A Baseline for Assessing Biometrics Performance Robustness: A Case Study across Seven Iris Datasets

Yooyoung Lee, James J. Filliben, Ross J. Micheals, Michael D. Garris, and P. Jonathon Phillips
Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD, 20899 USA
{yooyoung, filliben, rossm, mgarris, jonathon}@nist.gov

Abstract

We examine the robustness of algorithm performance over multiple datasets collected with different sensors. This study provides insight as to whether an algorithm performance derived from traditional controlled environment studies will robustly extrapolate to more challenging stand-off/real-world environments. We argue that a systematic methodology is critical in assuring the validity of algorithmic conclusions over the broader arena of applications. We present a structured evaluation protocol and demonstrate its utility by comparing the performance of an open-source algorithm over seven diverse datasets, spanning six different sensors (three stationary, one handheld, and two stand-off). We also provide baseline results for the ranking of the seven datasets as measured by four performance metrics. Finally, we compare our protocol-based ranking with a parallel ranking based on an independent survey of biometrics experts, with high correlation between the two rankings being demonstrated.

1. Introduction

As the iris-based biometrics technology grows, a broad array of iris sensors (e.g., stationary, handheld, stand-off) have been introduced and numerous recent evaluations of automated recognition algorithms have been done [1-10].

These evaluations primarily focused on performance competitions of multiple algorithms. A complementary evaluation would be to characterize the robustness of an algorithm performance across the multiple datasets with a broad range of sensor types. To date, algorithmic robustness studies have been limited by two key factors. The first is the lack of the algorithm's capability to accommodate images captured from the full range of iris sensors. The second is the utilization of only a small number of datasets with narrow image quality diversity (e.g., datasets with only near-ideal iris images captured in well controlled environments with cooperative subjects).

In this regard, our study quantifies the robustness of algorithm performance across seven diverse iris datasets collected from the six different sensors—the seven datasets consist of ICE2005[1], two subsets from MBGC2008 [2], and four subsets from NDSpring2011[3][4]. Our study also examine whether algorithm performance based on traditional controlled environment studies will extrapolate to stand-off/real-world environments (image data severity). As a case study, we use the open-source research algorithm, VASIR (Video-based Automatic System for Iris Recognition)—that can process a broad range of sensor types—to provide the baseline performance results needed for other research evaluation purposes. We demonstrate the analysis using a structured protocol for carrying out robustness assessment. In addition, we determine if correlation exists between VASIR's performance ranking and expert survey-based ranking about such sensor performance.

The specific contributions of this paper are as follows:

1) **Datasets:** provide ranked list of seven datasets consisting of six sensors (three stationary, one handheld, and two stand-off)
2) **Baseline:** provide baseline performance results using the open-source algorithm VASIR that has capability of handling all three sensor categories.
3) **Methodology:** demonstrate the analysis methodology using a structured evaluation protocol for carrying out robustness assessment of a biometrics system
4) **Experts:** determine the degree of correlation between the resulting VASIR performance ranking and expert opinion ranking over the six sensors.

For the iris-based biometrics research and academic community, our evaluation baseline is necessary to rapidly examine the performance of other algorithms robustness over multiple datasets, taken by multiple sensors, and/or collected under various environments. This performance evaluation baseline via an open-source algorithm will give researchers an opportunity for giving insight of the algorithm, for comparing components/modules of their algorithms, and for educationally advancing biometrics technology in general.

2. Related Work

Table 1 shows the comparison of previous studies and our study—# algs & # orgs defines the total number of algorithms provided by the total number of organizations.

Table 1 Comparison of previous evaluations and our study

Study	Year published	Study type	# algs & # orgs	# sensors	# datasets	Sensor category		
						Stationary	Handheld	Stand-off
ITIRT [5]	2005	Evaluation	1 & 1	3	1	o		
IRIS06 [6]	2007	Evaluation	1 & 1	3	1	o		
ICE2005[1]	2008	Challenge	13 & 10	1	1	o		
MBGC [2]	2009	Challenge	0 & 0	3	3	o		o
Bowyer et al [7]	2009	Cross/Experiment	1 & 1	2	2	o		
IREX-I [8]	2009	Evaluation	19 & 10	3	3	o		
ICE2006[9]	2010	Evaluation	3 & 3	1	1	o		
Connaughton et al [4]	2012	Cross/Experiment	3 & 3	3	3	o	o	
IREX-III [10]	2012	Evaluation	82 & 11	3*	1	o	o	
Our study	2013	Experiment	1 & 1	6	7	o	o	o

* This collection used primarily multiple models of the L1 Pier, L1 Hiide, and Crossmatch SEEK sensors. Several other sensors were used infrequently.

As given in the table above, many evaluation studies had the goal of comparing a collection of algorithms and determining their relative performance [1][4-10].

Such evaluations played a major role to advance current iris recognition technology, even while the experiment for these comparisons was frequently limited to a narrow range of sensors and narrow dataset diversity. Some of the datasets consisted of only near-ideal iris images captured indoors from stationary sensors.

