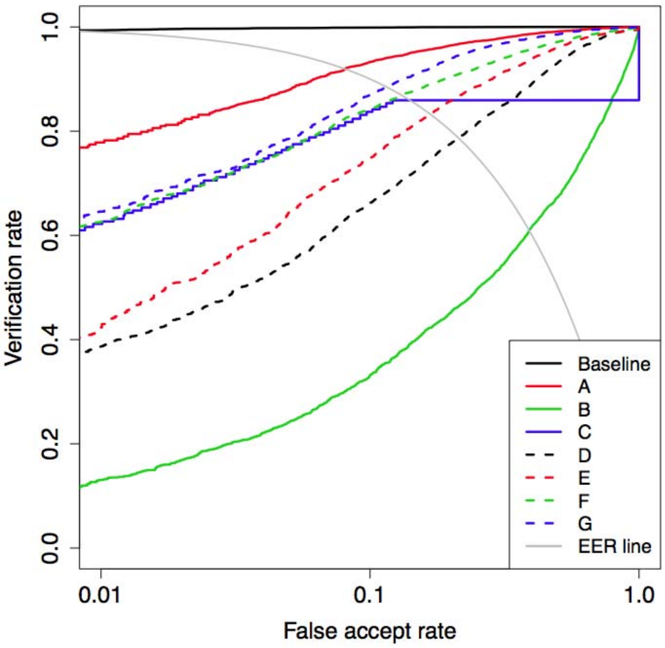


Fig. 6. ROC Curve for Same Day Female Gender.

TABLE III
EER RESULTS FOR SAME DAY GENDER

Alg.	Probe-Gallery Conditions Male	 Female
Baseline	<0.1%	0.7%
A	4.1%	8.1%
B	39.4%	39.1%
C	7.3%	35.1%
D	22.3%	21.3%
E	14.1%	16.7%
F	9.8%	13.1%
G	6.7%	11.5%
$\mu \pm \sigma$	14.8 ± 12.4%	20.7 ± 12.0%

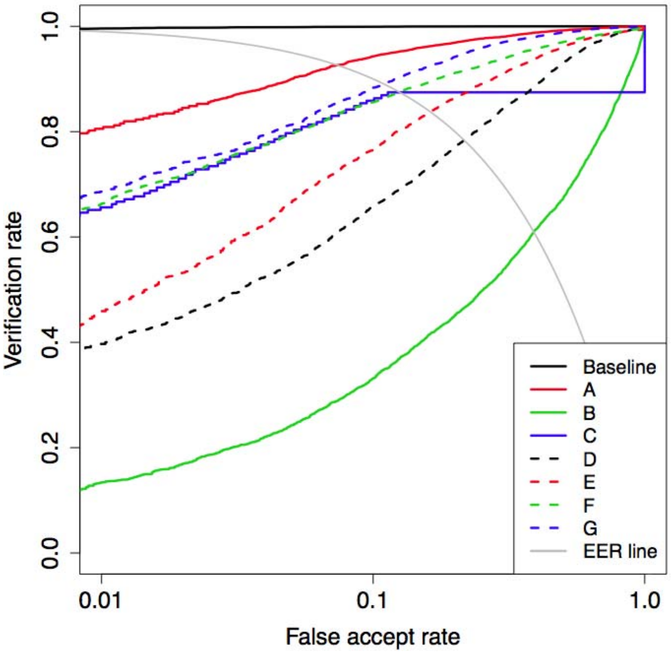


Fig. 7. ROC Curve for Same Day >40 Age.

TABLE IV
EER RESULTS FOR SAME DAY AGE

Alg.	Probe-Gallery Conditions <=40	 >40
Baseline	0.8%	0.6%
A	9.6%	7.4%
B	38.4%	39.0%
C	15.5%	34.1%
D	24.3%	21.6%
E	19.4%	21.6%
F	14.5%	12.5%
G	13.5%	11.0%
$\mu \pm \sigma$	19.3 ± 9.6%	21.0 ± 11.9%

or ambient, illumination were acquired from the outdoor setup during the 2009 acquisition. For a pair of images forming the probe and gallery images, there are three possible combinations of illumination: (i) Controlled-Controlled, (ii) Controlled-Uncontrolled, and (iii) Uncontrolled-Uncontrolled.

The EER results for each algorithm under each probe-gallery combination are shown in Table I. The last row in the table is the mean average and standard deviation of all seven algorithms for each condition. The ROC curves for the Studio-Studio illumination can be seen in Fig. 4. The baseline algorithm has an EER of 0.2% for the Studio-Studio conditions and 1.1% for Ambient-Ambient conditions. The experimental algorithms all have significantly higher equal error rates than the baseline but exhibit the same trend as the baseline.

