Exploring the Upper Bound Performance Limit of Iris Biometrics Using Score Calibration and Fusion

Dmitry O. Gorodnichy, Elan Dubrofsky, Richard Hoshino,
Wael Khreich, Eric Granger, Robert Sabourin

***Abstract*—Researchers now acknowledge that the ultimate goal for biometric technologies to be error-free may never be achieved for any biometric modality. The key interest therefore for any biometric modality is to know its current performance limits. For the iris modality, which is intensively used for trusted traveller programs in many countries, the question of the iris recognition limitations is of particular importance, as it affects security risk mitigation strategies employed by the programs. In this paper, we provide the answer to this question, based on the recent large-scale evaluations of state-of-the-art iris biometrics systems conducted by the National Institute of Standards and Technology (NIST) and the Canada Border Services Agency (CBSA) and two performance-improving post-processing methods developed by the CBSA and its academic partners: one based on score recalibration and the other based on fusion of decisions from multiple systems. Particular emphasis of the paper is on the description of datasets used in iris evaluations and the presentation of the new large-scale iris dataset created for the purpose at the CBSA. The importance of proper evaluation metrics and methodologies used in iris evaluations, including the subject-based analysis, is discussed.**

I. Introduction

How reliable is the iris biometric modality? What is the current upper-bound performance limit for this modality? — These are the key questions for iris biometrics users, such as CBSA [1], USA Department of Homeland Security [2], UK Home Office [3] and other European governments [4], [5], [6], [7] who use iris systems in Pre-approved Traveller Programs, the answers to which influence major business and operational decisions.

The only way to answer these questions is to conduct a large-scale performance evaluation that involves testing of the biometric products available on the market with a significantly large dataset.

Dmitry O. Gorodnichy (corresponding author) and Elan Dubrofsky are with Science and Engineering Directorate, Canada Border Services Agency, 14 Colonnade Road, Ottawa, Ontario, Canada, K2E 7M6. Richard Hoshino is with National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan. Eric Granger, Wael Khreich, Robert Sabourin are with Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Univérsite de Québec, Montreal, Canada.

Citation: Dmitry O. Gorodnichy *et al.*, "Exploring the upper bound performance limit of iris biometrics using score calibration and fusion", In Proc. IEEE Symposium Series in Computational Intelligence (SSCI 2011), Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM), 11-15 April 2011, Paris - France.

In most cases, a biometric user has to rely on external evaluations conducted elsewhere, the most referred of which are conducted by NIST [8].

In certain cases however, an organization may be interested in conducting iris evaluation itself in order to obtain the results pertaining to the particular system or data used by the organization, which is what CBSA has done using the in-house developed multi-order score analysis methodology [9], [10], [11].

The reason for doing this is seen from the fact that *the results of biometric performance evaluations are both 1) product-specific and 2) dataset-specific.*

To illustrate this point, refer to Figures 1.a-c which show the Detection Error Tradeoff (DET) curves obtained in the IREX iris evaluation recently conducted by NIST [8], in which ten different iris products were tested using three different datasets. One can observe that iris recognition performance varies significantly from vendor to vendor, as well as from one dataset to another. To highlight the difference in performance, Table II-a shows False Non-Match Rates (FNMR) obtained at fixed False Match Rates (FMR) by the same products on three different datasets.

Therefore, to better understand the value of the reported evaluation results, one needs to have a good understanding of the datasets used in obtaining those results.

Additionally, ones needs to know what testing protocol has been used in obtaining the results. — If tuning of the systems prior to testing is allowed, which is the case with IREX evaluation, this will produce better performance numbers, compared to when the default system settings have been used with no code/setting adjustment performed, which is how the systems were examined by CBSA.

Furthermore, certain products may allow the improvement of their performance through additional score analysis (referred to as Order-3 analysis [10], [11]) and post-processing techniques. This is demonstrated in Table II-b, which shows the performance of the same commercial systems on the same dataset without any post-processing (column 1) and with two post-processing techniques described in this paper (columns 2 and 3). The first of these techniques is based on the score calibration proposed in [12], [13] and the other is based on the fusion of decisions from multiple systems proposed in [14], [15].

Therefore, *relying on DET curves and FMR/FNMR metrics published in external reports may not be sufficient to fully understand the limitations and capabilities of a biometric*

system or modality.

The paper is organized as follows. First, we describe the datasets used in the large-scale iris evaluations conducted to date, including those used by NIST and the datasets created by CBSA, and present the results obtained on those datasets (Section II and III). Then we describe the score calibration and fusion techniques and show the recognition results obtained using these two techniques (Sections IV and V). The paper concludes with the discussion and the presentation of the results from the higher-order score analysis and the subject-based performance analysis conducted by the CBSA with iris systems.

