Distinguishing Identical Twins by Face Recognition

P. Jonathon Phillips, Patrick J. Flynn, Kevin W. Bowyer, Richard W. Vorder Bruegge,
Patrick J. Grother, George W. Quinn, Matthew Pruitt

Abstract— **The paper measures the ability of face recognition algorithms to distinguish between identical twin siblings. The experimental dataset consists of images taken of 126 pairs of identical twins (252 people) collected on the same day and 24 pairs of identical twins (48 people) with images collected one year apart. Recognition experiments are conducted using three of the top submissions to the Multiple Biometric Evaluation (MBE) 2010 Still Face Track [1]. Performance results are reported for both same day and cross year matching. Performance results are broken out by lighting conditions (studio and outside); expression (neutral and smiling); gender and age. Confidence intervals were generated by a bootstrap method. In terms of both the number of paris of twins and lapsed time between acquisitions, this is the most extensive investigation of face recognition performance on twins to date.**

I. Introduction

In the face recognition community, the conventional wisdom is that distinguishing between identical twins is one of the most challenging problems in face recognition. This paper presents the first detailed study of the ability of face recognition algorithms to distinguish between identical twins. The data in this study includes face images of 126 pairs of identical twins (252 people) collected on the same day and images from 24 pairs of identical twins (48 people) collected one year apart. Recognition performance is reported for three of the top submissions to the Multiple Biometric Evaluation (MBE) 2010 Still Face Track [1].

Experiments report the ability of the algorithms to distinguish between identical twins under five experimental conditions. These conditions include elapsed time between image acquisition. Performance is measured for images collected on the same day and separated by a year. Images were collected in both a mobile studio environment and in outside ambient lighting. Images of a subject were collected with both a neutral and smiling expression. Performance is also broken out by gender and age. Ninety percent confidence intervals were generated by a bootstrap method.

The number of experimental conditions allows for a robust assessment of the ability to distinguish between identical twins. The recognition results from images taken on the same day in the studio environment show performance under ideal conditions. The one-year time lapse recognition experiments provide a glimpse of potential performance under operational conditions.

The authors are only aware of one paper that examines the ability of face recognition algorithms to recognize identical twins. Sun et al. [2] conducted a study of biometric identification of identical twin siblings using the face, iris, and fingerprint modes as well as a fusion of these modalities. The data they used was collected in 2007 at the fourth Annual Festival of Beijing Twins Day. All images were collected during a single session. The data set used for experimentation consisted of samples from 134 subjects: 64 pairs of twins and two sets of triplets. Face recognition experiments were performed by a Cognitec FaceVACS system. The main result from the face recognition experiments was that the "identical twin impostor" distribution (i.e., the set of scores matching images of identical twin siblings) was "more similar to the [match] distribution than to the general impostor distribution." The main conclusion is that, for the Cognitec FaceVACS system, there is greater overlap between the match distribution and the non-match distribution consisting of identical twin sibling face images than a general impostor distribution.

II. Data

Data supporting these experiments were collected at the Twins Days festival [3] in Twinsburg, Ohio in August 2009 and August 2010. The Twins Days festival is a weekend event with a typical attendance of between 1500 and 2000 twin sibling pairs along with other multiple-birth sibling groups and their family members. Twins attending the festival range in age from newborn to elderly. Attendees at the festival represent a variety of different ethnic groups and races, and Caucasians are the largest single group. Research groups are hosted in a designated area on the festival grounds and research groups who provide advertisement copy to the festival organizers receive publicity in the Festival's printed program. Prior to acceptance by the Festival, the data acquisition protocol was reviewed and approved by the Festival organizers. All data was collected under the approved protocol and subjects completed a consent form prior to each acquisition. If subjects wished, they were allowed to participate on both days of the festival.

Our twins data collection involved a half-day of setup and equipment testing the day before the Festival opened, followed by two full days of data collection, followed by equipment tear-down and departure. Data validation and enrollment consumed several weeks after the completion

P. J. Phillips, P. J. Grother, and G. W. Quinn are with the National Institute of Standards and Technology, 100 Bureau Dr., MS 7740 Gaithersburg MD 20899, USA (e-mail: jonathon@nist.gov). Please direct correspondence to P. J. Phillips.