With the exception of Algorithm G, every other algorithm exhibits an increase in EER when more uncontrolled images are involved. The best results are always achieved using a probe and gallery image acquired in a controlled setting. Algorithm B has the worst performance overall while the remaining algorithms have an EER within 10% of each other. However, the controlled-uncontrolled condition sees an increase in the standard deviation of the algorithms' performance and an increase in the range of EER. Algorithm A had the best performance for every condition among the algorithms.

C. Expression

Two different facial expressions were examined to understand the effects of expression on facial recognition. Those two expressions are a neutral (sometimes referred to as a blank stare) expression and a smiling (or happiness) expression. During the 2009 acquisition, subjects were asked to pose with a neutral expression in the first picture and then with a smiling expression in the second picture both in the controlled and uncontrolled settings. The 2010 acquisition had subjects pose with a neutral and smiling expression in the controlled setting. For a pair of images, there are three possible combinations of expression: (i) Neutral-Neutral, (ii) Neutral-Smiling, and (iii) Smiling-Smiling.

The EER results for each algorithm under each probe-gallery combination are shown in Table II. The ROC curves

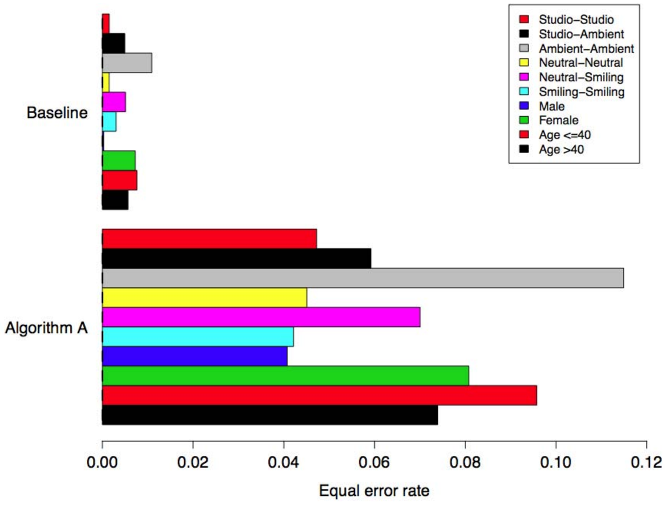


Fig. 8. EER results for Same Day.

for the Smiling-Smiling expression can be seen in Fig. 5. The baseline performs best when the probe and gallery expressions agree with the best performance occurring under a Neutral-Neutral expression. Again, all of the algorithms see a significant increase in EER compared to the baseline.

The experimental algorithms also exhibit the same behavior as the baseline. The lowest EER is observed when the expression of the probe images matches the expression of the gallery image, whether it be neutral or smiling. Of the two expressions, a smiling expression performs slightly better than a neutral expression but not significantly better. Algorithm A had the best performance for every condition among the algorithms.

D. Gender

Results broken down by gender are shown in Table III. The ROC curves for Females can be seen in Fig. 6. The baseline algorithm has an EER of less than 0.1% for males and slightly higher EER for females. While there is no significant difference between males and females, five of the seven experimental algorithms have a lower EER for twins that are male. The results for all the experimental algorithms are significantly higher than the baseline.

E. Age

The final covariate analyzed is the effect of age on performance. Subjects range in age from 18 years old to 79 years old. The age is broken into two categories: (i) over 40 years old - born before 1969 and (ii) 40 years old and younger - born in 1969 or later.

The EER results for each algorithm under each probe-gallery combination are shown in Table IV. The ROC curves for subjects over forty years of age can be seen in Fig. 7. The baseline has a slightly lower EER for those subjects older than 40 years of age. All of the algorithms have a significantly higher EER than the baseline. There is no significant difference between the two age groups. Four of the seven algorithms have a lower EER for the over forty age group while the remaining three algorithms have a lower EER for those subjects 40 years old or younger. Algorithm A has the best performance for both age groups among the algorithms.

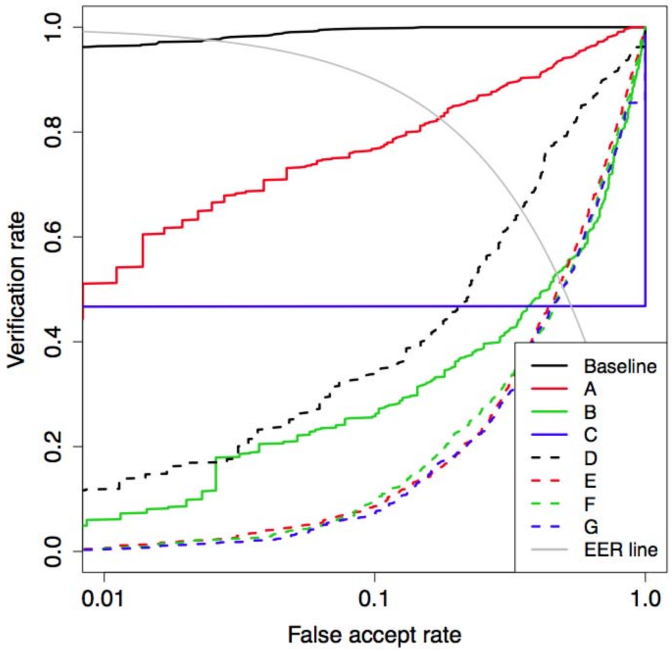


Fig. 9. ROC Curve for Cross-Year Controlled-Controlled Illumination.