TABLE I
SUMMARY OF IRIS DATASETS USED IN LARGE SCALE EVALUATIONS.

Dataset	Origin	#subjects	#images(enrolled+passage)
ICE	U. Notre Dame	249	249*3*2
BATH	U. of Bath	664	23025
OPS	operational	8160	(8160+8160)*2
G-500	operational	500	500+3000
G-4000	operational	4000	4000+24000

II. IRIS DATASETS USED IN LARGE-SCALE EVALUATIONS

This section and Table I summarize the specifics of the iris datasets used in the large-scale iris evaluations conducted to date. The first three datasets are those used by NIST in the IREX evaluations, summarized from [8]. The fourth one is developed by CBSA for its large-scale iris examination, the protocol of which and the representative anonymized results of which have been presented in our previous work [9], [10].

Figure 1 and Table II show the DET curves and FMR/FNMR rates obtained on these four datasets.

A. The ICE dataset

The ICE corpus has been created by the U. of Notre Dame [16]. It consists of left and right iris images collected from a university population over six semesters running from 2004 to 2006. The images are 480x640 in resolution, with the diameter of the iris in the image exceeding 200 pixels for most "good" images.

The images are acquired using an LG EOU 2200 iris scanner, which is a complete acquisition system that has automatic image quality control checks. They are stored with 8 bits of intensity, but every third intensity level is unused, which is the result of a contrast stretching automatically applied within the LG EOU 2200 system.

The system takes images in "shots" of three, with each image corresponding to illumination of one of the three infrared (IR) light emitting diodes (LED)s used to illuminate the iris.

For a given subject at a given iris acquisition session, two "shots" of three images each are taken for each eye, for a total of 12 images. The system provides a feedback sound when an acceptable shot of images is taken. An acceptable shot has one or more images that pass the LG EOU 2200's built-in quality checks, but all three images are saved. If none of the three images pass the built-in quality checks, then none of the three images are saved. At least one third of the iris images pass the system quality control checks, and up to two thirds do not pass. A manual quality control step at Notre Dame was performed to remove images in which the eye was not visible at all, for example, due to the subject having turned their head.

Issues and operational relevance: The use of ICE dataset proved controversial in the ICE 2006 evaluation because the suppression of the camera's quality control apparatus caused operationally non-representative images (e.g., eyes closed, non-axial gaze, blur) to be present in the dataset. The presence of degraded images adversely affected iris recognition accuracy, and while larger error rates give better statistical significance to FNMR estimates, the test results have less relevance to operational reality.

B. The BATH dataset

The BATH dataset has been created by the University of Bath in the United Kingdom [17]. The images were collected using a computer vision camera (not a commercial iris product) at a high resolution such that the uncompressed greyscale eight bit raster images have resolution of 1280 x 640 pixels across the peri-ocular region. The dataset is comprised of 29525 images from 800 individuals. This does not include the images held in directories labeled NonIdeal which were ignored throughout.

The images were downsampled to 640 x 480 via 2 x 2 neighborhood averaging. The iris diameter distribution is bimodal, with the average iris diameter equaling 275 pixels. Images with an iris diameter in excess of 340 pixels were omitted from the IREX sample. The effect of this operation reduced the number of images to 23055 and the number of subjects to 664.

Issues and operational relevance: Because the BATH dataset images have been collected using a regular computer vision camera, they were not required to pass the quality check that is normally implemented in commercial iris capture systems. This makes this dataset more challenging for testing commercial systems and may produce weaker results due to the presence of iris images that do not conform to the required quality measurements.

C. The OPS dataset

The OPS dataset is an operational dataset [8]. It consists of two captures of the left and right irises of 8160 individuals. This gives a total of 32640 distinct images. The images were collected using the PIER 2.3 camera from Securimetrics, now a division of L1 Identity Solutions. The files were extracted from a large multimodal dataset so that only the images of "matched" subjects are extracted. A subject is considered "Matched" when either of the following two conditions is met: either one of their eyes (either left or right) strongly match or both of the eyes weakly match using the production system.

The OPS dataset might be considered easy, because many of the images will never be involved in a failed comparison. Therefore, a smaller dataset in which errors are concentrated

TABLE II
IRIS PERFORMANCE RESULTS OBTAINED ON FOUR DIFFERENT DATASETS.

a)

FMR	ICE: 1	2	3	4	OPS: 1	2	3	4	BATH: 1	2	3	4
0.0001%	0.023	0.023	0.025	0.050	0.002	0.005	0.006	0.008	0.010	0.018	0.030	0.031
0.01%	0.009	0.010	0.014	0.015	0.0018	0.004	0.005	0.007	0.004	0.007	0.013	0.025

b)