The major difference in our study is that we address the inverse problem; that is, for a given single algorithm, is its performance robust over a collection of datasets that includes a full range of iris sensors?

3. Datasets

This section describes the seven datasets that were used for VASIR's performance evaluation. These seven datasets were captured by six different sensors and were taken using seven different acquisition procedures; all datasets were collected at the University of Notre Dame over several years. Table 2 provides details for each of the seven datasets for both the left and right eye.

The Iris Challenge Evaluation (ICE) [1] program provides a dataset (*D1*) of traditional iris-still images captured by an LG2200 system. LG2200 is a single iris capture at a time sensor. It uses three Near-Infrared (NIR) LEDs, and the subject distance must be within 3-10 inches (8-25 cm) to the camera. The system contains software that uses voice to prompt the subject to adjust his/her distance and eye position to the camera. The LG2200 system normally selects one image out of three images from a shot, and discards the other two images using a built-in image quality control system. The ICE2005 dataset, however, contains all three images to support a broad band of image qualities for research. Therefore, one image out of three meets the built-in quality checks, while the other two images may or may not meet such requirements. The total number of images for both left and right eyes is 2,953 and encompasses a total of 132 subjects.

The Multiple Biometrics Grand Challenge (MBGC) [2] program includes two datasets (D2 and D3).

The *D2* dataset was collected by the same sensor (LG2200) as D1 and the same acquisition approach was used. Note that multiple models of LG2200s might have been used for this dataset, which led to blank regions on some images—see the difference of Figure 1 (a) and (b). The total number of images for this dataset is 8,038 collected from 485 subjects.

Table 2 Summary of the seven datasets (D1 – D7)

Source	**# Set**	**Code**	**Sensor Type/ Iris Image Type**	**Eye Pos.**	**# Images**	**# Subjects**	**Image size**
ICE [1] (2005)	D1	LG2200-ICE	LG 2200	Left	1,527	119	640*480
			Traditional iris-still	Right	1,426	124	
MBGC [2] (2008)	D2	LG2200-MBGC	LG 2200	Left	4,025	485	640*480
			Traditional iris-still	Right	4,013	485	
	D3	IOM	Iris on Move (IOM)	Left	586 of 628	136 of 143	Varying
			Stand-off facial video	Right	586 of 628	136 of 143	
NDSpring [3][4] (2011)	D4	LG4000	LG Iris Access 4000	Left	6,319	554	640*480
			Traditional iris-still	Right	6,300	554	
	D5	LGiCAM	LG iCam (TD100)	Left	4,549	536	640*480
			Handheld iris-still	Right	4,557	536	
	D6	IGAD100	IrisGuard (AD100)	Left	5,994	532	640*480
			Traditional iris-still	Right	5,892	533	
	D7	CFAIRS	CFAIRS IR Cannon	Left	13,507	541	960*640
			Stand-off iris-still	Right	13,648	544	

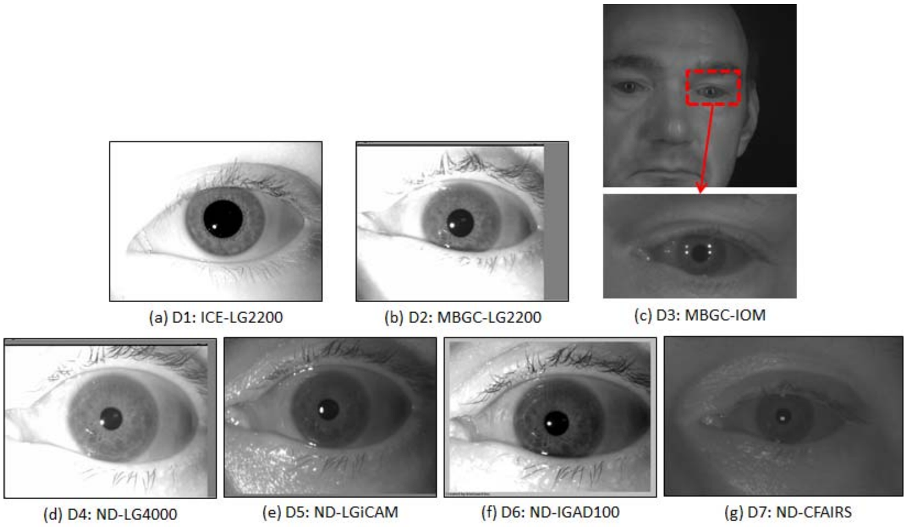

Figure 1 An iris image sample for each dataset from the same subject except the iris image (a)

D3 contains NIR face-visible video sequences, with a resolution of 2048x2048 frame pixels, captured by a Sarnoff Iris On Move (IOM) system [11]. The system takes face videos while a person walks at a distance through a portal with normal walking speed. The portal itself contains multiple NIR LEDs to illuminate subject's face or body. The subject starts to walk from approximately 118 inches (~3 meters) and the subject looks forward while walking. In this dataset, there are a total of 628 video sequences collected from 143 subjects. Out of the total of 628 videos, there are 30 videos in which either no eye is visible or only one eye is visible across all frames. Using VASIR's algorithm, eye-pairs were detected in 586 videos out of the 598 and then for a video sequence, the best-left and best-right iris images were selected. Hence, 1,196 (= 586 x 2) iris images with 136 subjects were used for this study.