P. J. Flynn, K. W. Bowyer, and M. Pruitt are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA.

R. W. Vorder Bruegge is with the Federal Bureau of Investigation, Science & Technology Branch, Digital Evidence Laboratory, Building 27958A, Quantico, VA 22135, USA.

of data collection activities. In 2009, our data acquisitions performed for each subject included

- 2D face still photography with simultaneous HD video recordings, captured inside the rented tent under studio lighting, and with ambient outdoor lighting in an area adjacent to the tent; and
- iris video capture using an LG2200 EOU system attached to a digital video recorder;

In 2010, data acquisitions were conducted jointly by Notre Dame and West Virginia University, and included the following modes for each subject:

- 2D face still photography with simultaneous HD and 4k x 2k video recordings, captured inside the rented tent under studio lighting;
- 3D face stills captured from a Minolta 910 range scanner;
- iris still images captured with an LG4000 iris camera; and
- fingerprints captured with a CrossMatch sensor.

The 2009 collection yielded 17,486 face stills from 252 twin subjects (126 pairs), of whom 34 (17 pairs) appeared in each of the two days of the Festival. Figure 1 shows an example set of images for a subject who participated in both the 2009 and 2010 collections. In 2010, data collection yielded 6863 face stills from 240 twin subjects (120 pairs), of whom 10 (5 pairs) came both days. There were 48 twins (24 pairs) who participated in both 2009 and 2010 acquisitions, and two twin subjects (one pair) participated in both days of both years. Finally, one set of identical triplets participated in 2010.

III. Methods and Materials

A. Algorithms

Performance is reported for three of the top submissions to the Multiple Biometric Evaluation (MBE) 2010 Still Face Track [1]. The algorithms were run in verification mode. To emphasize the potential for algorithms to distinguish between identical twins, the algorithms are de-identified and labelled 'A', 'B', and 'C'.

B. Reporting Performance

The goal of this study is to measure the ability of algorithms to distinguish between identical twins. The primary performance statistics reflect this goal. A match face pair consist of two images of the same person. In this paper, a match face pair consists of two images of a person who has an identical twin in this study. The performance statistics false reject rate (FRR) and verification rate (VR) are computed from the match face pairs. Unless explicitly stated otherwise, in this paper, a non-match face pair consists of one image from each person in a pair of identical twins. From the non-match pairs, false accept rates (FAR) are computed. When the non-match face pairs are identical twins the analysis measures the ability of an algorithm to distinguish between identical twins.

The primary statistic for reporting performance will be the equal error rate (EER). The EER is the point where the FRR and the FAR are equal. Distinguishing between identical twins is similar to a two-alternative force choice paradigm and is related to the area under the receiver operating characteristic (ROC) [4],[5]. However, the EER was selected instead of the area under the ROC statistic because the EER has a direct relationship to the classical performance measures verification and false accept rates.

Confidence intervals are generated by a bootstrap method. The sampling method is based on the subset bootstrap technique applied to biometrics [6]. The bootstrap samples at the level of pairs of identical twins. When a pair of identical twins is sampled, all match scores and non-matches for that pair of identical twins is selected. If a pair of identical twins were sampled n times, then n copies of all the pairs match and non-match scores for that pair of identical twins were extracted.

IV. Experiments

A. Same Day

The first experiment measures the ability to distinguish between identical twins when their images are collected on the same day. Because the images were collected on the same day, they provide an upper bound on performance.

All images in this experiment were collected during Twins Days 2009. Results are reported for six experimental conditions. In the first condition, all images were collected in the mobile studio and subjects had a neutral expression. In Figure 2 this condition is label 'Studio Neutral.' This experimental condition measures performance under the best possible environment. Figure 2 gives the EERs for all of the six conditions in the same day experiments. The EERs are plotted along with a 90% confidence interval. Table I gives one confidence interval for each of the experimental conditions in Figure 2.

TABLE I

This table provides one confidence interval for each of the six experimental conditions in Figure 2. The confidence intervals are at a 90% level for the EER. The confidence interval provided is for the algorithm with the smallest lower endpoint.

	Confidence Interval	
Experimental condition	Lower endpoint	Upper endpoint
Studio Neutral	0.01	0.04
Studio Neutral-Smile	0.04	0.07
Studio-Ambient Neutral	0.03	0.07
Studio-Ambient Neutral-Smile	0.05	0.10
Ambient Neutral	0.12	0.21
Ambient Neutral-Smile	0.12	0.16

In the second condition, all images were collected with ambient outside lighting and subjects had a neutral expression (labeled 'Ambient Neutral' in Figure 2). This condition measured the ability to distinguish twins when both images were collect under ambient lighting on the same day.

