Overview of

Meta-Analysis of Third-Party Evaluations of Iris Recognition

Elaine M. Newton, *Member, IEEE* and P. Jonathon Phillips, *Senior Member, IEEE*

Abstract— **Iris recognition has long been widely regarded as a highly accurate biometric, despite the lack of independent, large-scale testing of its performance. Recently, however, three third-party evaluations of iris recognition were performed. This paper compares and contrasts the results of these independent evaluations. We find that despite differences in methods, hardware, and/or software, all three studies report error rates of the same order of magnitude: observed false non-match rates (FNMRs) from 0.0122 to 0.03847 at a false match rate (FMR) of 0.001. Further, the differences between the best performers' error rates are an order of magnitude smaller than the observed error rates.**

I. Introduction

DESPITE the prior lack of third-party testing of iris matching recognition, the conventional wisdom in the biometrics community has been that iris recognition is highly accurate – even the most accurate biometric. One of many examples of this belief is a comparative table in a seminal biometrics book, which ranks various types of biometrics' abilities based on the perception of three biometrics experts [1]. The table ranks the iris biometric as having "High" performance, along with DNA, fingerprint, and retina. Another example is this statement from a biometric newsletter: "There is no denying that iris recognition is the most accurate biometric technology,…" [2].

Between May 2005 and March 2007, three major tests on iris recognition were released – the first of their kind. These tests were the Independent Testing of Iris Recognition Technology (ITIRT) conducted by the International Biometric Group (IBG) [3], the Iris RecognItion Study 2006 (IRIS06) conducted by Authenti-Corp (AC) [4], and the Iris Challenge Evaluation (ICE 2006) conducted by the National Institute of Standards and Technology (NIST) [5].

This paper gives an overview of [6], which discusses these evaluations, their similarities and differences, and most importantly summarizes performance across the evaluations. To compare performance across evaluations, performance statistics selected for this meta-analysis took into account evaluation type, failure to enroll and failure to acquire, sensor quality software, and subject variability. Based on the selection criteria, across all three evaluations, reported false non-match rate (FNMR) at a false match rate (FMR) of 0.001 ranged from 0.0122 to 0.03847. At an FMR of 0.001, the range of FNMR for the best performers in each test was 0.0122 to 0.0175.

E. M. Newton and P. J. Phillips are with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (corresponding author Newton's phone: 301-975-2532; fax: 301-975-5287; e-mail: enewton@.nist.gov; Phillips' e-mail: jonathon@nist.gov).

II. Background

A. ITIRT Study

IBG's ITIRT study was funded by the US Department of Homeland Security (DHS) and began in July 2004. Final results were released in May 2005 [3].

IBG tested match rates, enrollment and acquisition failure rates, interoperability, and level of effort needed for transactions using three sensors: Panasonic BM-ET300, Oki Irispass, and LG 3000. Results of these tests include "cross-visit recognition," which is a one-to-one comparison of an enrollment iris template from an initial visit against the iris template captured during a second visit. Template matching software from Iridian performed matching tests on the collected biometric samples.

B. IRIS06 Study

Authenti-Corp's study was funded jointly by the US Department of Justice, National Institute of Justice (NIJ) and the US DHS Transportation Security Administration (DHS/TSA) and kicked off in December 2005. The draft final report was released in March 2007 [4].

Authenti-corp also tested match rates, enrollment and acquisition failure rates, and level of effort needed for transactions using three sensors, but it did not identify the tested sensors, simply referring to them as Products A, B, and C. IRIS06 utilized a matching algorithm provided by Professor Daugman of the University of Cambridge.

Authenti-corp provided the authors with a breakdown of how many subjects and iris samples were used in testing for the results used in this analysis, as provided in Table I. The report provides performance results from both visits combined in the form of false non-match rates with upper and lower 95% confidence intervals (CIs).

C. ICE 2006

The ICE 2006 study conducted by NIST was funded jointly by the US DHS's Science and Technology Department and TSA, the US Director of National Intelligence's Information Technology Innovation Center, the US Federal Bureau of Investigation, the NIJ, and the Technical Support Working Group (TSWG). The study began in December 2003. The final report was released in March 2007 [5].

The ICE 2006 reported error rates for the left and right irises separately for three different matchers using the same images from a single sensor for data collection, at a single operating point – false accept rate (FAR) = 0.001. The matching algorithms tested were supplied by Sagem-Iridian (SG-2), Iritech (Irtch-2), and Cambridge (Cam-2). A modified LG EOU 2200 was used to collect iris images from 240 subjects. The LG EOU 2200 was modified so that the automatic quality check for capturing iris images was overridden. This allowed for up to two out of three captured images not to meet the built-in quality checks.