The four datasets D4 to D7 were classified by the sensor type from the NDSpring (2011) dataset [3][4]. *D4* consists of traditional iris-still images, captured by an LG IrisAccess 4000 (LG4000) system. LG4000 is a dual iris capture sensor (both irises capture at the same time). It uses two clusters of 12 NIR LEDs with varying wavelengths, and the subject distance is within 14 inches (36 cm). The sensor contains voice instruction software and a built-in image quality control system. In this dataset, three or four sets of images were captured after one sensor prompt. A total of 12,619 images was collected from 554 subjects. *D5* includes the traditional still images captured by LG iCAM TD100 (LGiCAM). LGiCAM is a face and dual iris capture sensor and can be a handheld camera or can be mounted on a tripod for stationary use—both handheld and stationary approaches were used to collect data. The sensor uses multiple clusters of NIR LEDs to illuminate the iris and the subject must be positioned about 13 inches (33 cm) away. This sensor does not have automatic voice instruction software. In total 9,106 images were collected from 536 subjects. *D6* also contains traditional still images captured by the IrisGuard AD100 (IGAD100) system. IGAD100 is a dual iris capture sensor using two clusters of six NIR LED illuminators. The subject distance must be 8-12 inches (20-30 cm) away from the camera, and the system includes voice prompt software to position the subject's eyes. The sensor has a built-in motion measure function, detects the use of glasses and contact lens, and adjusts illuminations automatically. In this dataset, four images were captured after one sensor prompt. The total number of iris images is 11,886 with 533 subjects. *D7* is the set of stand-off iris images captured by a Honeywell Combined Face and Iris Recognition System (CFAIRS). The sensor uses multiple NIR LEDs illuminators and the range of iris scans is within 157 inches (4 meters) away. In this dataset, a total of 27,155 left and right iris images were collected from 544 subjects.

Figure 1 shows a sample of iris images for each dataset—all iris images are from the same subject except (a) LG2200-ICE because there is no common subject in the dataset. Five datasets (see a, b, d, e, and f) out of the seven contain traditional NIR iris still images (stationary or handheld used) and two datasets (c and h) include stand-off video/still imagery that are challenging problems for iris recognition. The iris diameter of the five datasets typically exceeds 200px for optimal images, while the iris size for IOM and CFAIRS datasets varies markedly (with average ~100px). In summary, a total of 72,921 images collected from 796 subjects were used in our evaluation and analysis.

4. Method

4.1. Algorithm (VASIR) Case Study

To demonstrate our algorithm robustness assessment methodology, we used a particular iris recognition algorithm/system VASIR [12][13], which was developed using a structured design and analytic approach that allowed for algorithm evaluation, characterization, and optimization. VASIR is a NIST-developed research algorithm featuring Near-Infrared (NIR) face-visible video-based iris recognition. VASIR is an open-source fully-automated algorithm which is capable of handling three sensor categories: 1) traditional stationary, 2) handheld, and 3) stand-off. The algorithm addresses the challenge of recognizing a person in less-than-optimal environments, while coping with both high and low image quality—see Lee [12] for detailed methods and procedures for optimizing VASIR's sub-components.

As with most iris recognition algorithms, VASIR has a number of input parameters, but VASIR's performance is driven primarily by only six parameters (that can be adjusted for a given dataset): 1) scale factor for the iris size, 2) pupil circle ratio, 3) threshold for detecting pupil boundary, 4) Hough transform scaling factor for detecting the iris boundary, and 5) closing and 6) opening iteration numbers for reducing noise within an image.

For the seven datasets, these parameters have been optimized in a statistically rigorous fashion by using orthogonal fractional factorial experiment designs. Details of this extensive optimization and sensitivity analysis study for biometric systems are given in Lee et al [14].

Note that the six parameter values for each dataset are near-optimal, with opportunities still remaining for continued optimization.

4.2. Evaluation Protocol

In this section, we briefly describe the protocol that we used for our evaluation. The iris recognition performance is conventionally evaluated by four metrics [15][16]. The True Accept Rate (TAR)—also equivalently known as "Verification Rate (VR)—is a proportion of image pairs where two biometric samples are indeed of the same person's eye. The True Reject Rate (TRR) is where the algorithm correctly decided that two samples (genuine and imposter) are not of the same person. The False Accept Rate (FAR) is defined as the rate where samples are accepted as the same identity while in reality they are different. The False Reject Rate (FRR) is where two samples from the same person's eye are rejected. Note that FAR and FRR are often referred as FMR (False Match Rate) and FNMR (False Non-Match Rate), respectively. FAR and FRR share an overlapping area between genuine (mated) and imposter (non-mated) distributions and a smaller overlapping area is considered to be a better performance.

The gallery (query) set is the biometric sample database of known individuals for a specific implementation or evaluation experiment. The probe (target) set is the submitted biometric sample to compare against one or more references in the gallery [17].