In the third condition, all images were collected in the mobile studio. The algorithms compared two images where the face in one image had a neutral expression and in the

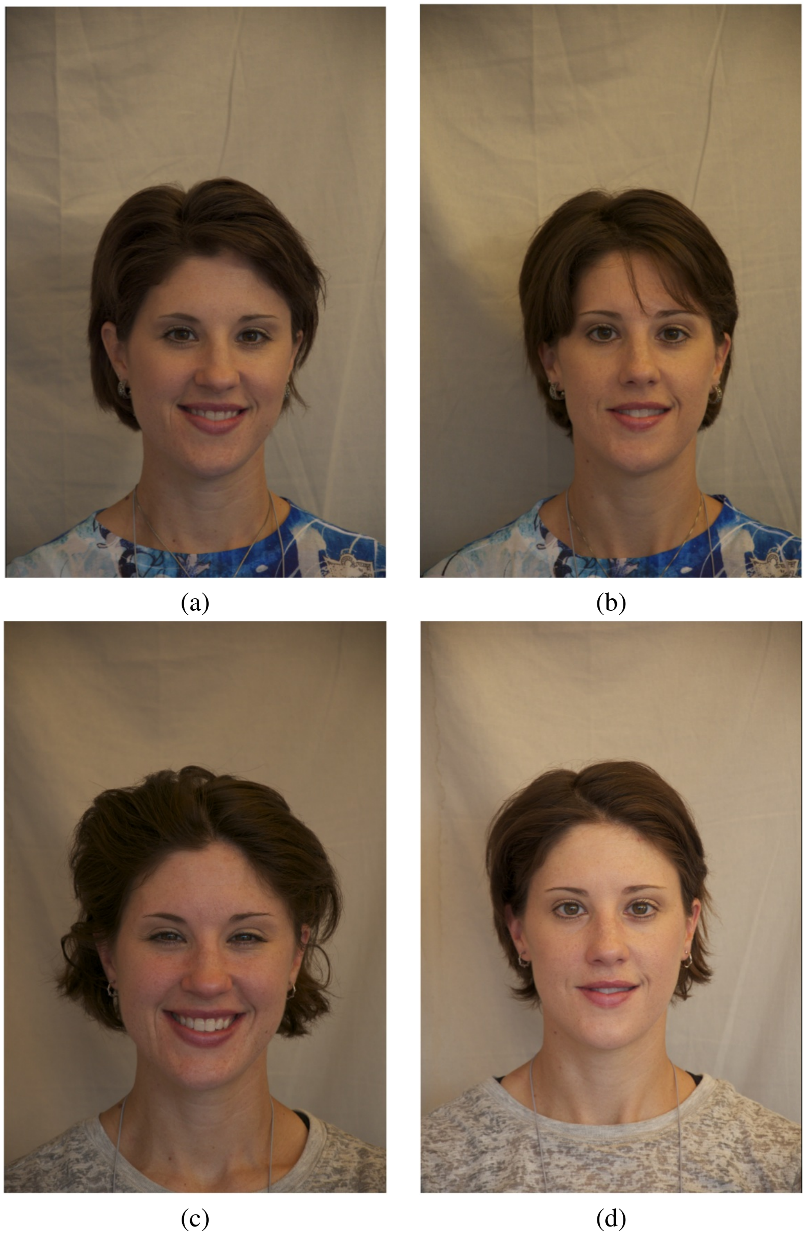

Fig. 1. Example images from a pair of identical twins acquired in both 2009 and 2010. Images (a) and (b) were collected in August 2009, and images (c) and (d) were collected in August 2010. Images (a) and (c) are of the same twin, and images (b) and (d) are of the same twin. All four images were taken in the mobile studio environment.

second image, the face had a smile (labeled 'Studio Neutral-Smile'). In the fourth condition, all images were collected with ambient outside lighting. The algorithms compared two images where the face in one image had a neutral expression and in the second image, the face had a smile (labeled 'Ambient Neutral-Smile'). These two conditions measure the impact of a change in expression on performance. The experiment examined the effect of a change in expression for both studio and ambient lighting factors.

The next two experimental conditions measure the effect of changing the image capture environment. In both conditions, the algorithms compare images where one image was acquired in the studio and the other was acquired under ambient lighting. In the fifth condition, all faces had a neutral expression (labeled 'Studio-Ambient Neutral'). In the sixth condition, one face had a neutral expression and the second face was smiling (label 'Studio-Ambient Neutral-Smile').

For all algorithms, at the 90% confidence level, the 'Studio Neutral' condition has the best performance. For Algorithms B and C, they have the highest EER for 'Ambient Neutral' condition. For Algorithm A, there is no statistically significant difference in the EER for the 'Ambient Neutral' and 'Ambient Neutral-Smiling' conditions.