TABLE V

EER RESULTS FOR CROSS-YEAR ILLUMINATION

	Probe-Gallery Conditions	
Alg.	Cont.-Cont.	Cont.-Uncont.
Baseline	0.8%	2.4%
A	12.8%	17.4%
B	43.3%	47.0%
C	41.9%	51.7%
D	29.8%	34.7%
E	49.3%	48.5%
F	49.3%	50.0%
G	49.5%	49.7%
$\mu \pm \sigma$	39.4 ± 12.6%	42.7 ± 11.6%

F. Results

For a pair of images of identical twins acquired on the same day, every algorithm has a varying degree of difficulty classifying the pair as match or nonmatch. The equal error rate of the experimental algorithms is significantly higher than the baseline algorithm under every covariate condition. It is easiest to distinguish identical twins with controlled illumination. Twins are easier to identify if both the probe image and gallery image have the same expression. There is no significant difference when comparing genders or age groups.

The baseline algorithm has the lowest EER when considering only male subjects (<0.1%) and the highest EER when the images are acquired with uncontrolled illumination (1.1%). This difference is a range of 1% in equal error rate for any covariate value. However, when distinguishing between

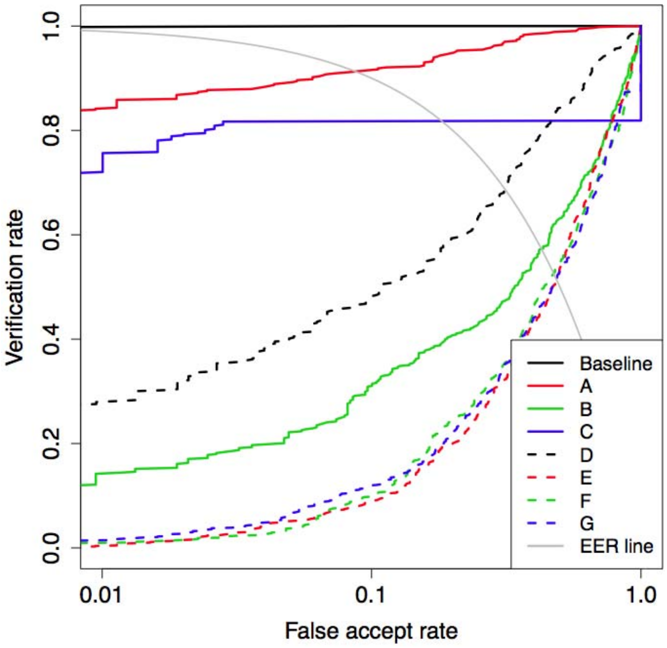


Fig. 10. ROC Curve for Cross-Year Neutral-Neutral Expression.

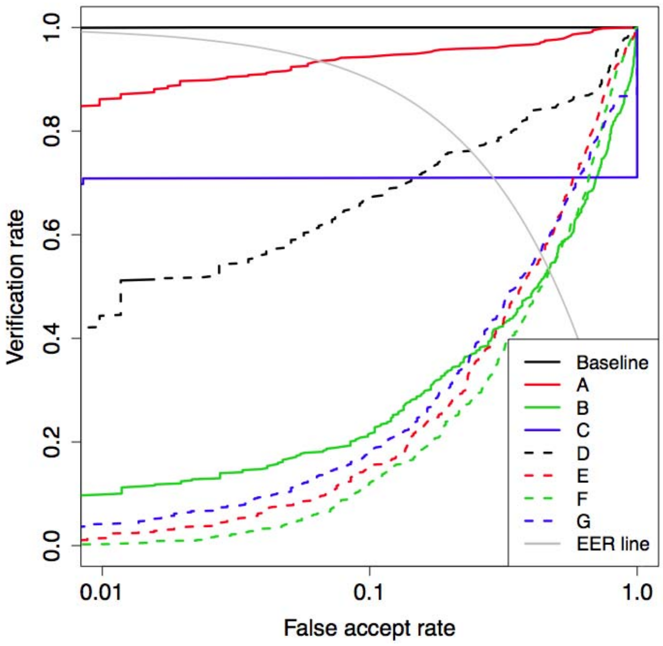


Fig. 11. ROC Curve for Cross-Year Male Gender.