The four metrics are not independent, in fact, the TRR and FRR can be computed as (TRR = 1 – FAR) and (FRR = 1 – TAR), respectively. Based on these rates, Error Equal Rate (EER) can be computed as the rate in which FAR and FRR are equal. We used the mated (genuine) scores as threshold to calculate the four metrics.

For benchmarking purposes, we decided in our study to have both the gallery set and the probe set to be the same. Hence, this allows us to efficiently calculate only half of the off-diagonal elements of the similarity matrix [18]. For non-mated scores, the gallery images can be compared against all probe images whereby the probe subject is a different person. This full matching (one to all others) can be time-consuming and is not necessary since the non-mated scores are sufficiently consistent for virtually all probe subjects [18]. Our protocol randomly selects a representative subsample of all probe samples, all of a different person for getting the non-mated results. We find 50 subjects to be a sufficient subsample [18]—a different size subsample may be required depending on the number of subjects and number of images. This not only will reduce execution time but also allows us to characterize the performance more easily.

VASIR performance for verification is evaluated by VR at fixed values of FAR, and by EER. In particular, performance is measured by four criteria: 1) VR at FAR = 0.001, 2) VR at FAR = 0.01, 3) VR at FAR = 0.1, and 4) EER. High VR values and a small EER value indicate superior algorithm performance. Note that if images failed to load/process due to the file corruption, these images (< .04 %) were excluded in our evaluations.

4.3. Experts (Opinion) Case Study

In addition to the above algorithm evaluation, we conducted a complementary reference study by surveying seven biometrics experts in the iris recognition field. The purpose of this study is to provide a ranked list of the iris recognition performance of the six sensors (see the sensor type in Table 2) based on their expert knowledge and experience. Some of the experts had experience with all six sensors, while others only with some of the sensors. The experts were instructed to use their professional judgment to rank the six sensors and to use information from other sources (e.g., Internet) as necessary. They were also instructed to assume that performance is to be defined as FRR at FAR=0.001. Further, they should assume that there is no cross sensor matching, and that the sensors are operating under normal conditions. The expert judges were given randomly ordered sensor lists along with a questionnaire that included the information about sensor model, sensor class, single or dual iris capture function, and handheld state. This study was done simultaneously with the algorithm evaluations.

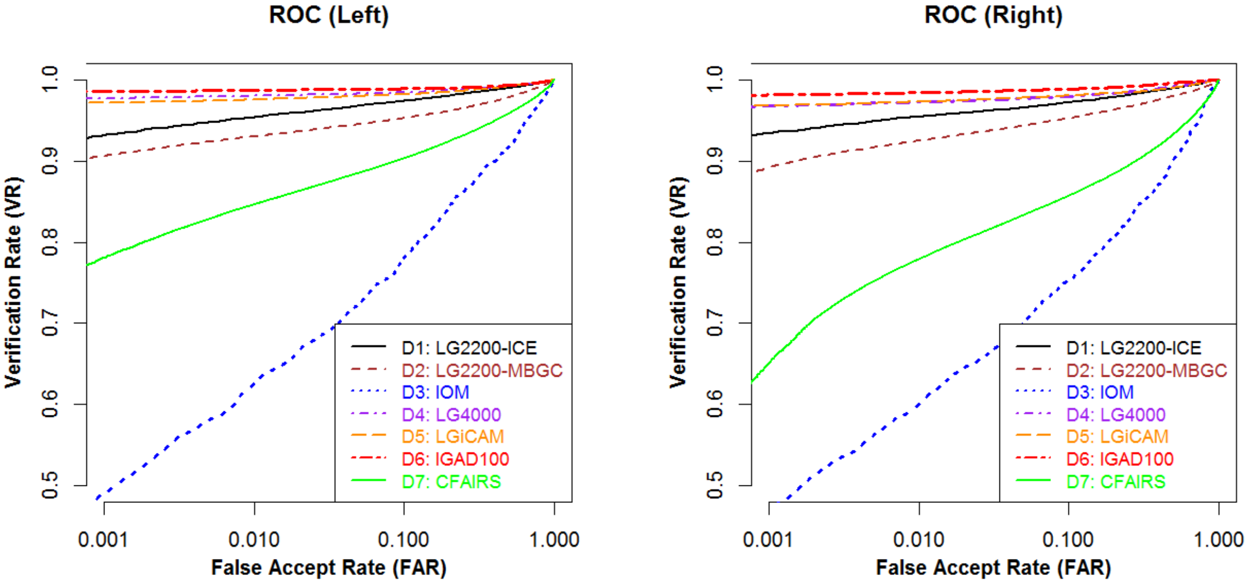

Figure 2 Comparisons of VASIR's VR performance over the seven datasets

5. Results

5.1. Algorithm (VASIR) Experiment Results

This section describes the analysis of the VASIR performance results carried out on the seven datasets spanning six different sensor types. For the VASIR case study, a summary of the number of mated scores and non-mated scores and the performance results with four criteria are given in Table 3.