In the Studio to Studio comparisons, for all three algorithms, a change in expression yields a statistically significant change in the EER. Also, in the Studio to Ambient lighting

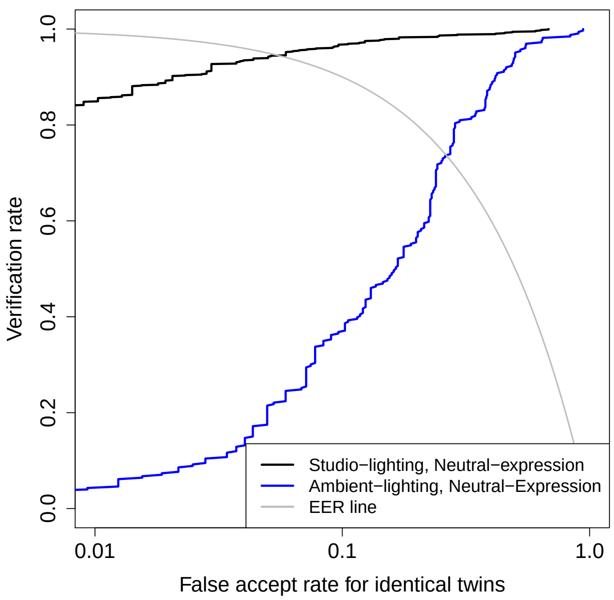

Fig. 3. A ROC showing the extremes in performance on the same day.

comparisons, for all three algorithms, a change in expression yields a statistically significant change in the EER.

For all three algorithms, the EERs are significantly higher for the Ambient light conditions than either the Studio-to-Studio or Studio-to-Ambient light conditions.

The same day experiments show that the 'Studio Neutral' condition produces the best EERs and the ambient lighting conditions have the worst EERs.

The last analysis in this section looks at the range of performance for the same day conditions on a ROC. Figure 3 shows the ROC for the 'Studio Neutral' and 'Ambient Neutral' conditions for Algorithm C. The horizontal axis is the false accept rate where the non-match face pairs are identical twins. The vertical axis is the verification rate where the match face pairs are subjects who have an identical twin in the study. The 'Studio Neutral' ROC is an upper bound on Algorithm A's performance and the 'Ambient Neutral' is a lower bound for same day performance. This ROC shows the large range of performance possible for distinguishing between identical twins.

B. Cross Year

This experiment measures the ability to distinguish between identical twins from frontal images taken one year apart. For the cross-year experimental conditions, performance was computed from 24 pairs of identical twins (48 subjects) and 126 pairs of identical twins (256 subjects) for the same-day conditions. Performance is measured under two conditions. The first is comparing two faces when both were taken in the mobile studio (labeled 'Studio-studio in Figure 4). The second compared two images when one was taken in the mobile studio and the other was taken under ambient lighting conditions (labeled 'Studio-ambient'). In the control condition for this experiment, both images were taken on the same day (labeled 'same day'). Performance is reported for the same two experimental conditions in the cross year case. Table II gives one confidence interval for each of the experimental conditions in Figure 4. The 90% confidence intervals are longer for the cross-year conditions because there are fewer subjects in the cross-year condition than the same-day case.

TABLE II

THIS TABLE PROVIDES ONE CONFIDENCE INTERVAL FOR EACH OF THE FOUR EXPERIMENTAL CONDITIONS IN FIGURE 4. THE CONFIDENCE INTERVALS ARE AT A 90% LEVEL FOR THE EER. THE CONFIDENCE INTERVAL PROVIDED IS FOR THE ALGORITHM WITH THE SMALLEST LOWER ENDPOINT.

	Confidence Interval	
Experimental condition	Lower endpoint	Upper endpoint
Studio-studio same day	0.06	0.08
Studio-ambient same day	0.06	0.10
Studio-studio cross-year	0.12	0.21
Studio-ambient cross-year	0.15	0.27

At the 90% confidence level, there is no significant difference in performance for the three algorithms in both cross-year experimental conditions. For Algorithm B, there is a significant difference in EER between the 'Studio-ambient' and 'Studio-studio' case for the cross-year condition. For algorithms A and B, there is a significant drop in performance when going from the same day to cross-year comparisons. For algorithm C, the difference is not significant between the 'Studio-ambient same day' and 'Studio-studio cross-year' conditions. For algorithm C, there was significant differences in the EERs between the 'Studio-studio cross-year' condition and the two same year conditions.

C. Covariates

The next set of experiments look for effects of gender and age on performance. For each covariate, the effect of the covariate is reported for the studio-to-studio and studio-to-ambient lighting conditions. Performance is only reported for images collected on the same day in 2009. There were not enough subjects to measure covariate effects on the cross-year data.