TABLE VI
EER RESULTS FOR CROSS-YEAR EXPRESSION

Alg.	Probe-Gallery Conditions Neutral-Neutral	 Neutral-Smiling	 Smiling-Smiling
Baseline	0.4%	2.3%	0.6%
A	8.6%	15.8%	8.1%
B	42.7%	46.0%	44.9%
C	34.8%	55.5%	35.7%
D	31.2%	32.0%	25.7%
E	48.8%	47.4%	50.1%
F	47.4%	50.0%	52.0%
G	48.4%	50.0%	49.8%
$\mu \pm \sigma$	37.4 ± 13.4%	42.4 ± 12.8%	38.0 ± 15.0%

TABLE VII
EER RESULTS FOR CROSS-YEAR GENDER

Alg.	Probe-Gallery Conditions Male	 Female
Baseline	0.6%	2.0%
A	6.5%	13.7%
B	46.7%	43.6%
C	39.7%	46.9%
D	23.6%	29.0%
E	43.9%	50.8%
F	47.1%	50.3%
G	43.5%	51.3%
$\mu \pm \sigma$	35.9 ± 14.1%	40.8 ± 13.2%

identical twins even the best performing algorithm exhibits a range of 7%. Overall, the algorithms have an average range of 12% between the best performer and worst performer. The increase in range is evidence of the difficulty algorithms have in distinguishing identical twins.

When considering a general population that is absent of twins, the baseline algorithm has an equal error rate of 1.1% or less. Once identical twins are introduced into the population, the average algorithm has an equal error rate between 15-20%. Even with a pair of images taken minutes or hours apart on the same day, face recognition algorithms have varying degrees of difficulty distinguishing between a pair of identical twins. Algorithm A consistently had the best performance on the twins dataset and Fig. 8 shows the equal error rates of the baseline and Algorithm A for each covariate condition.

VI. CROSS-YEAR EXPERIMENTS

The next set of experiments aim to distinguish between identical twins from images taken one year apart. The same day experiments represented an ideal scenario of images taken within a short time span, but the cross-year experiments present a more realistic scenario where one image has been stored in a database and the probe image was recently acquired.

All images in these experiments are from the 2009 and 2010 collections where subjects participated in both years. Performance was comported from images of 48 subjects (24 pairs of identical twins). Each experiment was run on the baseline algorithm as well as the seven experimental algorithms. The four conditions examined are the effects of illumination, expression, gender, and age.

A. Baseline

As with the Same Day Experiments, the baseline algorithm for the Cross-Year Experiments will be run under the same conditions as each algorithm. Again, the baseline is run using only Twin A with Twin B omitted from the population to simulate a general impostor distribution. The baseline algorithm used is Algorithm A.

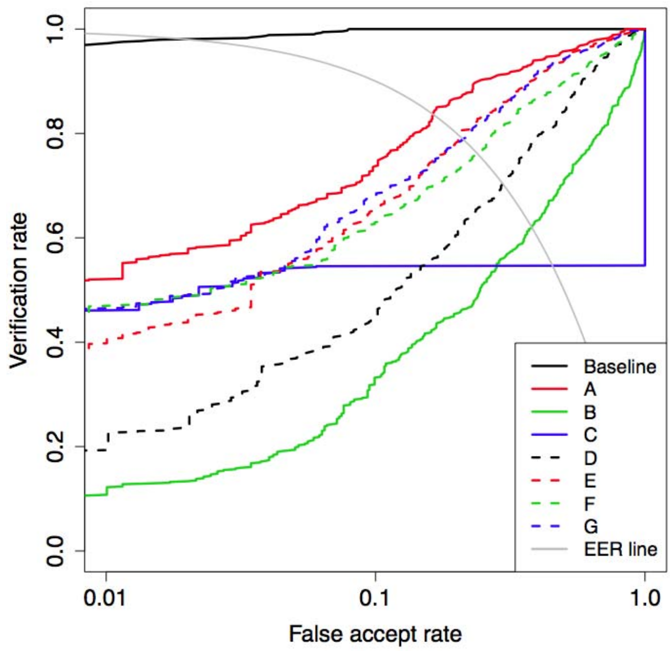


Fig. 12. ROC Curve for Cross-Year <40 Age.

TABLE VIII
EER RESULTS FOR CROSS-YEAR AGE

Alg.	Probe-Gallery Conditions <=40	>40
Baseline	2.0%	1.6%
A	16.3%	14.8%
B	38.0%	44.5%
C	49.2%	45.3%
D	29.7%	31.9%
E	20.4%	23.3%
F	23.9%	22.8%
G	21.1%	20.1%
$\mu \pm \sigma$	28.4 ± 10.6%	28.9 ± 11.1%

B. Illumination

During the 2009 acquisition, images were acquired in the controlled tent and uncontrolled ambient environments. The 2010 acquisition acquired images only in a controlled tent environment. Therefore, there are only two combinations of illumination for the cross-year experiments: (i) Controlled-Controlled using images from 2009 and 2010 (ii) Controlled-Uncontrolled where the uncontrolled images are from 2009 only.