ICE 2006 divided its test set into 30 random test sets and reported results via boxplots for each algorithm, reporting a maximum, third quartile, median, first quartile, and minimum false reject rate (FRR).

III. Methods

In this section, we summarize the rationale for choosing the points of comparison from each test, which included issues of how quality, enrollment and acquisition, glasses, and timing between visits were handled.

Table I summarizes the algorithms, sensors, and the total number of subjects and biometric samples (both genuine and impostor) used in each of the tests used in this analysis.

All studies used volunteer test subjects who were informed of the testing taking place. Each study used attendees to operate equipment to capture subjects' biometric samples. All studies also collected iris images indoors in a fixed environment within each study (but not necessarily the same across the studies).

While there is not enough information to discuss time to match irises across all three studies, none of the studies indicated an imposed timing constraint for offline matching experiments.

It is important to note that the Iridian and University of Cambridge algorithms are all based on John Daugman's work. In other words, all of the matching algorithms here except one – Iritech's algorithm in the ICE 2006 evaluation – have the same genesis.

A. Types of Errors

This paper discusses four types of errors, some of which were defined differently across studies. They are defined for the purposes of this paper as follows:

The false match rate (FMR) is the rate at which a matching algorithm incorrectly determines that an impostor's biometric sample matches an enrolled sample.

The false non-match rate (FNMR) is the rate at which a matching algorithm incorrectly fails to determine that a genuine sample matches an enrolled sample.

The failure to enroll rate (FTE) is the rate at which a biometric system fails to enroll a subject's biometric sample.

The failure to acquire rate (FTA) is the rate at which a biometric system fails to capture a subject's biometric sample for the purpose of recognition of the subject.

Note that neither the FMR nor the FNMR include FTE or FTA in their definitions. FMR and FNMR are strictly statistics of the capabilities of a matching algorithm. Further, the ICE 2006 study reports false accept rates (FARs) and false reject rates (FRRs) in its study, where the terms FMR and FNMR as defined above are more apropos given that FTE and FTA are not taken into account. This is further discussed in the Quality section below.

B. Comparison at FMR = 0.001

Given the number of samples in the studies as well as the data presented in the main body of the ICE 2006 report, we chose to compare these studies by comparing the FNMRs at the point that FMR is 0.001.

ICE 2006 did not take into account FTEs or FTAs. Hence, the results reported as false reject rates (FRRs) and false accept rates (FARs) are treated here to be the same as FNMR and FMR, respectively.

Numerical results for the FNMR figures in the boxplots at (FMR of 0.001) for ICE 2006 were made available for this paper. Results for the left and right iris were averaged together for the purpose of comparison with the other studies, which did not report results separately for the left and right irises.

ITIRT's results are listed by Hamming distance. FNMR results were collected for cross-visit recognition where the FMR was nearly 0.001. When it was unclear which FMR to choose, we chose the higher FMR value (hence, the lower FNMR).

IRIS06 results are presented graphically in the report, but for our analysis, Authenti-corp provided the exact results in Figure 1.

IV. Results

The results of this paper are presented graphically by comparing the observed FNMRs. Additional anaylsis and discussion is available in [6].

Fig. 1 shows the results of the three tests in order of when the studies began: oldest on the left (ICE 2006, ITIRT in the middle, and most recent on the right (IRIS06).

ICE 2006 results were presented in box plots. The horizontal line in the middle of the box is the median. The top and bottom of the box correspond to the 1st quartile

(25th percentile) and 3rd quartile (75th percentile) values of the observations, respectively. The dashed lines above and below the box, called "whiskers," end with a short horizontal line, which mark minimum and maximum data values.

ITIRT results are single values.

IRIS06 results give performance values with a range of estimated uncertainty in the form of 95% confidence intervals, computed using the logit beta-binomial method for the results used here.

Collectively, these data points range from 0.00473 to 0.0465. The observed FNMR values range from 0.0122 (ICE 2006, SI-2) to 0.03847 (ITIRT, LG 3000).

Fig. 1 suggests two conclusions about the range of results observed over three evaluations. First, the difference among the best performers in each evaluation is an order of magnitude less than the observed performance. The best performer had an FNMR of approximately 0.01 at an FMR of 0.001, and the differences among the three best FNMRs were on the order of 0.001. Second, the range of FNMRs for all performers and the differences among all performance statistics was of the same order of magnitude.