The effect of gender on EER is reported for males and females. Performance is reported in Fig 5 with 90% confidence intervals. With the exception of the result for algorithm A on 'Studio-to-ambient' lighting condition, there was not a significant effect of gender on performance.

The age is broken into categories. The first is over 40 years old (born before 1969) and the second is 40 years old or younger (born 1969 or later). Performance is reported in Fig 6. For the studio-to-studio lighting condition, there is an age effect with performance better on the over 40 year old age group. For the studio-to-ambient lighting condition, there is not an age effect.

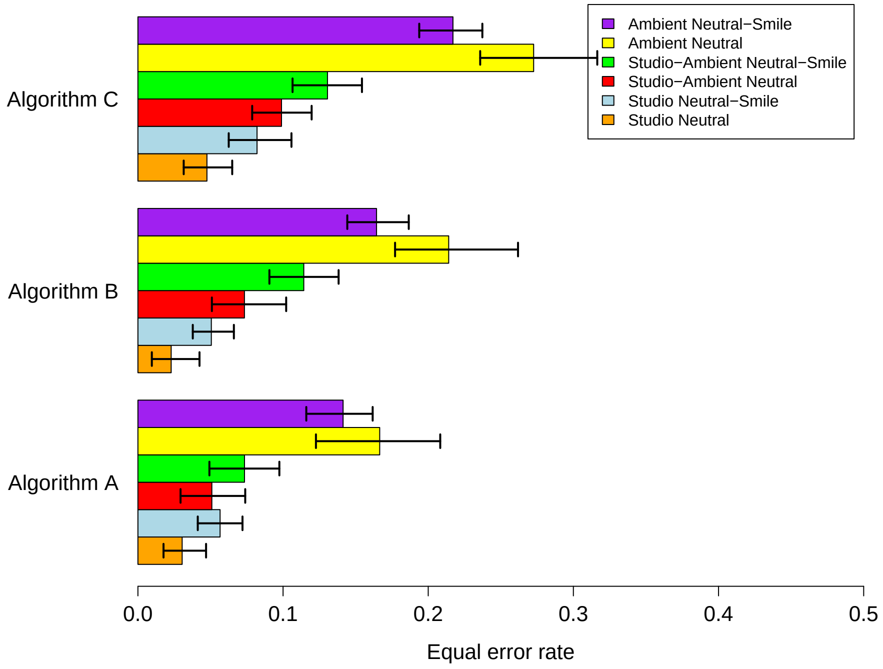

Fig. 2. This barplot summarizes performance when all images for each set of identical twins are collected on the same day. Performance is reported under six different experimental conditions. The EERs are plotted with error-bars for a 90% confidence interval.