The EER results for each algorithm under each probe-gallery combination are shown in Table V. The ROC curves for the Studio-Studio illumination can be seen in Fig. 9. The baseline has a lower EER for the Studio-Studio condition and compared to the same day illumination results the EER is slightly. The EER for all of the experimental algorithms are significantly higher than the baseline EER and also are significantly higher than the corresponding same day EER for each algorithm. Algorithm A has the best performance among all the algorithms. Five of the seven algorithms have great trouble differentiating between identical twins and their EER is at or near 50%.

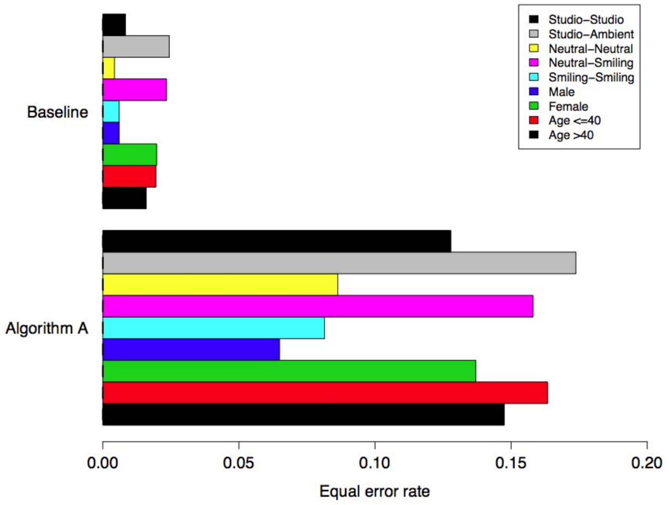


Fig. 13. EER results for Cross-Year.

For every condition, the EER of each algorithm is significantly higher than the baseline and the corresponding algorithm in the same day experiment. The experimental algorithms exhibit the same trend as the baseline whereas performance is best when the probe and gallery images have the same expression. The best performance is achieved with Algorithm A. There are four algorithms that have an EER at or near 50% for every expression condition.

C. Expression

In both 2009 and 2010, images were acquired of subjects with both a neutral expression and smiling expression. The EER results for each algorithm under each probe-gallery combination are shown in Table VI. The ROC curves for the Neutral-Neutral condition are shown in Fig. 10. As with the same day experiments, the baseline EER is lowest when both the probe and gallery images have the same expression. The EER for the neutral-neutral condition is slightly lower than the smiling-smiling condition.

D. Gender

The EER results for each algorithm under each probe-gallery combination are shown in Table VII. The ROC curves for Males are shown in Fig. 11. The baseline algorithm found males easier to distinguish than females and the EER increased compared to the same day results. All of the algorithms had a significantly higher EER than the baseline. Males had a slightly lower EER for all but one algorithm but there is no preference for one gender over the other. Algorithm A had the lowest EER and five of the seven algorithms had an EER at or near 50% for both genders.

E. Age

The EER results for each algorithm under each probe-gallery combination are shown in Table VIII. The ROC curves

