A Cross-Sensor Evaluation of Three Commercial Iris Cameras for Iris Biometrics

Ryan Connaughton, Amanda Sgroi, Kevin W. Bowyer, Patrick Flynn
University of Notre Dame
Department of Computer Science and Engineering
Notre Dame, IN 46556
rconnaug,asgroi,kwb,flynn@nd.edu

Abstract

As iris biometrics increasingly becomes a large-scale application, the issue of interoperability between iris sensors becomes an important topic of research. This work presents experiments which compare three commercially available iris sensors and investigates the impact of cross-sensor matching on system performance. The sensors are evaluated using three different iris matching algorithms, and conclusions are drawn regarding the interaction between the sensors and the matching algorithm in a cross-sensor scenario.

1. Introduction

As the field of biometrics grows and becomes a part of daily life, the technology used to capture biometric data also advances. Many more companies are producing sensors for capturing iris data, and the pre-existing companies continue to experiment and improve existing sensors. This poses the question of whether these systems are interoperable. Several studies have investigated the interoperability of both face and fingerprint sensors [10],[8]. Additionally, some researchers have reported on sensor safety, illumination, and ease-of-use for iris recognition systems [4],[9]. Nevertheless, few studies have been conducted to investigate the interoperability of iris sensors from varying manufacturers using multiple available matching algorithms.

This work compares three commercially available iris sensors, observes the effects of cross-sensor comparisons, and investigates the impact of the recognition algorithm on single and cross-sensor performance. Each of the three sensors examined in this work is used for iris enrollment and recognition in the field today. Further, three different algorithms are used to evaluate the recognition rates of the single and cross-sensor comparisons. The use of multiple algorithms provides an unbiased experiment to compare the sensors, as well as offers some insight into the effect the algorithm has on cross-sensor performance.

The remainder of this paper is organized as follows. Section 2 discusses previous work done regarding the study of interoperability of iris sensors. Section 3 describes the experimental setup and the dataset used to analyze the three iris sensors. Results produced by the three matching algorithms are shown and discussed in Section 4, and Section 5 concludes with general observations.

2. Related Work

Because iris biometrics is increasingly becoming a large-scale application in which data is kept and used for long periods of time, the interoperability between iris sensors has become a recent topic of interest. Bowyer et al. investigated cross-sensor and cross-session comparisons using two iris sensors [3]. The authors found that the older of the two systems provided a less desirable match score distribution, which led to an even less desirable cross-sensor match score distribution. They concluded that if the newer sensor were used to collect enrollment data and the older sensor used to collect probe data, the system would achieve higher recognition rates than if the older sensor were used for enrollment and the newer sensor were used to acquire probe data.

In 2005, IBG evaluated the performance of four of the currently available and most widely used iris acquisition and recognition systems [7]. Through four criteria - false accept and reject rate, failure to enroll and acquire, acquisition time and subject usability, and performance over time- the authors evaluated each system in order to report which system the U.S. Department of Homeland Security should employ. The investigation showed that the sensor with the lowest failure to enroll rate had less robust matching over time than the sensor with a higher failure to enroll rate. Overall, the less robust option was recommended due to better performance in other categories of evaluation.

Many factors are known to affect the accuracy of most

iris biometrics systems. These factors include, but are not limited to, pupil dilation [6], time lapse between enrollment and recognition [1], and contact lenses [2].Some sensors may be more or less sensitive to these factors than others, according to the optics, illumination technique, hardware, and software of each sensor. In the IBG evaluation and other reports such as that performed by Du, the illumination schemes, sensor optics, and usability were evaluated [4].

Similarly, the matching algorithm used to perform iris recognition also plays a large role in determining the performance of the system. In 2009 NIST released the IREX report comparing several algorithms using three iris datasets [5]. The IREX evaluation was concerned mostly with algorithm performance regarding timing estimates and expense of computations. However, the authors did report that the choice in the iris recognition algorithm is more influential on the outcome given standardized iris images than in other biometrics, such as fingerprint. Although many algorithms take a Daugman-like approach, they provide different segmentation and preprocessing steps, which affect the outcome of the comparisons.