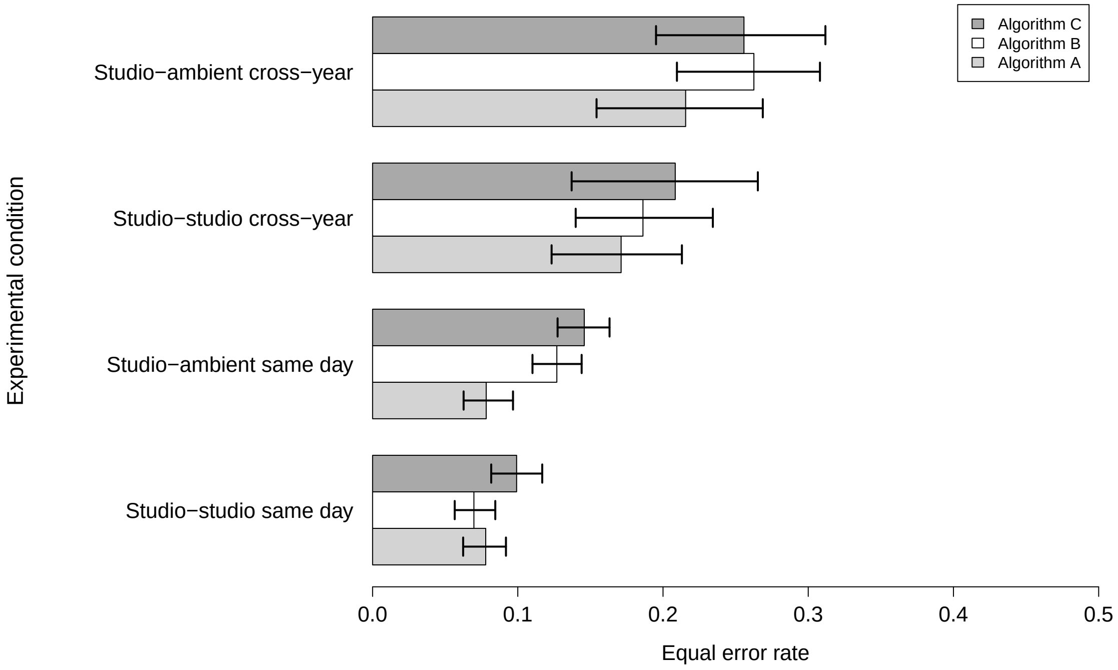

Fig. 4. This barplot shows the effect on performance with a one-year time lapse between images of identical twins. The EERs are plotted with error-bars for a 90% confidence interval. Performance is reported for studio-to-studio lighting matching conditions for images taken on the same day and for images taken a year apart. Performance is reported for similar conditions for studio-to-ambient lighting matching.

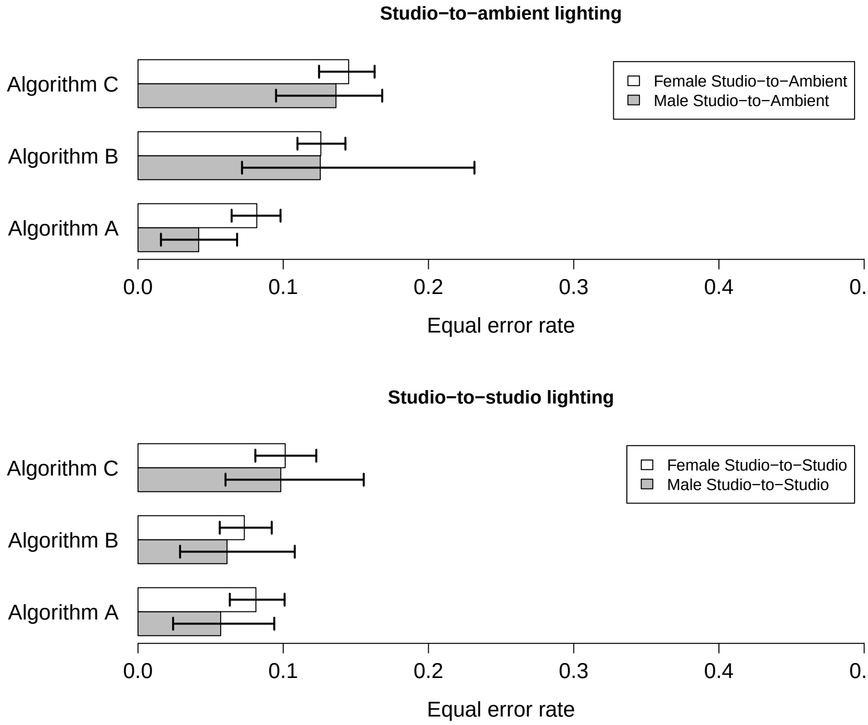

Fig. 5. These two barplots show performance broken out by gender. The top barplot shows performance for the three algorithms for the studio experimental condition. The bottom barplot shows performance for the three algorithms for studio against the ambient lighting condition. Both barplots show error-bars for a 90% confidence interval.