FMR	G-500: 1	2	3	4	+Calib: 1	2	3	4	+Fused(all)	+Fused(two)
0.0001%	0.09 (0.28)	0.27	0.18	0.11	0.06 (0.16)	0.24	0.99	0.05	0.28	0.11
0.001%	0.06 (0.21)	0.16	0.17	0.07	0.04 (0.10)	0.09	0.99	0.06	0.07	0.08
0.01%	0.05 (0.17)	0.10	0.17	0.06	0.03 (0.06)	0.07	0.17	0.05	0.02	0.06
0.1%	0.04 (0.12)	0.06	0.16	0.05	0.02 (0.03)	0.04	0.16	0.05	0.01	0.01
FTA.E	0.008	0.132	0	0	-	-	-	-	-	-
FTA.P	0.011	0.233	0.001	0.012	-	-	-	-	-	-

The table shows False Non-match Rate (FNMR) at fixed False Match Rate (FMR) for the best four performing systems: a) from IREX evaluation, with the effect of FTA counted for; and b) from anonymized CBSA examination, with the effect of FTA not counted for – without post-processing and with post-processing using score calibration (+Calib) and using fusion (+Fused). The table also shows Failure to Acquire (FTA) of the systems tested on the G-500 dataset (FTA.E is FTA for Enrolled images, FTA.P is FTA for Passage images. FTA = FTA.E + FTA.P) and the FNMR computed using the subject-based analysis (in brackets for one of the systems). All FNMR measurements are obtained approximately from the DET curves by the interpolation at the particular FMR points.

has been created. This is comprised of 1335 genuine image pairs from 1144 subjects. Unless otherwise stated, these are used by taking the first image of the pair and comparing with all members of the OPS dataset including the second mated member of the pair.

Issues and operational relevance: The fact that the operational OPS dataset has already been matched (at some threshold) means the images are clean - false rejection will be less frequent than if no matcher had been used. That being said, if the original OPS collection policy had embedded a matching phase, then localization failures on the resulting corpus would be much rarer and performance would likely be better than for a collection that did no such thing.

Because the OPS dataset has been obtained using a commercial product used in production of the images, it is expected that a bias could be present towards this product.

The images are likely to be more representative of enrollment samples in which care had been taken to produce a pristine and matchable image. This makes the dataset less attractive for evaluating the performance of systems in less controlled environments.

TABLE III
NUMBER OF GENUINE AND IMPOSTOR COMPARISONS PERFORMED.

System	1	2	3	4	
	G-500				Total
Genuine	2,942	2,049	2,997	2,963	3,000
Impostor	1,468,194	996,585	1,495,503	1,478,537	1,497,000
FTA.E	58	951	3	37	
FTA.P	28,806	500,415	1,497	18,463	
FTA	28,864	501,366	1,500	18,500	

For each of the systems tested on the G-500 dataset (Refer to Table II-b), the FTA numbers indicate the number of comparisons that are not performed due to Failure To Acquire of either Enrolled images (FTA.E) or Passage images (FTA.P).

III. THE CBSA "G-" IRIS DATASETS

The CBSA datasets have been created to facilitate the examinations of market products using multi-order score analysis described in [9], [10], [11]. The CBSA datasets, named G-100, G-500, G-1000, and G-4000, are made of enrolled and passage images, corresponding to the same individuals. Particularly, a G-N dataset has N "enrolled" images and 6N "passage" images corresponding to N enrolled travellers, where each enrolled passenger has exactly 6 passage images. Only right eye images are used.

TABLE IV
IRIS RECOGNITION PERFORMANCE WITH AND WITHOUT POST-PROCESSING SCORE CALIBRATION, REPORTED USING TRANSACTION-BASED AND SUBJECT-BASED ANALYSIS (FROM [11]).

Failure to Acquire (Enrolled & Passage Images): FTAe = 0.8%, FTAp=1.1%					
False Match (Accept) Rate: FMR (%)	Failure of Confidence Rate: FCR (%)	False Non-Match (Reject) Rate: FNMR (%)			
		Using system original scores		Using Gorodnichy-Hoshino score calibration	
		Transaction based	Subject based	Transaction based	Subject based
0.0001 *	**0**	**10**	**28**	**6**	**16**
0.000136	0.03	8.0	26.8	5.2	14.0
0.001 *	**0.6**	**6**	**22**	**4**	**10**
0.00102	0.57	6.2	21.0		
0.00109				4.1	10.2
0.01 *		**5**	**17**	**3**	**6**
0.0101				2.9	5.8
0.0174		4.4	15.6		
0.1 *		**4**	**12**	**2.3**	**4**
0.103				2.2	2.6
0.146		3.2	11.4		

* Estimated interpolated

The numbers are obtained on the G-500 dataset with the best performing system (Refer to Table II-b).

By design, G-500 and G-4000 datasets are created not to overlap each other, however G-100 and G-1000 datasets are the smaller subsets of G-500 and G-4000, respectively.