Table 3 VASIR performance results on the seven datasets for left (L)/right (R) and four performance metrics

Set	Code	L/R	#	#	VASIR Performance (%)			
			Mated	Nonmated	VR001	VR01	VR1	EER
D1	LG2200 -ICE	L	14,653	76,350	93.2	95.3	97.4	3.6
		R	12,221	71,300	93.4	95.4	97.2	3.7
D2	LG2200 -MBGC	L	96,341	201,250	90.6	93.0	95.2	5.4
		R	95,779	200,650	89.2	92.5	95.2	5.5
D3	IOM	L	1,594	29,300	48.8	62.8	78.0	17.3
		R	1,594	29,300	47.3	60.1	75.2	19.5
D4	LG4000	L	43,317	315,950	97.7	98.0	98.4	1.9
		R	43,093	315,000	96.8	97.3	98.0	2.5
D5	LGiCAM	L	22,599	277,450	97.2	97.5	98.2	2.3
		R	22,655	277,850	96.8	97.3	98.0	2.5
D6	IGAD100	L	40,298	299,700	98.5	98.7	98.9	1.3
		R	38,993	294,600	98.1	98.4	98.8	1.5
D7	CFAIRS	L	211,127	675,351	78.1	84.6	90.3	9.7
		R	214,991	682,400	65.2	77.9	85.7	13.2

In this evaluation, the total number of mated scores is 859,255 and the total number of non-mated scores is 3,746,453 across all seven datasets. For datasets D1 to D6, the table shows that the performance results are similar for both left and right eyes, but with average results from the left eye is slightly better than the right eye across all four criteria (except D1). For the D7 dataset, the left eye results are significantly higher than right differencing by 12.9 % for VR at FAR=0.001 and by 3.5 % for EER. This D7 result prompts the need for further investigation.

For both left and right eyes, the overall results show that the D3 (Sarnoff IOM) dataset has the poorest performance with the lowest average VR at FAR=0.001 of 48.0 % and the highest average EER of 18.2 %, while the D4 (IrisGuard AD100) dataset has the best performance with the highest average VR at FAR=0.001 of 98.3 % and the lowest average EER of 1.4 %.

Figure 2 shows ROC curve comparison of VASIR's VR performance results over the seven datasets. Although VASIR's performance varies over the seven markedly diverse datasets, nonetheless, this study serves as a valuable baseline for both algorithmic performance and dataset severity. As with the Table 3 results, the "most challenging" dataset for VASIR is D3 (blue dotted line) collected from the Sarnoff IOM sensor. The "easiest" dataset is the D6 (red dotted line) collected from the IrisGuard AD100 sensor.

Figure 3 illustrates (via a block plot) the ranking of the seven datasets for each of the four criteria and for both eye positions. For both left and right eyes, the results of VASIR performance ranked list of datasets are consistent over the four criteria—therefore, we conclude that the VASIR performance ranking of the seven datasets is robust across both eye position and all four performance metrics. Further, the block plot shows that the results for left are higher than the results for right across all datasets except D1 (LG2200-ICE)—the right result for D1 has slightly better performance and this result is consistent with the conclusion from [1].

In summary, the order of the seven datasets (from easiest to hardest) based on the average VR and the EER value is given in Table 4.

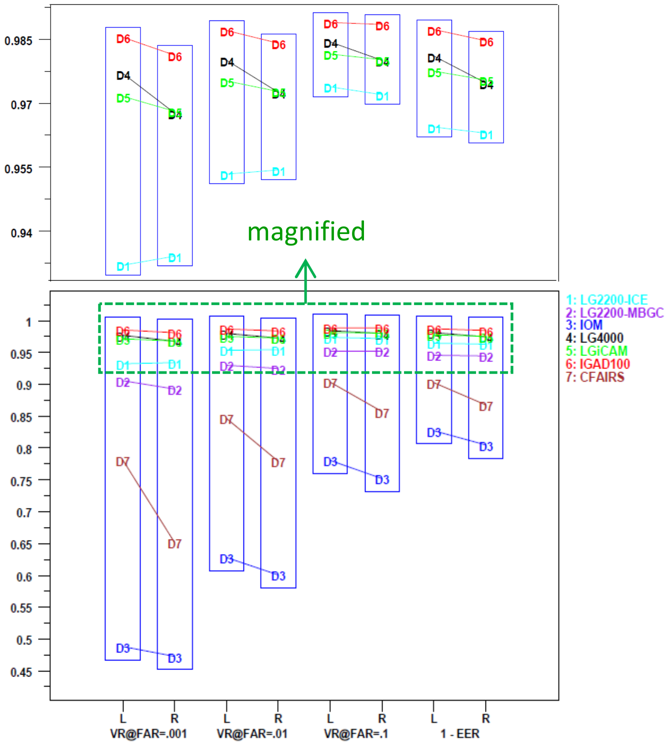

Figure 3 The rank of VASIR performance on the seven datasets for all four criteria and the left (L)/right (R) eye position