3. Methods

In this work, three commercially available iris sensors (S1, S2, and S3) are compared in a cross-sensor and cross-session context to evaluate both the performance of each individual sensor and the interoperability between the sensors. Each sensor was used to collect left and right iris images for the same set of subjects over a span of several weeks under a human subjects protocol approved by the Notre Dame Human Subjects Institutional Review Board. Each sensor was mounted on a tripod and adjusted to each subject's height, and the tripods were placed in a row with equal spacing such that all three sensors were equidistant from the ambient visible light sources in the acquisition studio. The controlled environment was designed to provide similar acquisition conditions for each sensor. Subjects approached sensor S1, S2, and S3 in order with little time between each acquisition. Images were acquired using the default settings for each sensor, and each sensor had its own illumination technique and internal image quality control. Additionally, in all acquisition sessions, multiple images were acquired for each subject using each sensor, and the amount of time between successive image acquisitions varied from near-instantaneous to several seconds depending on the sensor. All three sensors produced images of the same size, 640 pixels by 480 pixels.

The final dataset was collected in four acquisition sessions, which spanned a total of 12 weeks. In total, 23,444 iris images were collected, spanning 510 unique subjects (1,020 unique irises). Table 1 shows a detailed breakdown of the number of iris samples collected and subjects involved in each of the four acquisitions sessions. Examples of images of the same iris acquired by sensors S1, S2, and S3, are shown in Figure 1.

Three iris matching algorithms (A1, A2, and A3) were used to compare irises acquired from different acquisition sessions. One of the algorithms is an in-house iris matcher, while the remaining two are commercially available matchers. None of the iris matchers are affiliated with the manufacturers of the sensors. The algorithms were used to compare S1, S2, and S3 in both single-sensor and cross-sensor scenarios.

4. Results

In a preprocessing stage, each of the three matching algorithms was used to extract an iris template from each of the original images. To create these templates, the irises were segmented and features extracted using techniques specific to each algorithm. In some cases, the algorithms were unable to produce templates for particular images. In total, A1 failed to produce templates for 4 images, A2 failed for 17 images, and A3 failed for 84 images. More specifically, A1 failed for 1 image from S1, 1 image from S2, and 2 images from S3. A2 failed for 0 images from S1, 6 images from S2, and 11 images from S3. Finally, A3 failed for 0 images from S1, 66 images from S2, and 18 images from S3. Thus, images from sensor S3 produced the most template failures for algorithms A1 and A2, while sensor S2 produced the most failures for algorithm A3. However, it should be noted that successful template generation does not guarantee correct segmentation.

Table 2 shows the average pupil and iris radius, and the average dilation ratio for each sensor as detected by A2 and A3. This information could not be determined using A1. The dilation ratio is calculated using the equation

$$Dilation = \frac{PupilRadius}{IrisRadius} \tag{1}$$

and in general, a smaller dilation ratio is considered to be better for most traditional iris biometrics systems. In summary, sensor S3 consistently had the smallest pupil and iris radius, while the dilation ratio for S3 was between the ratios for S1 and S2. Between S1 and S2, S1 had both the larger iris radius and the smaller dilation ratio. Additionally, sensor S1 produced the most variation in iris size across the entire dataset.

Algorithms A1, A2, and A3 were each used to compare irises from different acquisition sessions. The matching results presented in this section divide the iris comparisons based on the sensors used to acquired the images. Specifically, there were three single-sensor experiments (S1vS1, S2vS2, and S3vS3), and three cross-sensor experiments (S1vS2, S1vS3, and S2vS3). Experiment S1vS1, for example, compares all pairs of images collected using sen-

Table 1. DETAILED ACQUISITION SUMMARY

	Session 1	Session 2	Session 3	Session 4	Total
S1	2080 Samples 265 Subjects	1671 Samples 212 Subjects	2345 Samples 302 Subjects	2191 Samples 278 Subjects	8287 Samples 491 Subjects
S2	1606 Samples 269 Subjects	1584 Samples 266 Subjects	2457 Samples 302 Subjects	2651 Samples 319 Subjects	8298 Samples 506 Subjects
S3	1579 Samples 269 Subjects	1568 Samples 267 Subjects	1799 Samples 301 Subjects	1913 Samples 320 Subjects	6859 Samples 509 Subjects