Similar to the OPS dataset, the images used in the CBSA datasets are the images of the "matched" subjects only. Particularly, each passage image used in the dataset have been already matched by the operational system to its corresponding enrolled image. The letter "G" in the naming of the datasets comes from "Genuine" to indicate the all images in the dataset come from genuine transations.

The images are captured by a commercial iris acquisition system and have to pass the image quality check applied by the system. However, the enrolled images are normally of better quality than the passage data, since they are captured in a controlled environment at the time of enrollment under the guidance from an enrolling officer, while the passage data are captured in the airport with no guidance.

The captured images are securely saved using the system's proprietary format, which cannot be read by other systems.

The "Import" function is used to extract the images from their original proprietary format into JPEG format, which results in degrading the image quality. However, to mitigate the effect of such conversion on the evaluation results, the following procedure has been used. First, all captured (enrolled and passage) anonymized iris images available in the operational database are imported to the JPEG format, using the system's "Import" function. Then the compressed version of each image is compared to its original using the image quality function provided by the system, which can read both compressed and original images. If the image quality numbers of both (compressed and original) versions of the image are the same, then the image is marked as "not degraded". Only these "not degraded" images are used. The number of such images was sufficient to create datasets with up to N=4000 enrollees. This however was one of the factors in limiting the number of passage images to 6.

According to the CBSA testing protocol, for which the G-N datasets have been created, all $6N$ passage images are matched to all N enrolled images, resulting in $6N$ genuine comparisons, and $6N(N-1)$ impostor comparisons. The actual number of comparisons performed is often less than that due to a percentage of images that are rejected by the system, due to the system's Failure of Acquire (FTA). This is illustrated in Table III, which shows the total umber of attempeld Genuine and Impostor comparisons as well as the number of comparisons that triggered a failure to acquire (FTA) for enrolled images (FTA.E) and passage images (FTA.P). The data is obtained for the same four anonymized productsshown in Table II-b and Figure 1.d tested on the G-500 dataset.

Issues, values and C-BET evaluation

Further analysis on the effect of image compression in the CBSA datasets as well as the fact that all their passage images have been already matched by the system may need to be further conducted. Nevertheless, these datasets have become instrumental in exposing the limitations of the existing biometric performance metrics, such as those defined by international standards [18], [19], [20] and conventionally used by industry and academia [21], [22], [23], [24].

These datasets have also allowed us to develop and test a new evaluation methodology and metrics based on the multi-order score analysis, collectively referred to as the *C-BET (Comprehensive Biometrics Evaluation Toolkit)* evaluation framework, which provided the scientific and biometric user communities with guidelines on all-inclusive reporting of biometric system performances [9], [10], [11].

As described in [11], contrary to the conventional biometric methodology, the C-BET methodology is modality- , design-, and application- agnostic. That is, it can be applied to any system design (1-to-1 vs. 1-to-many), any application mode (instant fully automated recognition as in access/border control vs. semi-automated recognition as in forensic investigation) and any capture environment (constrained and cooperative vs. unconstrained and uncooperative).

The CBSA datasets have also made it possible to better understand the inner properties of biometric systems, which are often treated as "black boxes", and to develop new post-processing techniques to improve the performance of those "black box" systems. Two of such post-processing techniques are presented in the next two sessions. The G-500 dataset is used to demonstrate and validate the results.

IV. Improving the performance using score calibration

In [12], [13], Gorodnichy & Hoshino have proposed a post-processing score calibration algorithm to improve the performance of commercial off-the-shelf biometric systems. The improvement is achieved by making the use of the a-priori knowledge of the genuine and impostor score distributions generated by the system, which is obtained in advance using the Order-0 score analysis, and which is used to derive the posterior probabilities of a probe belonging to a particular enrolled person. The theory and empirical results show that by replacing the original iris scores, such as those computed using the L1 norm (Hamming Distance), with the posterior probabilities, one can attain the best achievable performance for a system. The algorithm is further summarized below.

Assume that the genuine and impostor matching scores distributions are binomial and denote them as $G \sim Binom(\hat{m}, \hat{u})$ and $I \sim Binom(m, u)$, where $\hat{u}$ and u are the means of genuine and impostor score distributions, and $\hat{m}$ and m are the distributions' degrees-of-freedom, computed from $\hat{u}$ and u and the score standard deviation $\hat{\sigma}$ and σ as

$$m = \frac{u(1-u)}{\sigma^2}.$$

Let $\{x_1, x_2, \ldots, x_n\}$ designate the set of enrolled people and X designate a person X arriving at the kiosk.

For each $1 \le i \le n$, define $s_i = s_i(X, x_i)$ to be the matching score of x_i. Thus, person X produces the n-tuple $S = (s_1, s_2, \ldots, s_n)$, the vector of matching scores.