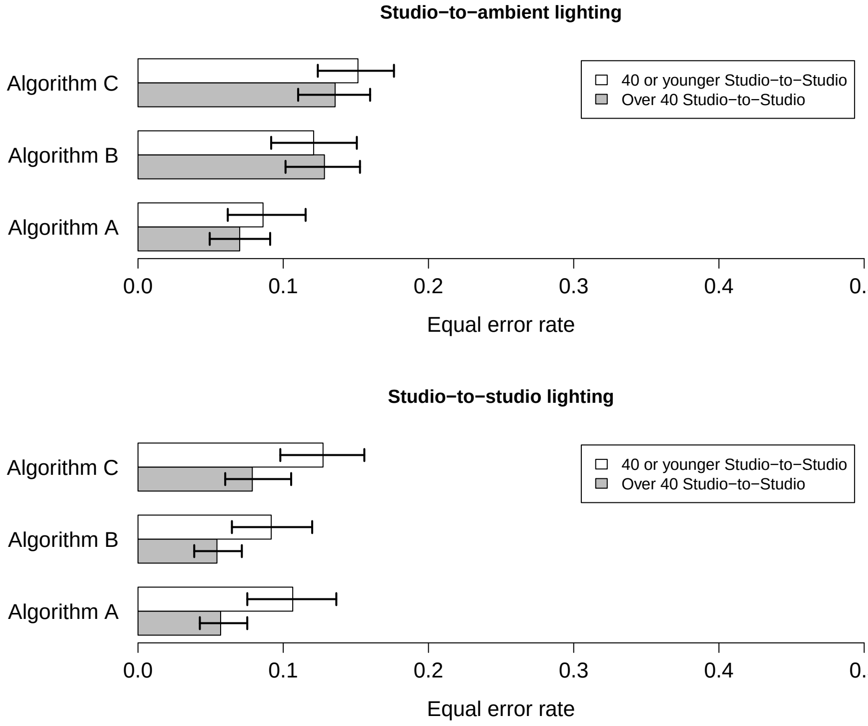

Fig. 6. These two barplots show performance broken out by age. The top barplot shows performance for the three algorithms for the studio experimental condition. The bottom barplot shows performance for the three algorithms for studio against the ambient lighting conditions. Both barplots show error-bars for a 90% confidence interval.

The covariate results for identical twins generally agree with the results for the general population [7]. There is not a gender effect, and older people are easier to recognize.

D. Relationship to Non-twin Performance

The previous experiments have examined the ability of algorithms to distinguish between identical twins. In standard analysis, performance is measured on an algorithm's ability to distinguish between people who are not twins. In this experiment we examine the relationship between the similarity score threshold required to distinguish between identical twins and standard non-match face pairs.

To compare identical twin and standard non-match face pairs, similarity scores were computed from a set of standard non-match pairs. For the studio lighting and neutral expression conditions, the set of standard non-match face pairs consisted of all non-match pairs that are not twins. The EER for the ROC with the identical twin match pairs of faces and the standard non-match face pairs is 0.00. Back-to-back histograms of the all similarity scores distributions are plotted in Figure 7.

A similar analysis was performed for ambient-to-ambient lighting and neutral-to-neutral conditions. In this case, the EER when using standard non-match face pairs is 0.005. Back-to-back histograms of the three corresponding similarity score distributions are plotted in Figure 8.

The histograms in Figures 7 and 8 show significant overlap between the match and non-match distributions for identical twins. By contrast, there is minimal overlap between the match and non-match distributions for the standard non-match cases. This shows that increasing the sensitivity of a system to detect identical twins would result in a substantial increase in the false reject rate.

V. Conclusion

This paper presents the first detailed looked at the ability to distinguish between identical twins. Experimental results measured the performance when faces were collected on the same day and a year apart. The results also measured the effect of changes in expression and lighting. Also, an experiment examined the effect of gender and age on performance.

There was a significant range of performance. The best performance was observed when all images of a pair of identical twins were taken on the same day in the studio environment and the twins had a neutral expression. For this case, the best performing algorithm had a 90% confidence interval for the EER from 0.01 to 0.04. The corresponding confidence interval for the same day with ambient light and neutral expression was 0.12 to 0.21. For cross-year recognition, the best 90% confidence interval for the EER was from 0.15 to 0.27. Performance also showed that gender does not effect performance and that there is an age effect. The results showed that it is easier to distinguish twins over 40 year old than twins under 40.

Our results show that there is promise for distinguishing identical twins under ideal conditions (same day, studio lighting and neutral expression). However, under less than ideal conditions, the problem is very challenging. New research ideas are needed to help improve performance on recognition of identical twins in realistic imaging contexts.

VI. Acknowledgments

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 authors, with the exception of RWVB, acknowledge the support of the Federal Bureau of Investigation.

References

[1] P. J. Grother, G. W. Quinn, and P. J. Phillips, "MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms," National Institute of Standards and Technology, NISTIR 7709, 2010.

[2] Z. Sun, A. A. Paulino, J. Feng, Z. Chai, T. Tan, and A. K. Jain, "A study of multibiometric traits of identical twins," in *Biometric Technology for Human Identification VII*, B. V. K. V. Kumar, S. Prabhakar, and A. A. Ross, Eds., vol. Proc. SPIE 7667, 2010.

[3] [Online]. Available: http://www.twinsdays.org

[4] J. P. Egan, *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

[5] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*. Cambridge: Cambridge University Press, 1991.

[6] R. M. Bolle, N. K. Ratha, and S. Pankanti, "Error analysis of pattern recognition systems—-the subsets bootstrap," *Computer Vision and Image Understanding*, vol. 24, no. 13, pp. 2105–2113, 2003.

[7] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in *Proceedings Third IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2009.