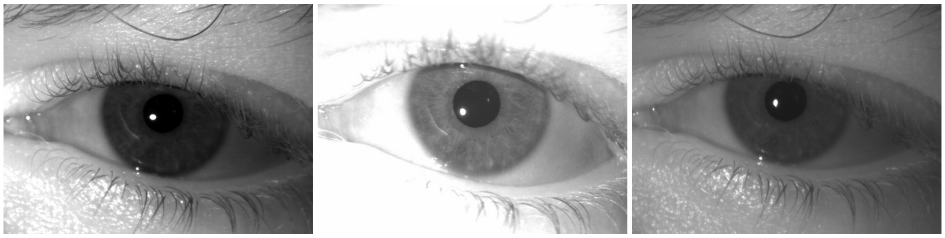

Figure 1. Images of the same iris taken by S1 (**Left**), S2 (**Middle**), and S3 (**Right**)

sor S1, omitting image pairs that originated from the same session. Similarly, experiment S1vS2 compares all images collected using sensor S1 to all images collected using sensor S2, omitting image pairs acquired during the same session. Each experiment was repeated using each of the three matching algorithms.

The number of match and non-match comparisons used in each experiment varied depending on both the sensor and matching algorithm. In some cases, the matching algorithms filtered out comparisons which could not generate match scores above a particular confidence threshold; these comparisons are omitted from the experiment results presented in this section. The total number of match and non-match comparisons used in each experiment are shown in Table 3.

The match and non-match score distributions for all experiments using algorithm A1 are shown in Figure 2. The output of algorithm A1 is a similarity score for each iris comparison, where a higher score indicates a better match. In Figure 2, the non-match score distributions for each experiment are nearly identical. From this graph, it can be seen that the match score distributions for single-sensor experiments were generally further to the right than the cross-sensor match score distributions. Figure 3 shows the ROC curves for all experiments using A1. Comparing the ROC curves in Figure 3 at a FAR=0.01, S1vS1 performed the best and S3vS3 performed the worst of the single-sensor experiments, and the cross-sensor experiments achieved TAR's that fall between those of the corresponding single-sensor experiments.

Figure 4 shows the match and non-match score distributions for all experiments using algorithm A2. The output of algorithm A2 is a distance score for each iris comparison, where a lower score indicates a better match. While all of match score distributions are near zero, the match distributions from the cross-sensor experiments consistently extend further to the right than the single-sensor experiments. Figure 5 shows the ROC curves for the experiments using algorithm A2. Comparing these ROC curves at FAR=0.01, S2vS2 performed the best and S3vS3 performed the worst of the single-sensor experiments. Also, unlike the results from A1, for A2 the ROC curves of the cross-sensor experiments do not strictly fall between the ROC curves of the corresponding same-sensor experiments. The performance of S1vS2 was consistently worse than the performance of either S1vS1 or S2vS2.

Finally, the match and non-match score distributions for the experiments using A3 are shown in Figure 6. The output of A3, like A2, is a distance score, so a lower score indicates a better match. The match score distributions for the single-sensor experiments are generally shifted to the left of the cross-sensor experiments in Figure 6. Figure 7 shows the ROC curves for the experiments using algorithm A3. Comparing the ROC's at FAR=0.01, S1vS1 performed the best and S3vS3 performed the worst of the single-sensor comparisons. Interestingly, S1vS3 actually achieved the second best performance of all of the experiments using A3, and S2vS3 performed the worst.