Table 4 VASIR performance ranking (easiest to hardest) over the seven datasets

Rank	Sensor code	Set	VASIR Performance (%)			
			VR @ FAR=.001	VR @ FAR=.01	VR @ FAR=.1	EER
1	IGAD100	D6	98.3	98.6	98.9	1.4
2	LG4000	D4	97.3	97.7	98.2	2.2
3	LGiCAM	D5	97.0	97.4	98.1	2.4
4	LG2200-ICE	D1	93.3	95.4	97.3	3.7
5	LG2200-MBGC	D2	89.9	92.8	95.2	5.5
6	CFAIRS	D7	71.7	81.3	88.0	11.5
7	IOM	D3	48.1	61.5	76.6	18.4

From Table 4, D6 (IGAD100) has the highest performance closely followed by D4 (LG4000) and D5 (LGiCAM). These sensor models have multiple NIR LEDs cluster to illuminate the iris region and contain a built-in image quality control software—especially, IGAD100 system has the function to measure motion and to adjust the illumination of the iris image. The eye distance with the camera for IGAD100 is within 12 inches, for LG4000 is within 14 inches, and for LGiCAM is within 13 inches.

On the other hand, D3 (IOM) has the lowest rank—the iris images were isolated from the face-visible video frame and the system deals with moving subject at a distance. As shown samples in Figure 4, out of the seven datasets, the D3 dataset contains the most challenging iris images for the VASIR algorithm, mainly, due to the poor image quality by poor illumination, out-of-focus, and other deficiencies.

To gain understanding into the ranking of the seven datasets, Figure 5 provides the distributions of mated (green) and non-mated (red) scores from the seven datasets (in rank order). Note that the overlap region between the mated and non-mated distribution significantly increases when going from the easiest to hardest performance. We also observe that while the mated distribution shape (green) changed from one plot to the next, the shape of the non-mated distribution (red) stayed relatively consistent over all seven datasets. For the mated case, the lower bound on the Hamming Distance (HD) value range tends to increase with the ranked dataset (from 0.06 to 0.22), while the upper bound remains invariant at approximately 0.52. For the non-mated case, the HD value ranges are stable over the given datasets except D3.

Based on the observation, we arrive at the important conclusion that the main driver of the classification between mated and non-mated distribution in iris recognition is not the non-mated score, but rather the mated score. This conclusion is consistent with the study on the analysis of the iris image quality score levels done by Lee (see p62 in [12]).

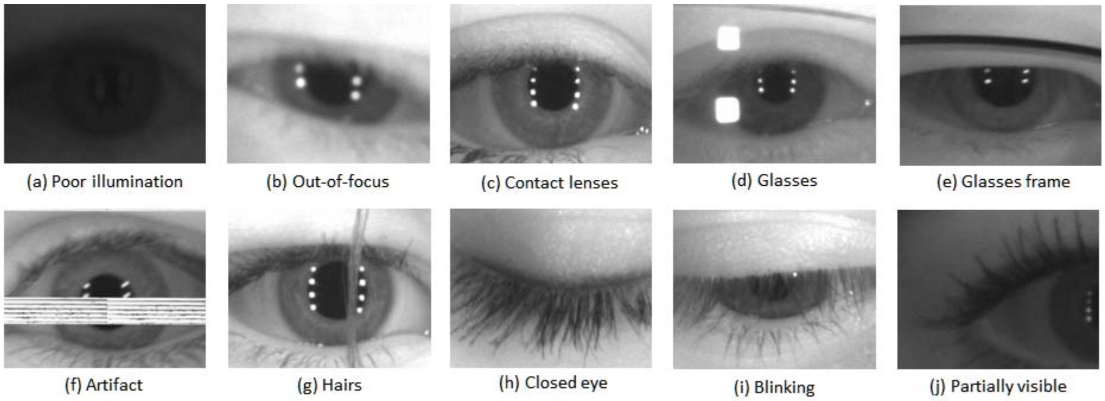
(a) Poor illumination
(b) Out-of-focus
(c) Contact lenses
(d) Glasses
(e) Glasses frame
(f) Artifact
(g) Hairs
(h) Closed eye
(i) Blinking
(j) Partially visible

Figure 4 Demonstration of poor quality images selected as the best image from videos captured by an IOM sensor (D3)

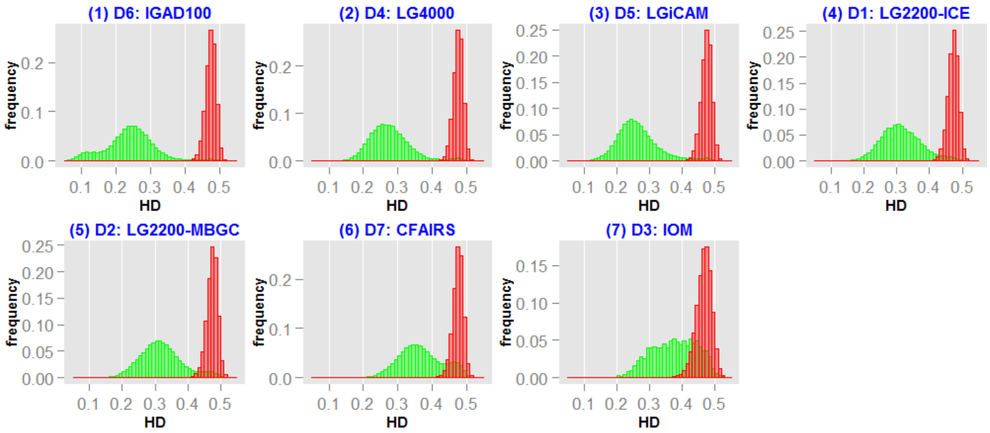

Figure 5 Mated (green) and non-mated (red) distributions of Hamming Distance (HD) scores on the seven datasets (ordered by easiest to hardest)