Define $c_i = P(\{X = x_i\} \mid S)$ the probability that X is passenger x_i, given the n-tuple S. The probability vector $C = (c_1, c_2, \ldots, c_n)$ defines the calibrated confidence scores. Compute new scores c_i according to the *Score Calibration Function* (SCF) given below:

$$c_i = \frac{p_i z_i}{\sum_{i=1}^{n} p_i z_i + q \cdot \dfrac{(1-u)^m}{(1-\hat{u})^{\hat{m}}}}, \quad \text{where} \tag{1}$$

$$z_i = \frac{\binom{\hat{m}}{\hat{m}s_i}}{\binom{m}{ms_i}} \cdot \left(\frac{\hat{u}^{\hat{m}}(1-u)^m}{u^m(1-\hat{u})^{\hat{m}}} \right)^{s_i}, \tag{2}$$

where $p_i = P(X = x_i)$ is the a-priori probability that an individual arriving at the kiosk is person x_i, and $q = 1 - \sum_{i=1}^{n} p_i$ is the probability that the individual is unenrolled.

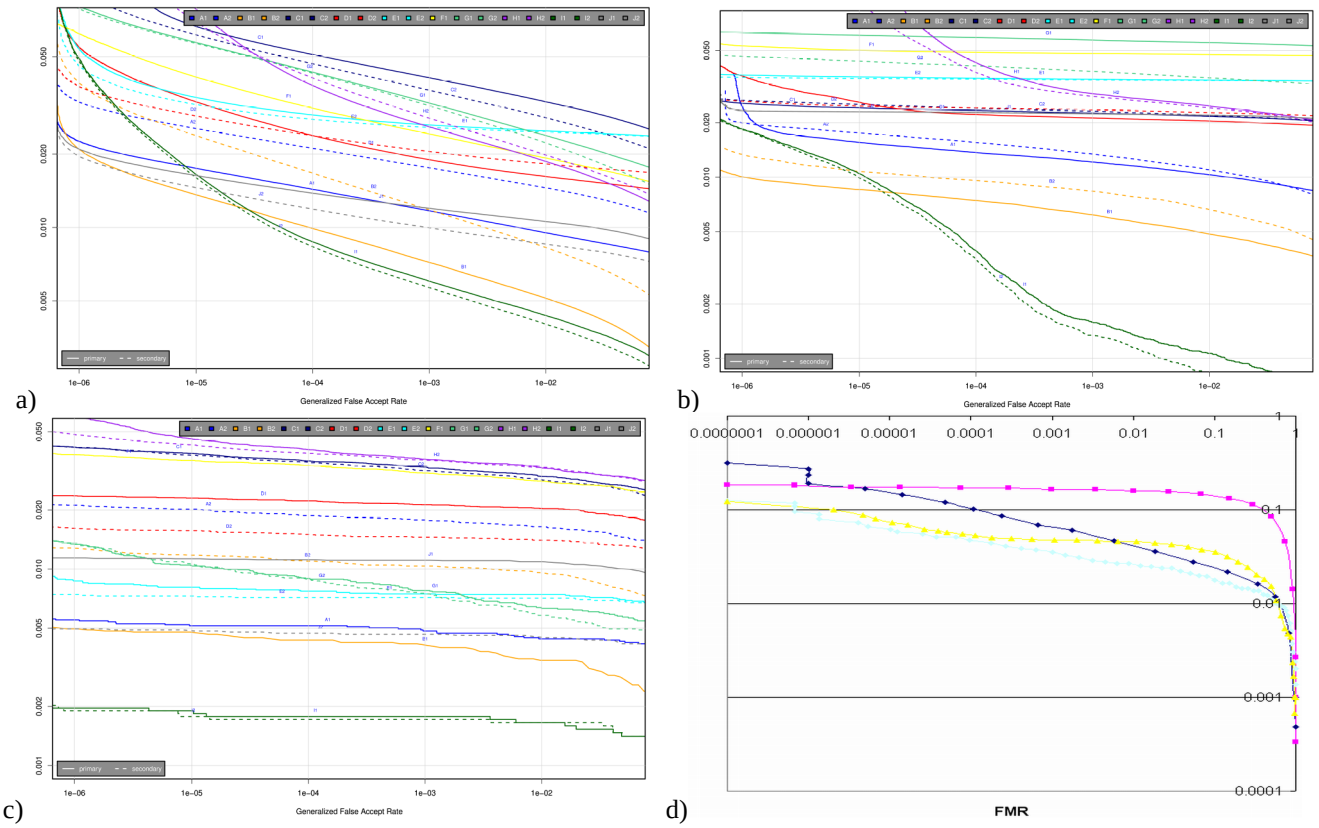

Fig. 1. DET curves obtained on four different datasets: ICE (a), BATH (b)and OPS (c) used by NIST (taken from[8] for uncompressed images), and G-500 (d) used by CBSA.