Evaluating these results in the context of a sensor comparison, it is clear that images from sensor S3 performed the worst of the three sensors for all three algorithms in a single-sensor scenario. Out of the three single-sensor experiments considered, sensor S3 had the lowest performance at FAR=0.01 for all three algorithms. The relatively lower performance of sensor S3 may be explained by Table 2, which shows that the average iris radius was generally

Table 2. AVERAGE PUPIL RADIUS, IRIS RADIUS, AND DILATION RATIO

	A2	A3
S1	Pupil = 44.6 pixels ($\sigma = 6.9$) Iris = 130.8 pixels ($\sigma = 10.3$) Dilation = 0.342 ($\sigma = 0.051$)	Pupil = 45.8 pixels ($\sigma = 6.8$) Iris = 134.0 pixels(($\sigma = 11.5$) Dilation = 0.343 ($\sigma = 0.052$)
S2	Pupil = 47.6 pixels ($\sigma = 8.2$) Iris = 124.3 pixels ($\sigma = 8.6$) Dilation = 0.383 ($\sigma = 0.062$)	Pupil = 48.6 pixels ($\sigma = 8.5$) Iris = 125.1 pixels ($\sigma = 8.7$) Dilation = 0.389 ($\sigma = 0.065$)
S3	Pupil = 43.2 pixels ($\sigma = 7.8$) Iris = 115.8 pixels ($\sigma = 7.5$) Dilation = 0.373 ($\sigma = 0.066$)	Pupil = 44.7 pixels ($\sigma = 7.6$) Iris = 118.2 pixels ($\sigma = 8.6$) Dilation = 0.379 ($\sigma = 0.064$)

Table 3. NUMBER OF MATCH AND NON-MATCH COMPARISONS IN EACH EXPERIMENT

	A1	A2	A3
S1vS1	Match = 28,207 Non-Match = 12,783,542	Match = 28,207 Non-Match = 12,786,533	Match = 28,188 Non-Match = 12,772,638
S2vS2	Match = 27,485 Non-Match = 12,645,155	Match = 27,453 Non-Match = 12,629,379	Match = 26,835 Non-Match = 12,330,790
S3vS3	Match = 18,921 Non-Match = 8,775,488	Match = 18,903 Non-Match = 8,752,069	Match = 18,836 Non-Match = 8,732,954
S1vS2	Match = 55,534 Non-Match = 25,542,660	Match = 55,494 Non-Match = 25,530,015	Match = 54,817 Non-Match = 25,199,025
S1vS3	Match = 46,051 Non-Match = 21,208,102	Match = 46,031 Non-Match = 21,181,421	Match = 45,927 Non-Match = 21,123,435
S2vS3	Match = 45,772 Non-Match = 21,147,811	Match = 45,721 Non-Match = 21,105,682	Match = 45,071 Non-Match = 20,848,573

smaller than the irises acquired using S1 and S2. Interestingly, S1vS3 had the best of the cross-sensor performances using algorithm A3, achieving better performance than two of the single-sensor experiments as well. In both the single-sensor and cross-sensor scenarios, Sensor S1 achieved the best relative performance using algorithm A1, and sensor S2 achieved the best relative performance using algorithm A2.

Evaluating the results to summarize the impact of using cross-sensor comparisons, several conclusions can be made. From the previously presented score distributions and ROC curves, it is clear that the relative performance of the sensors is sensitive to the algorithm choice. In other words, when designing an iris biometrics system, sensor selection should not be made independent of the algorithm choice; instead, the two factors should be evaluated in combination. Additionally, the performance of cross-sensor comparisons is also dependent on both the sensors and the matching algorithm. In the experiments using A1, for example, the TAR of the cross-sensor experiments consistently fell between the TAR of the corresponding single-sensor experiments (at FAR=0.01). This might suggest that given a system which currently uses an older sensor, the introduction of a newer sensor that achieves higher performance may be able to increase overall system performance even if cross-sensor comparisons are used to utilize legacy data. However, the experiments using algorithm A2 show that this is not always the case. Using A2, S2vS2 consistently performed better than S1vS1, but cross-sensor comparisons between S1 and S2 actually degraded performance below the performance of either single-sensor experiment.