V. Conclusion

One of the hallmarks of science is the repeatability of experiments. Here we have compared experiments from three independent sets of third-party testers. They independently designed tests, collected data from different populations, and conducted experiments on iris recognition systems. Despite the differences in collection efforts, sensors, matching algorithms, protocols, and other factors, these three tests produced consistent results and demonstrate repeatability.

Because of the strong agreement among these tests, we conclude that these evaluations represent an accurate assessment of the state of the art in iris recognition as of Spring 2006. (Spring 2006 marks the last algorithm submission within this set of evaluations.)

There are two possible reasons why the results of these evaluations are so similar. First, all but one of the tested algorithms was based on the work of Professor John Daugman. Second, because there is a symbiotic relationship between the development of sensors and iris recognition algorithms, the dominance of the Daugman-based algorithms in the market place may also decrease variation in the output of different sensors.

The FNMRs examined here are all the same order of magnitude, and as observed in Fig. 1, there is a fair amount of overlap of the boxplots of ICE 2006 and the CI's of IRIS06. The mean differences show that the difference between each tests' data points is no greater than the order of magnitude of the FNMRs. Further, the differences between the best performers (i.e., the lowest FNMR scores) for all test-pairs is an order of magnitude less than the error rates.

Acknowledgment

The authors thank Roger Cottam and Valorie Valencia of Authenti-Corp and Michael Thieme of IBG for reviewing this paper.

References

[1] A.K. Jain, R. Bolle, and S. Pankanti, "Introduction to biometrics," in *Biometrics: Personal Identification in Networked Society*, A.K. Jain, R. Bolle, and S. Pankanti, eds., Boston: Kluwer Academic, 1999, pp. 1-41.

[2] M. Lockie, "Comment," *Biometric Technology Today*, p. 12, November/Decmeber 2004.

[3] International Biometric Group. (2005, May). Independent testing of iris recognition technology (ITIRT) - Final Report. [Online]. Available: http://www.ibgweb.com/reports/public/ITIRT.html.

[4] Authenti-Corp. (2007, Mar. 31). Draft Final Report: Iris Recognition Study 2006 (IRIS06). version 0.40. [Online]. Available: http://www.authenti-corp.com/iris06/report/IRIS06_draft_report_v0-40p_20070331.doc

[5] P. J. Phillips, *et. al.*(2007, Mar.) FRVT 2006 and ICE 2006 Large-Scale Results. National Institute of Standards and Technology. Gaithersburg, MD. Technical Report NISTIR 7408. [Online]. Available: http://iris.nist.gov/ice/FRVT2006andICE2006LargeScaleReport.pdf

[6] E. M. Newton and P. J. Phillips, "Meta-Analysis of Third-Party Evaluations of Iris Recognition," National Institute of Standards and Technology, Gaithersburg, MD, Technical Report NISTIR 7440, Aug. 2007.

TABLE I
SUMMARY OF TEST SENSORS, ALGORITHMS, AND NUMBER OF BIOMETRIC SAMPLES USED IN EACH EVALUATION.

Evaluation	Sensor	Matching Algorithm	Total Number of Tested: Subjects / Samples
ICE 2006	LG EOU 2200	Sagem-Iridian (SG-2)	240 / 59,558 (left & right eyes)
		Iritech (Irtch-2),	
		Cambridge (Cam-2)	
ITIRT	Panasonic BM-ET300 (Pan)	Iridian's KnoWho OEM SDK v3.0	458 / 12,238 (left & right eyes)
	OKI IRISPASS-WG (OKI)		458 / 12,587 (left & right eyes)
	LG IrisAccess 3000 EOU & ROU (LG)		458 / 16,826 (left & right eyes)
IRIS06	Product A	Daugman algorithm	285 / 4397 (left & right eyes)
	Product B		285 / 4467 (left & right eyes)
	Product C		284 / 4696 (left & right eyes)

FIG. 1
FNMR RESULTS FROM THREE STUDIES (AT FMR = 0.001). ICE 2006 REPORTS RESULTS FOR ALGORITHMS ON BOXPLOTS; ITIRT REPORTS A SINGLE PERFORMANCE STATISTIC FOR EACH SENSOR; AND IRIS 06 REPORTS ESTIMATED FNMR WITH A 95% CONFIDENCE INTERVALS.