5.2. Expert Survey Results

This section describes the results of a parallel study in which a survey was conducted from seven biometrics experts. The focus of the survey was to gather expert opinions on the relative (perceived) quality of the given six sensors that were used in the previous algorithmic (VASIR) experiments. Motivation for this survey was to determine the correlation between VASIR experiment-based rankings and expert survey-based rankings.

Table 5 Comparison of algorithm VASIR-based and expert-based ranking (easiest {1} to hardest {6}) on the six sensors

Set	Sensor type	VASIR									Expert							
		VR001		VR01		VR1		EER		Avg	E 1	E 2	E 3	E 4	E 5	E 6	E 7	Avg
		L	R	L	R	L	R	L	R									
D1D2	LG2200	4	4	4	4	4	4	4	4	4	4	3	4	4	3	4	3	3.3
D3	IOM	6	6	6	6	6	6	6	6	6	6	6	6	6	6	5	6	5.9
D4	LG4000	2	2	2	2	2	2	2	3	2.1	2	2	1	2	2	1	1	1.6
D5	LGiCAM	3	2	3	2	3	2	3	2	2.5	3	4	3	3	5	3	4	3.6
D6	IGAD100	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	2	1.4
D7	CFAIRS	5	5	5	5	5	5	5	5	5	5	5	5	5	4	6	5	5.0

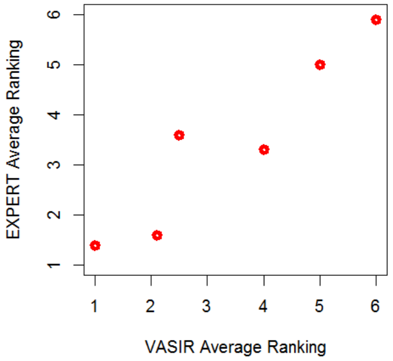

Figure 6 Correlation (94 %) of VASIR ranking and expert ranking

For this study, we surveyed the ranked list (easiest {1} to hardest {6}) from seven experts ($E_1 \ldots E_7$) for the iris recognition performance on the six sensors—note that D1 and D2 were collected by the same LG2200 sensor, so we used the coded dataset D1D2 for the LG2200 sensor for comparison. Table 5 summarizes the comparison of the VASIR experiment-based ranking results and the expert survey-based ranking result; we averaged the LG2200 ranking results for datasets D1 and D2.

The expert-based results show that ranking from easiest to hardest is quite consistent across the seven experts. For example, 6 out of 7 experts predicted the IOM system would perform worst in iris recognition and a majority of the experts (4 of 7) responded that the IGAD100 system would perform best followed by LG4000—in fact, the experts were struggling between IGAD100 and LG4000 for the best. Based on the overall results across the seven experts, the ranked list (easiest to hardest) in terms of iris sensor performance is 1) IrisGaurd AD100, 2) LG4000, 3) LG2200, 3) LGiCAM TD100, 5) CFAIRS, and then 6) IOM, which is very similar to the VASIR evaluation results except LG2200 and LGiCAM. We note that two (E1, E4) of the experts had identical rankings with VASIR's average ranking. Further, as shown in Figure 6, the experts' average ranking and VASIR's average ranking were highly correlated (94%)—hence, this would provide an interesting potential link whereby VASIR-based sensor ranking conclusions serve as a pre -dictor for human expert-based sensor ranking conclusion.

6. Conclusions

This study presented a structured evaluation protocol for algorithm robustness assessment. We used the open-source research algorithm VASIR to demonstrate this protocol. VASIR's performance was characterized by four performance metrics over seven diverse datasets captured from six different sensor types. Further, a parallel biometrics-expert survey was conducted to provide a ranked list of the given six sensors for iris recognition.

For the VASIR algorithm experiment, the results showed that the dataset collected by the Iris Guard AD100 (IGAD100) system had the highest performance with VR at FAR=0.001 of 98.3 % and EER value of 1.4 %, while the dataset with Iris on Move (IOM) had the lowest performance with VR at FAR=0.001 of 48.1 % and EER value of 18.4 %. The ranked list results for VASIR performance over seven dataset were robust across all four performance metrics and eye position. The order from highest to lowest was as follows: 1) D6: IrisGuard AD100, 2) D4: LG4000, 3) D5: LGiCAM TD100, 4) D1: LG2200-ICE, 5) D2: LG2200-MBGC, 6) D7: CFAIRS, and 7) D3: IOM. Further, we observed that the main driver of the classification between mated and non-mated distribution in iris recognition was not the non-mated score, but rather the mated score.