Further, the relative performance of single-sensor comparisons is not necessarily a reliable predictor of the relative performance of cross-sensor comparisons. In the experiments presented in this work, we find that the performance of the same-sensor experiments is a reliable predictor of cross-sensor performance for algorithms A1 and A2, but not for algorithm A3. For example, using algorithm A1, the relative order of same-sensor performance (from best to worst at FAR=0.01) was S1vS1, S2vS2, and S3vS3. Using the same algorithm, S1vS2 was the best of the cross-sensor experiments, followed by S1vS3, and finally S2vS3. In algorithm A2, the relative order of same-sensor performance was different (S2vS2, S1vS1, and then S3vS3); however, this relative order was preserved in the cross-sensor experiments, where the relative order was S1vS2, S2vS3, and S1vS3. Unlike algorithms A1 and A2, A3 did not preserve this relative ordering. For A3, the same-sensor per-

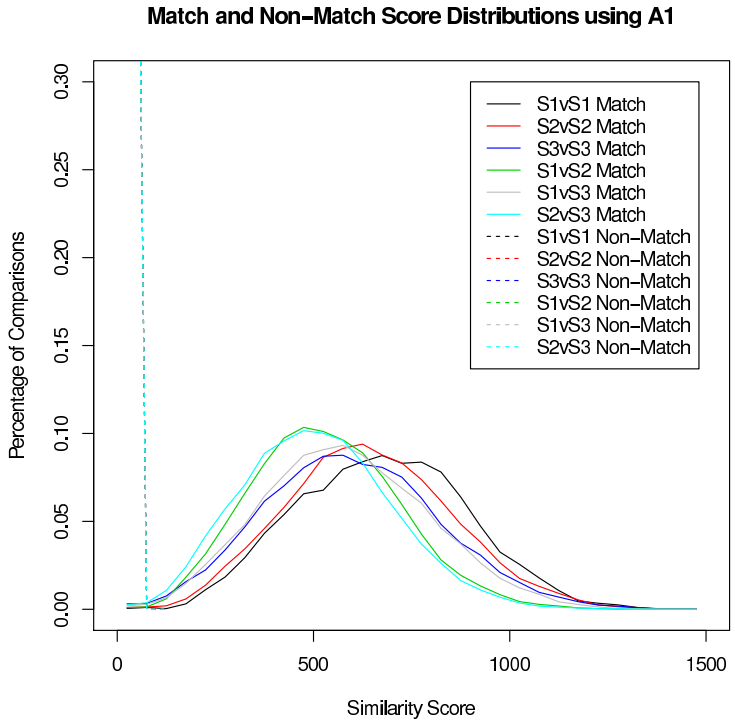

Figure 2. Match and non-match score distributions for all experiments using algorithm A1. A higher score indicates a better match. The non-match score distributions are nearly identical.

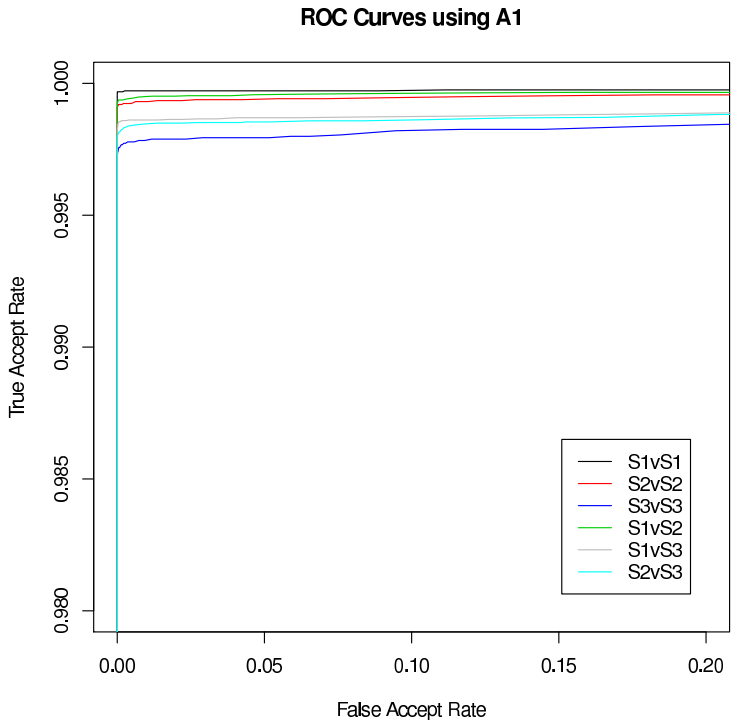

Figure 3. ROC curves for all experiments using algorithm A1.