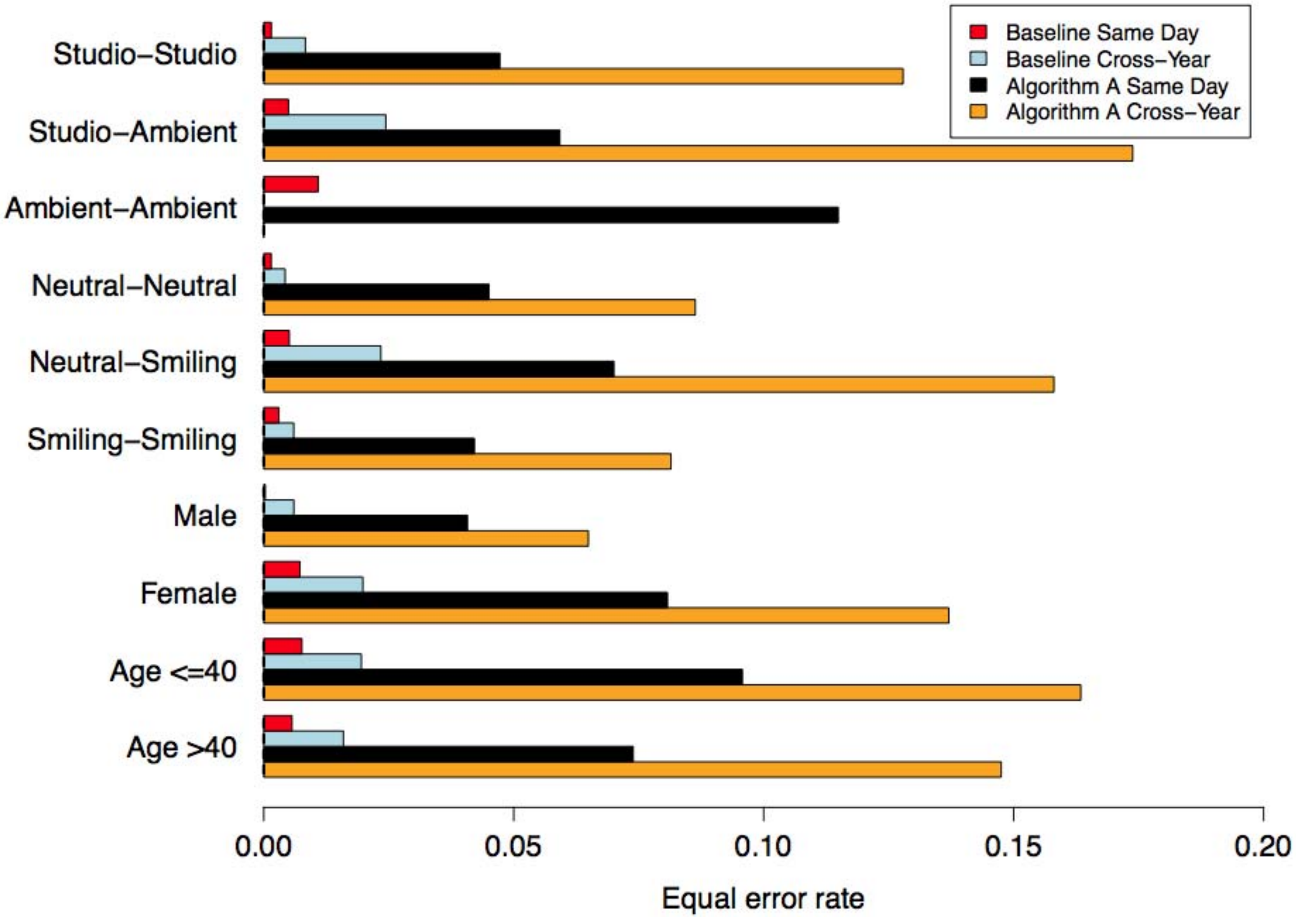


Fig. 14. EER results for same day and cross-year experiments. Comparing the baseline and top performing algorithm.

for subjects forty years old or younger can be seen in Fig. 12. The baseline algorithm has a slightly lower EER for subjects older than forty and slightly higher compared to the same day results. Every experimental algorithm has an EER that is significantly higher than the baseline but only slightly higher than the same day results. There is still no age group that is favored over the other. Algorithm A has the best performance for both age groups.

F. Results

For images acquired a year apart, every algorithm exhibited a decrease in performance compared to the same day results. The baseline algorithm had a slight increase in equal error rate compared to the same day results. The equal error rates for the experimental algorithms were significantly higher than the baseline in every covariate condition. The cross-year results of each algorithm are also significantly higher than the same day results for every covariate except age. When images are acquired one year apart, there is no value of any covariate that proves to be easier to differentiate twins under.

When considering a general population that is absent of twins, the baseline algorithm has an equal error rate of 2.4% or less. When the population consists of only identical twins, the typical algorithm has an equal error rate between 25-40%. As the elapsed time between image acquisition increases, face recognition algorithms have greater difficulty distinguishing between a pair of identical twins. Algorithm A consistently had the best performance on the twins dataset and Fig. 13 shows the equal error rates of the baseline and Algorithm A for each covariate condition.

The results of the baseline and top performing algorithm for both the same day and cross-year experiments are shown in Fig. 14. The equal error rates for Algorithm A are significantly higher than the equal error rates for the baseline in both the same day and cross-year experiments. The cross-year results are also significantly higher than the same day results for Algorithm A. For every condition of illumination, expression, gender, age, and time between acquisitions, it is more difficult to differentiate identical twins than a general population without twins.

VII. Conclusion

This paper has shown using the Twins Days Dataset [14] that differentiating between identical twins is a difficult problem. Facial recognition algorithms that are robust with a general population generally do not perform nearly as well on a population consisting of only identical twins. Experimental results measured the performance when faces were collected on the same day and one year apart. The results also examined the effects of changes in illumination and expression as well as the effect of gender and age on performance.

Performance varied significantly. While the baseline's equal error rate ranged from <0.1% to 2.4%, the best performing algorithm had an equal error rate from 4.1% to 17.4%. It is easier to distinguish identical twins when images have been acquired on the same day instead of one year apart. Studio illumination resulted in the highest performance and there was no preference of expression as long as the two images being compared had the same expression. The gender and age of the subjects do not affect the performance.