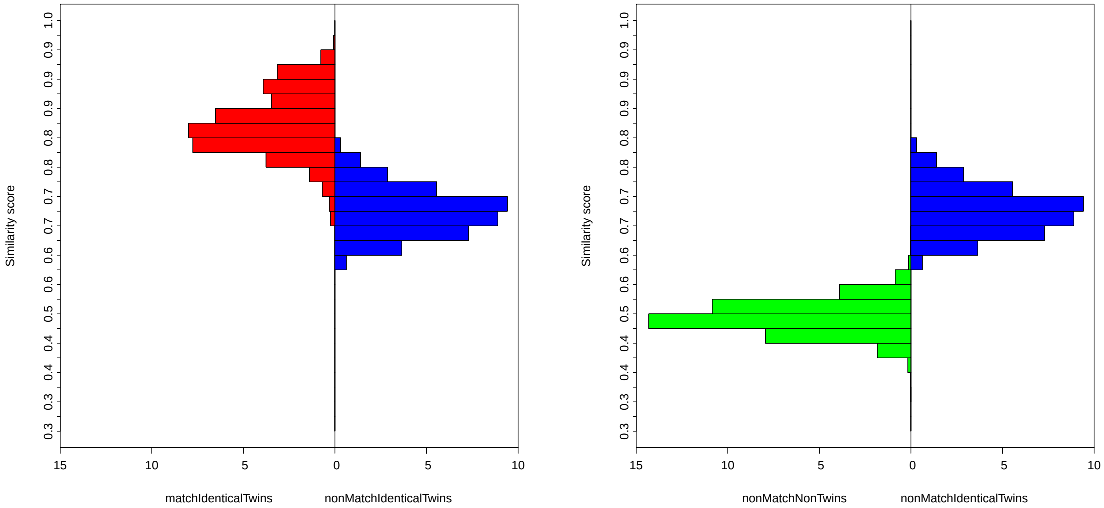

Fig. 7. Back-to-back histograms of the similarity score distributions for Algorithm C for the Studio-to-studio lighting and Neutral-to-neutral expression experiment. All three distributions are for face pairs collected on the same day. The distribution for match face pairs is plotted in red in the histogram on the left (labeled 'matchIdenticalTwins' in this Figure). The distribution for non-match pairs for identical twins is plotted in blue in both histograms (labeled 'nonMatchIdenticalTwins'). The distribution for non-match pairs for non-identical twins (standard non-match face pairs) is plotted in green in the histograms on the left (labeled 'nonMatchNonTwins').

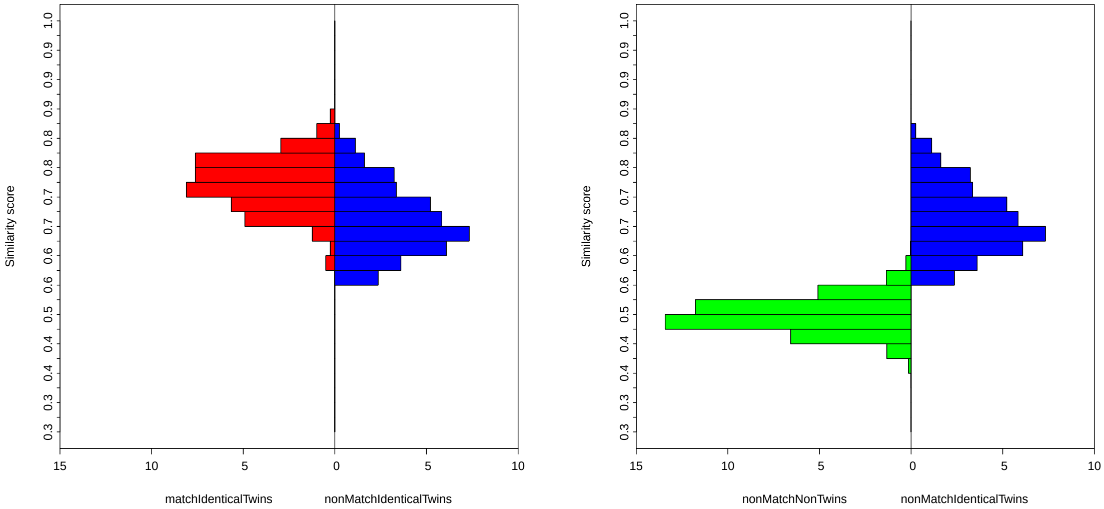

Fig. 8. Back-to-back histograms of the similarity score distributions for Algorithm C for the Ambient-to-ambient lighting and Neutral-to-neutral expression experiment. All three distributions are for face pairs collected on the same day. The distribution for match face pairs is plotted in red in the histogram on the left (labeled 'matchIdenticalTwins' in this Figure). The distribution for non-match pairs for identical twins is plotted in blue in both histograms (labeled 'nonMatchIdenticalTwins'). The distribution for non-match pairs for non-identical twins (standard non-match face pairs) is plotted in green in the histograms on the left (labeled 'nonMatchNonTwins').