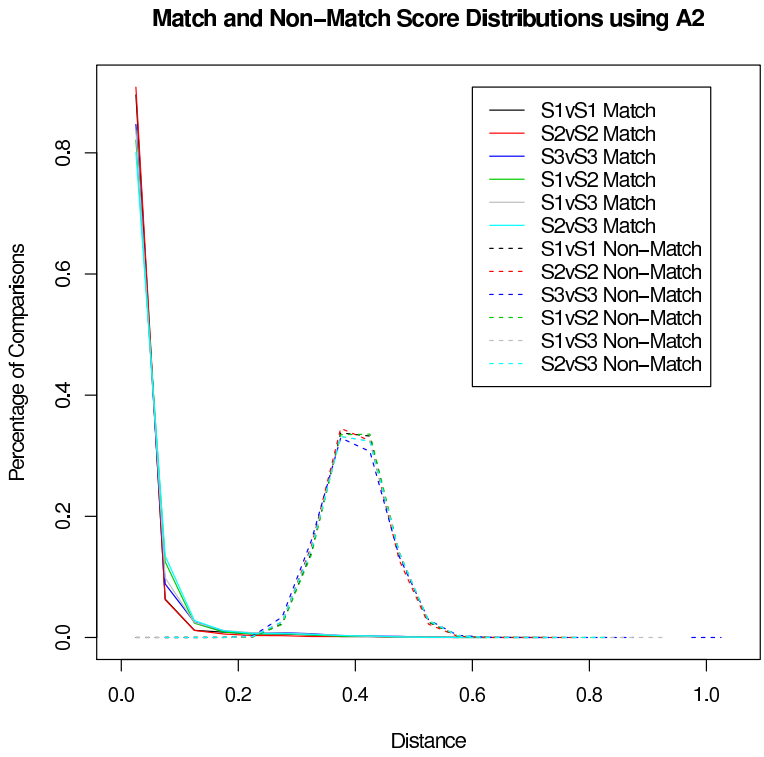

Figure 4. Match and non-match score distributions for all experiments using algorithm A2. A lower score indicates a better match.

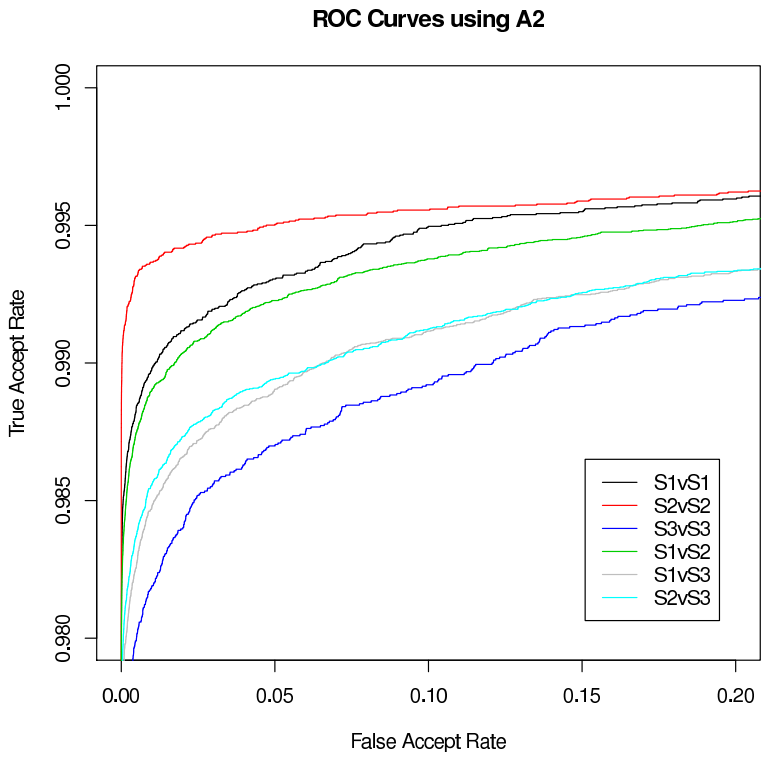

Figure 5. ROC curves for all experiments using algorithm A2.