For the biometrics expert study, the results showed that the ranking of the six sensors is surprisingly consistent across the seven participants. It is also of interest that the sensor average ranking from the experts was highly correlated with the average ranking based on the VASIR algorithm study. Hence, an unexpected derivative result from our robustness protocol was that the VASIR algorithm of the protocol may serve as a good predictor for the human-expert ranking of sensors in iris recognition.

The latest version of VASIR algorithm, the evaluation protocol and its scripts, and the study results are publically available at: http://www.nist.gov/itl/iad/ig/vasir.cfm

Our study will give academia, industry, and government researchers an opportunity for insightful understandings of the algorithm, for comparing user algorithms with the module-based algorithms (e.g., segmentation), and for educationally advancing iris-based biometrics technology in general.

Disclaimer

The identification of any commercial product or trade name does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] P. J. Phillips, K. W. Bowyer, P. J. Flynn, X. Liu, and W. T. Scruggs, “The iris challenge evaluation 2005,” in 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, 2008, pp. 1–8.

[2] P. J. Phillips, P. J. Flynn, J. R. Beveridge, K. W. Bowyer, W. T. Scruggs, A. O’Toole, and et al., “Overview of the Multiple Biometrics Grand Challenge,” National Institute of Standard and Technology, NIST Interagency/Internal Report (NISTIR) 7607, 2009.

[3] R. Connaughton, A. Sgroi, K. W. Bowyer, and P. Flynn, “A cross-sensor evaluation of three commercial iris cameras for iris biometrics,” in Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, 2011, pp. 90–97.

[4] R. Connaughton, A. Sgroi, K. Bowyer, and P. J. Flynn, “A Multialgorithm Analysis of Three Iris Biometric Sensors,” Information Forensics and Security, IEEE Transactions on, vol. 7, no. 3, pp. 919–931, 2012.

[5] Internation Biometirc Group, “Independent Testing of Iris Recognition Technology,” Technical Report, 2005.

[6] Authenti-Corp, “Iris Recognition Study 2006 (IRIS06)” Technical Report, 2007.

[7] K. W. Bowyer, S. E. Baker, A. Hentz, K. Hollingsworth, T. Peters, and P. J. Flynn, “Factors that degrade the match distribution in iris biometrics,” Identity in the Information Society, vol. 2, no. 3, pp. 327–343, Dec. 2009.

[8] Patrick Grother, E. Tabassi, G. W. Quinn, and W. J. Salamon, “IREX I: Performance of Iris Recognition Algorithms on Standard Images,” National Institute of Standards and Technology, NIST Interagency/Internal Report (NISTIR) NISTIR 7629, 2009.

[9] P. J. Phillips, W. T. Scruggs, A. J. O’Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, “FRVT 2006 and ICE 2006 Large-Scale Experimental Results,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 5, pp. 831–846, May 2010.

[10] P. Grother, G. W. Quinn, J. R. Matey, M. Ngan, W. J. Salamon, G. Fiumara, and C. I. Watson, “IREXIII - Performance of Iris Identification Algorithms,” NIST Interagency/Internal Report (NISTIR) 7836, 2012.

[11] J. R. Matey, O. Naroditsky, K. Hanna, R. Kolczynski, D. J. LoIacono, S. Mangru, M. Tinker, T. M. Zappia, and W. Y. Zhao, “Iris on the Move: Acquisition of Images for Iris Recognition in Less Constrained Environments,” IEEE Proceedings, vol. 94, no. 11, pp. 1936–1947, 2006.

[12] Y. Lee, “VASIR: Video-based Automatic System for Iris Recognition,” Dissertation, School of Computer Science and Engineering, Chung-Ang University, South Korea, 2012.

[13] Y. Lee, R. J. Micheals, J. J. Filliben, and P. J. Phillips, “VASIR: An Open-Source Research Platform for Advanced Iris Recognition Technologies,” Journal of Research of the National Institute of Standards and Technology, vol. 118, p. 218, Apr. 2013.

[14] Y. Lee, J. J. Filliben, R. J. Micheals, and P. Jonathon Phillips, “Sensitivity analysis for biometric systems: A methodology based on orthogonal experiment designs,” Computer Vision and Image Understanding, vol. 117, pp. 532–550, Jan. 2013.

[15] Flynn, PJ and Phillips, PJ, “ICE mining: Quality and demographic investigation of ICE 2006 performance results,” National Institute of Standards and Technology, Tech. Rep., 2008.

[16] P. Grother, R. Micheals, and P. J. Phillips, “Face Recognition Vendor Test 2002 Performance Metrics,” Fourth International Conference on Audio-Visual Based Person Authentication, 2003.

[17] NSTC Subcommittee on Biometrics, “Biometrics Glossary,” National Science and Technology Council (NSTC), 2006.

[18] Y. Lee, Phillips, P. J., Micheals, R. J., Filliben, J. J., and Sahibzada, H. A., “Ocular and Iris Recognition Baseline Algorithm,” National Institute of Standard and Technology, NIST Interagency/Internal Report (NISTIR) 7828, 2011.