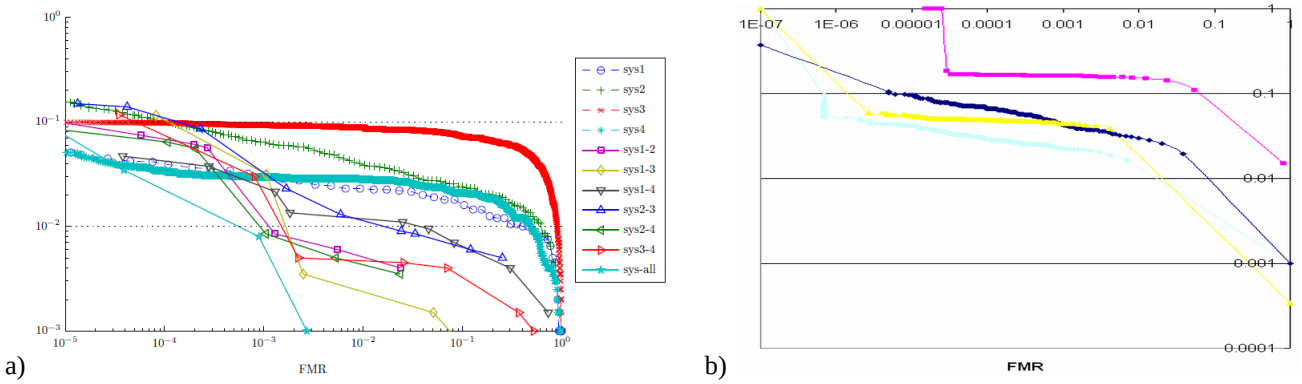

Fig. 2. DET curves obtained on the G-500 dataset with fusion (a) and score calibration (b). - Compare to Figure 1.d.

This SCF function replaces matching scores with meaningful confidence scores that are perfectly calibrated and normalized. This algorithm has been shown

(a) to replace matching scores with meaningful confidence scores that are perfectly calibrated and normalized, regardless of the size of the enrollment database or the nature of the distributions of the genuine and impostor matching scores.

(b) to produce a convex receiver operating characteristic (ROC) curve and DET curve, that dominates the ROC and DET curves of *any* other algorithm. Therefore, this approach of turning matching scores into calibrated confidence scores maximizes the overall accuracy of the biometric system, and cannot be improved any further.

(c) to effectively separate the genuine confidence scores from the impostor confidence scores, with the overwhelming majority of genuine comparisons receiving the maximum confidence score of $c = 100\%$ and nearly every impostor comparison receiving the minimum confidence score of $c = 0\%$.

The results of applying this algorithm to four different iris systems with the G-500 dataset are shown in Figure 2.b (compare to Figure 1.d) and Table II-b, as well as in Table IV, which shows detailed results for one of the tested system. As

seen, for several systems using the Gorodnichy-Hoshino score calibration allows one to decrease the false non-match rate by almost a factor of two.

V. Improving the performance using fusion

Fusion of the evidence from multiple different sources of information is another approach to improve accuracy and reliability of iris modality. Despite reducing information to binary decisions, integrating sources of information at this level provides a robust framework for combination that applies across different biometric modalities and systems, and eliminates issues related to score normalization.

An iterative Boolean combination (IBC) technique has been proposed in [14], [15] for efficient fusion of responses from multiple decision-level fusion technique for biometrics, where the Receiver Operating Curves (ROC) and DET curves may result from a wide range of biometric systems designed with different modalities, sensors, feature sets, classifiers, parameters, training data etc. It does not require any prior assumption regarding the independence of classifiers and the convexity of ROC curves. IBC has been successfully applied to combine the responses from a multiple classifier system, based on hidden Markov models (HMMs), for host-based intrusion detection. It has been shown to provide a significantly higher level of accuracy than related techniques, especially when classifiers are trained with limited data [14].

Given K classifiers, the input to the IBC algorithm are the score vectors, $S_1, S_2, \ldots S_K$, assigned by each classifier to validation samples. The IBC starts by applying each Boolean function[1] to combine the responses corresponding to each decision threshold from the first classifier to those from the second classifier. Fused responses are then mapped to vertices in the ROC space. Each vertex is the result of two decision thresholds combined with a Boolean function, where each threshold is selected from one classifier. Vertices that are superior to the ROC convex hull (ROCCH) of original classifiers are then selected, and the ROCCH is updated to include these emerging vertices. When all Boolean functions have been applied, the decision thresholds from each classifier along with the Boolean functions corresponding to the emerging vertices on the facets of the updated ROCCH are stored. The responses corresponding to each decision threshold from the third classifier are then combined with the responses of each emerging vertex, and so on, until the last classifier is reached. This corresponds to first iteration of the IBC algorithm [14].