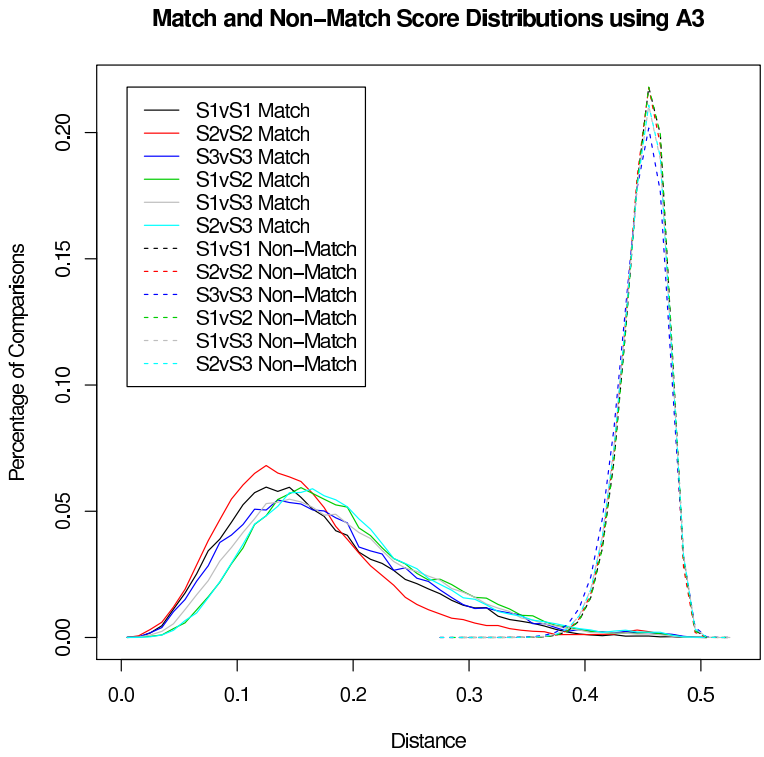

Figure 6. Match and non-match score distributions for all experiments using algorithm A3. A lower score indicates a better match.

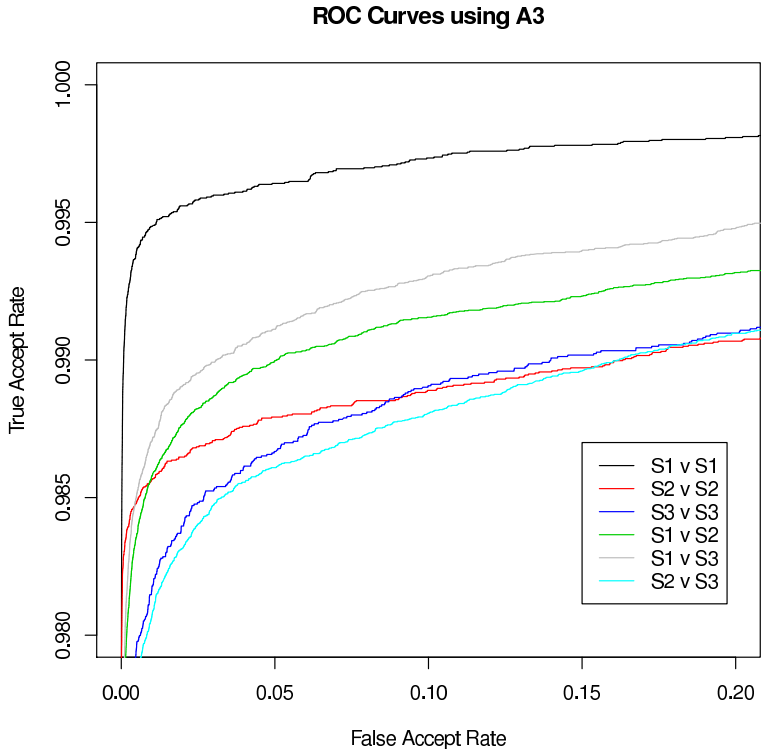

Figure 7. ROC curves for all experiments using algorithm A3.

formances from best to worst at FAR=0.01 were S1vS1, S2vS2, and S3vS3. The cross-sensor rankings, however, were S1vS3, S1vS2, and S2vS3. This may suggest that algorithm A3 is more sensitive to differences in particular acquisition factors (e.g. pupil dilation/size, illumination technique, specular highlights), and that these factors experience the greatest variability in a cross-sensor scenario.