The results show that it is possible to distinguish identical twins under ideal conditions (same day acquisition, studio-like illumination, consistent expression). However, when conditions are less than ideal, it is very challenging to distinguish identical twins. New research ideas are needed to help improve performance on recognition of identical twins in realistic imaging scenarios.

If face recognition algorithms strive to perform under the most difficult conditions, then the algorithms should be presented with the most difficult problems. Identical twins represent a very difficult recognition problem and the algoithms studied performed significantly worse than the baseline. The result from these experiments show that there is still room for improvement in face recognition. One specific area that needs additional attention is identical twins.

References

[1] A. Ariyaeeiniaa, C. Morrison, A. Malegaonkara, and B. Black, "A test of the effectiveness of speaker verification for differentiating between identical twins," *Sci. Justice*, vol. 48, no. 4, pp. 182–186, Dec. 2008.

[2] S. Biswas, K. W. Bowyer, and P. J. Flynn, "A study of face recognition of identical twins by humans," in *Proc. IEEE WIFS*, Dec. 2011, pp. 1–6.

[3] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Three-dimensional face recognition," *Int. J. Comput. Vis.*, vol. 64, no. 1, pp. 5–30, Aug. 2005.

[4] P. J. Grother, G. W. Quinn, and P. J. Phillips, "MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms," NIST, Gaithersburg, MD, USA, Tech. Rep. NISTIR 7709, 2010.

[5] K. Hollingsworth, K. Bowyer, and P. Flynn, "Similarity of iris texture between identical twins," in *Proc. IEEE Comput. Soc. Conf. CVPRW*, Jun. 2010, pp. 22–29.

[6] A. Jain, S. Prabhakar, and S. Pankanti, "On the similarity of identical twin fingerprints," *Pattern Recognit.*, vol. 35, no. 11, pp. 2653–2663, Nov. 2002.

[7] K. Kodate, R. Inaba, E. Watanabe, and T. Watanabe, "Facial recognition by a compact parallel optical correlator," *Meas. Sci. Technol.*, vol. 13, no. 11, pp. 1756–1766, Nov. 2002.

[8] A. W. Kong, D. Zhang, and G. Lu, "A study of identical twins' palmprints for personal verification," *Pattern Recognit.*, vol. 39, no. 11, pp. 2149–2156, Nov. 2006.

[9] U. Park and A. Jain, "Face matching and retrieval using soft biometrics," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 406–415, Sep. 2010.

[10] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, *et al.*, "Distinguishing identical twins by face recognition," in *Proc. IEEE Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 185–192.

[11] M. T. Pruitt, J. M. Grant, J. R. Paone, P. J. Flynn, and R. W. V. Bruegge, "Facial recognition of identical twins," in *Proc. 1st Int. Joint Conf. Biometrics*, Oct. 2011, pp. 185–192.

[12] Z. Sun, A. A. Paulino, J. Feng, Z. Chai, T. Tan, and A. K. Jain, "A study of multibiometric traits of identical twins," *Proc. SPIE, Biometric Technol ogy for Human Identification VII*, vol. 7667, pp. 76670T-1–76670T-12, Apr. 2010.

[13] (2013, May 24). *CVRL Data Sets* [Online]. Available: http://www.nd.edu/~cvrl/CVRL/Data_Sets.html

[14] (2013, May 24). *Twins Days Festival* [Online]. Available: http://www.twinsdays.org

[15] N. Ye and T. Sim, "Combining facial appearance and dynamics for face recognition," in *Computer Analysis of Images and Patterns* (Lecture Notes in Computer Science), vol. 5702. X. Jiang and N. Petkov, Eds. Heidelberg, Germany: Springer-Verlag, 2009, pp. 133–140.

Jeffrey R. Paone is currently a Post-Doctoral Researcher with the Oak Ridge National Laboratory. He received the Ph.D. degree from the University of Notre Dame with a dissertation on the biometric menagerie applied to face and iris recognition. His research interests include face and iris recognition, camera calibration, computer vision, and augmented reality.

Patrick J. Flynn (F'12) is a Professor of computer science and engineering and Concurrent Professor of electrical engineering with the University of Notre Dame. He received the Ph.D. degree in computer science from Michigan State University in 1990. His research interests include computer vision, biometrics, and image processing.

Dr. Flynn is an IAPR Fellow, an ACM Distinguished Scientist, a past Associate Editor-in-Chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), and a past Associate Editor of the IEEE Transactions on Information Forensics and Security, the IEEE Transactions on Image Processing, the IEEE TPAMI, *Pattern Recognition*, and *Pattern Recognition Letters*.