Further improvements in performance may be achieved by re-combining the emerging ROCCH vertices resulting from the previous iteration with those of the original ROC curves over several iterations. Once again, the algorithm begins a cumulative Boolean combination (BC) of each emerging vertex (on the previous ROCCH) with each decision threshold from each classifier, selecting only the thresholds and Boolean functions corresponding to superior vertices. The iterative procedure stops when there are no further improvements to the ROCCH or when a maximum number of iterations are performed [14].

The final composite ROCCH is the new maximum realizable ROC in the Newman-Pearson sense. The outputs are the vertices of the ROCCH, where each vertex is the results of the decision thresholds selected from between two to K classifiers, and combined with the corresponding Boolean functions. These thresholds and Boolean functions are stored and applied during operations.

The final ROCCH allows for visualization and selection of the vertex corresponding to the optimal operating point, according to a desirable false alarm rate, cost of errors, and prior probabilities. The thresholds and Boolean functions corresponding to the selected vertex are then applied during operations. When the desired operating point lies between two vertices on the ROCCH (each one activating different classifiers, thresholds and Boolean functions), the final fusion output is an interpolation of their responses. Retaining the ROCCH also allows for adjusting the operating point during operation to adapt to changes in the environment, such as changes in prior probabilities and cost of errors.

During the design phase, with K classifiers each comprising n distinct decision thresholds (on validation samples), the worst-case time complexity required by IBC is reduced to $\mathcal{O}(Kn^2)$ Boolean operations, compared to $\mathcal{O}(K2^{2^n})$ required with a brute force search for optimal combinations. During operations, the computational overhead of the activated Boolean functions is lower than that of operating the required number of classifiers. Therefore, the worst-case time complexity is limited to operating the K classifiers.

The IBC technique makes no assumptions regarding the independence of classifiers or convexity of their ROC curves. In addition, Boolean combination of responses within the ROC space does not require normalization of scores. In fact, by applying *all* Boolean functions to combine the responses from each decision threshold of each classifiers, the IBC implicitly accounts for the effects of correlation, and accommodates for the concavities in the curves. Concavities are typically caused by limited training data or poorly designed classifiers. In contrast related BC techniques in literature are based on only the AND and OR rules. These rules will not provide improvements for the inferior points that correspond to concavities. However, other Boolean functions, for instance those that exploit negations of responses, may emerge. In addition, the iterative procedure accounts for potential combinations that may have been disregarded during the first iteration, but are useful when provided with limited and imbalanced training data.

Figure 2.a displays the results achieved on the G-500 data by applying the IBC combination rules, obtained on the validation (G-100) data, to combine the responses from each pairwise and from the four different iris systems. For an unbiased evaluation and combination of the systems, the failure to acquire images have been filtered out.

As illustrated in Figure 2.a, the combination of responses from the four systems according to the IBC technique achieves

[1]There are ten distinct Boolean operations on two variables a and b: $a \wedge b, \neg a \wedge b, a \wedge \neg b, \neg(a \wedge b), a \vee b, \neg a \vee b, a \vee \neg b, \neg(a \vee b), a \oplus b, a \equiv b$.

the overall highest level of accuracy, while the combination of responses from systems 2 and 4, achieves the highest level of accuracy among the pairwise combinations.

VI. Further performance analysis

This paper explored the performance limits of iris biometrics by summarizing the results obtained to date and putting together the best reported DET curves and FMR/FNMR results obtained for this modality. The results are obtained using four different large-scale datasets, three of which (OPS, ICE and BATH) are used by NIST on its recent IREX evaluation and one (G-500) is used by CBSA in its own iris product examination. The percularities of each dataset are presented to a provide better understanding of the results obtained on these datasets.

The performance of the systems tested on the G-500 dataset is further improved by using two different post-processing computational intelligence techniques presented in this paper: one based on score calibration and the other based on decision-level fusion. The improved results are shown using DET curves and FMR/FNMR metrics.

A. Additional performance metrics

This paper presented the iris recognition results and performance limits using DET curves and FMR/FNMR metrics.

These metrics however do not provide "the entire story" about the system or modality performance. This is best demonstrated by referring to the results presented in this paper, in particular those shown in Figures 1.d and 2.b (and corresponding columns in Table II-b and Table IV), which show the DET curves for the same four products with and without post-processing score calibration. — A system with a worse DET curve may be better performing than the system with a better DET curve, because it may allow for further post-processing to improve its performance or because its other performance metrics are better. Particularly, in addition to FMR, FNMR and DET curves, other important performance metrics include Failure to Acquire (FTA) and Failure of Confidence Rate (FCR) introduced in [10], shown in Table IV.