5. Conclusion

In this work, we present a study designed to compare three commercially available iris cameras for iris biometrics. In addition to comparing the isolated performance of each of the three sensors (S1, S2, and S3), we analyze the performance of cross-sensor comparisons for the three sensors as well. Data was collected for the same subjects using each of the three sensors, and the iris images are compared using three different matching algorithms (A1, A2, and A3).

Summarizing the results of the sensor comparisons we found that sensor S1 performed the best under algorithms A1 and A3 at FAR=0.01, while Sensor S2 performed best under algorithm A2. Sensor S3 consistently performed the worst of the three sensors in the single-sensor experiments, although the cross-sensor experiment between S1 and S3 outperformed all but one of the single-sensor experiments and all other cross-sensor experiments using algorithm A2. Across all experiments using all sensors and matchers, the best performance at FAR=0.01 was achieved by algorithm A1, for which sensor S1 was the strongest of the three sensors.

We also draw the following conclusions from the cross-sensor experiment results:

- When selecting a sensor and algorithm for a biometric system, the two components should be evaluated in combination, rather than independently.
- The relative performance of cross-sensor experiments compared to the single-sensor experiments using the same sensors is dependent on both the sensors and the matching algorithm. In some cases, introducing a higher quality sensor to be used for cross-sensor comparisons may degrade performance rather than improve it.
- The relative performance of single-sensor comparisons is not necessarily a reliable predictor of the relative performance of cross-sensor comparisons using the same sensors. Thus, evaluating a new sensor on images it acquires may not predict cross-sensor compatibility to an older sensor.

In future work, we plan to investigate whether particular acquisition and sensor characteristics (e.g. illumination technique, pupil dilation, specular highlights) have greater impacts on cross-sensor comparison performance than others.

6. Acknowledgments

This work is sponsored under IARPA BAA 09-02 through the Army Research Laboratory and was accomplished under Cooperative Agreement Number 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 IARPA, 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.

References

[1] S. E. Baker, K. W. Bowyer, and P. J. Flynn. Empirical evidence for correct iris match score degradation with increased time-lapse between gallery and probe matches. *Advances in Biometrics*, 5558, 2009.

[2] S. E. Baker, K. W. Bowyer, P. J. Flynn, and P. J. Phillips. Degradation of iris recognition performance due to non-cosmetic prescription contact lenses. In *Computer Vision and Image Understanding*, volume 114, September 2010.

[3] K. W. Bowyer, S. E. Baker, A. Hentz, K. Hollingsworth, T. Peters, and P. J. Flynn. Factors that degrade the match distribution is iris biometrics. In *Identity in the Information Society*, volume 2, pages 327–343. Springer, 2010.

[4] Y. E. Du. Review of iris recognition: Cameras, systems, and their applications. *Sensor Review*, 26(1):66–69, 2006.

[5] P. Grother, E. Tabassi, G. W. Quinn, and W. Salamon. Irex 1: Performance of iris recognition algorithms on standard images. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=903606, 2009.

[6] K. Hollingsworth, K. W. Bowyer, and P. J. Flynn. Pupil dilation degrades iris biometric performance. In *Computer Vision and Image Understanding*, volume 113, pages 150–157. 2009.

[7] IBG. Independent testing of iris recognition technology. http://www.biometricscatalog.org/itirt/itirt-FinalReport.pdf, May 2005.

[8] S. K. Modi. Analysis of fingerprint sensor interoperability on system performance. http://www.cerias.purdue.edu/news_and_events/events/security_seminar/details.php?uid=it58pkph77c0lrrbathkqhn5lo@google.com, August 2008.

[9] E. Newton and P. Phillips. Meta-analysis of third-party evaluations of irisrecognition. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 39:4–11, January 2009.

[10] A. Ross and A. Jain. Biometric sensor interoperability: A case study in fingerprints. *Proceedings of the International ECCV Workshop on Biometric Authentication (BioAW)*, 3087:134–145, May 2004.