P. Jonathon Phillips (F'10) is a leading researcher in the fields of computer vision, biometrics, face recognition, and human identification. He is with the National Institute of Standards and Technology (NIST), where he works on designing grand challenges for advancing face recognition and visual biometric technology and science. His previous efforts include the Iris Challenge Evaluations, the Face Recognition Vendor Test (FRVT) 2006, and the Face Recognition Grand Challenge and FERET. From 2000 to 2004, he was assigned to the Defense Advanced Projects Agency as a Program Manager for the Human Identification at a Distance Program. He was a Test Director for the FRVT 2002. For his work on the FRVT 2002, he received the Department of Commerce Gold Medal. His work has been reported in print media of record, including the *New York Times* and the *Economist*. He has appeared on NPR's ScienceFriday. Prior to joining NIST, he was with the U.S. Army Research Laboratory. He received the Ph.D. degree in operations research from Rutgers University. From 2004 to 2008, he was an Associate Editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence and Guest Editor of an issue of the Proceedings of the IEEE on biometrics. In an Essential Science Indicators analysis of face recognition publication over the past decade, his work ranks at #2 by total citations and #1 by cites per paper. He received the Inaugural Mark Everingham Prize. He is a fellow of the IAPR.

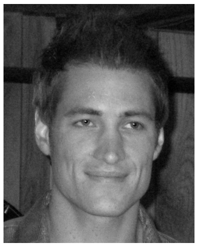
Kevin W. Bowyer (F'98) is the Schubmehl-Prein Professor and Department Chair of the Department of Computer Science and Engineering, University of Notre Dame.

Prof. Bowyer's most recent book is the *Handbook of Iris Recognition*, edited with Dr. M. Burge. He is serving as a General Chair of the 2015 IEEE International Conference on Automatic Face and Gesture Recognition. He has served as an Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE BIOMETRICS COMPENDIUM, as the General Chair of the 2007, 2008, and 2009 IEEE International Conference on Biometrics Theory Applications and Systems and the 2011 International Joint Conference on Biometrics, and as a Program Chair of the 2011 Automatic Face and Gesture Recognition Conference.

Prof. Bowyer is a fellow of the IAPR, and a Golden Core Member of the IEEE Computer Society. He has made advances in many areas of biometrics, including iris recognition, face recognition, and multibiometric methods. His research group has been active in support of a variety of government-sponsored biometrics programs, including the Human ID Gait Challenge, the Face Recognition Grand Challenge, the Iris Challenge Evaluation, the Face Recognition Vendor Test 2006, and the Multiple Biometric Grand Challenge.

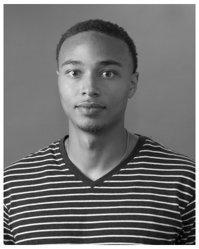
Richard W. Vorder Bruegge is a Senior Photographic Technologist with the Federal Bureau of Investigation (FBI), Science and Technology Branch, where he is responsible for overseeing science and technology developments in the imaging sciences. He received the B.S. degree in engineering, the M.S. degree in geological sciences, and the Ph.D. degree in geological sciences from Brown University in 1985, 1987, and 1991, respectively. He has been with the FBI since 1995, where he has performed forensic analysis of image and video evidence, testifying in state, federal, and international courts as an expert witness over 60 times. His research interests include the forensic analysis of image evidence, with a particular interest in face recognition. He was a Chair of the Scientific Working Group on Imaging Technology from 2000 to 2006 and is a Current Chair of the Facial Identification Scientific Working Group. He is a fellow of the American Academy of Forensic Sciences, and he was named a Director of National Intelligence Science and Technology Fellow in 2010.

Patrick J. Grother is a Staff Scientist with the National Institute of Standards in Technology responsible for biometric standards development, algorithm testing, and analysis. He runs the ongoing MINEX, IREX, and FRVT/MBE evaluations of fingerprint, iris and face recognition technologies. He leads development of biometric specifications for U.S. Government identity credentials, and was the recipient of the Department of Commerce Gold Medals in 2003 and 2007, for that work and for tests of border management technologies. He has interests in large scale biometric systems, performance testing, failure analysis, ageing, fusion, fingerprint minutia detection, and iris and face recognition.



George W. Quinn is a Computer Scientist with the National Institute of Standards and Technology. He was involved in various biometrics technology evaluations, including the iris interoperability exchange, the evaluation of latent fingerprint technology, and the multiple biometrics evaluation still face test. His research interests include biometric sample quality, biometric fusion, and statistical computing. He received the bachelor's degree in mathematics and computer science from the University of Maryland.

Matthew T. Pruitt, photograph and biography not available at the time of publication.



Jason M. Grant received the B.S. degree in computer engineering from the University of Maryland, Baltimore, in 2010, and the M.S. degree from the University of Notre Dame in 2013, where he is currently pursuing the Ph.D. degree. A Deans Fellowship recipient and GAANN Teaching Fellow, he currently researches topics in computer vision related to face and activity recognition.