Therefore, in order to get "the entire story" about the iris modality, CBSA-S&E has conducted the comprehensive biometric performance evaluation of iris systems using the in-house developed multi-order score analysis, the results of which are presented in [9], [10], [11].

B. Subject-based analysis

In order to further understand the limitations of a modality or a system, a subject-based performance evaluation, known as biometric menagarie or Doddington's zoo analysis [25], [8], [26], should also be conducted. Rephrasing the Doddington's (sheep-lamp-wolf-goat) zoo terminology into a biometric-enabled Trusted Traveller Program context, the biometric system performance may vary substantially for different types of travellers. In particular, four types of travellers are identified: 1) "happy and causing no risk", who rarely/never get False Match or Non-Match errors, 2) "happy but causing risk" users, who rarely/never experience False Non-Match, but who may cause frequent False Non-Match errors thus creating higher security risk in using the system, 3) "frustrated, but causing no risk", who frequently experience False Non-Match problem, but do not cause False Non-Match errors, and finally 4) "frustrated and causing risk" users, who frequently get both False Match or Non-Match errors.

The CBSA-S&E has conducted the subject-based analysis for the iris systems tested on the CBSA G-500 datasets. The results of this analysis, some of which are shown in Table IV, are very revealing. In particular, these results indicate that the percentage of registered travellers who would experience a problem with the system due to false rejection, i.e. subject-based FNMR, is higher than the conventionally reported transaction-based FNMR computed by averaging all comparison transactions. More analysis of this phenomenon is presented in [11] and the topic of our current research.

C. Pilot-based evaluation

The results presented in this paper are based on the off-line evaluation with pre-collected iris data. Very often however, the collected iris data are not fully representative of the performance of the system. For example, the collected iris data may not show "bad" (failed to acquire) images or images that did not match anyone in a database, or may have only the best image from a number of captured images.

In order to further understand the potential and the challenges of iris systems, it is recommended to conduct evaluation using real-life live pilots, within which all interactions between users and the system are logged and which is also the topic of our current research.

Acknowledgment

This paper is an extended version of the CBSA-S&E internal Technical Report (D.O. Gorodnichy. "Exploring the upper bound performance limit of iris biometrics", Canada Border Services Agency, Science and Engineering Directorate, Technical Report TR-10-10-G1, October 2010).

The valuable feedback from the colleagues from other directorates, in particular, Michael Chumakov, is gratefully acknowledged, as is the help of other colleagues and students, who have contributed to creating the datasets, conducting iris products evaluations and developing the C-BET software.

The development of the Comprehensive Biometric Evaluation Toolkit (CBET) is partially funded by the Defence Research and Development Canada's Center for Security Science (DRDC-CSS).

For the Government of Canada's Biometric Community of Practice users and partners, the results presented in this paper as well as other recommendations related to the all-inclusive evaluation of biometric systems are made available at the dedicated CBET portal *https://partners.drdc-rddc.gc.ca/css/Portfolios/Biometrics (Human ID Systems)/C-BET* maintained in partnership of the CBSA-S&E and DRDC-CSS.

DISCLAIMER

The results presented in this paper are intentionally made anonymous not to be associated with any production system or vendor product and are used solely for the tasks identified in this paper. In no way do the results presented in this paper imply recommendation or endorsement by the Canada Border Services Agency, nor do they imply that the products and equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] www.Nexus.gc.ca

[2] http://www.globalentry.gov/netherlands.html

[3] http://www.ukba.homeoffice.gov.uk/ travellingtotheuk/Enteringtheuk/usingiris/

[4] http://www.schiphol.nl/Travellers/AtSchiphol/PriviumIrisscan/ WhyPrivium/FastBorderPassageWithTheIrisScan.htm

[5] http://www.barin.nl/show_pubnews.php?publ_id=1501

[6] http://www.bundespolizei.de/nn_734694/EN/Home/ AutomatedBorderControls/procedure.html

[7] ISO SC 37 WD 29195, Technical Report on passenger processes for biometric recognition in automated border crossing systems, Last edition: 2010-08-12

[8] P. Grother, E. Tabassi, G. W. Quinn, W. Salamon. "IREX I Performance of Iris Recognition Algorithms on Standard Images" NIST Interagency Report 7629, September 20, 2009. http://iris.nist.gov/irex/.

[9] D. O. Gorodnichy. Evolution and evaluation of biometric systems. Proceedings of the IEEE Workshop on Applied Computational Intelligence in Biometrics, IEEE Symposium: Computational Intelligence for Security and Defence Applications (CISDA), Ottawa, July 8-10, 2009.

[10] D. O. Gorodnichy. Multi-order analysis framework for comprehensive biometric performance evaluation. Proceedings of SPIE Conference on Defense, Security and Sensing: track on Biometric Technology for Human Identification. Orlando, 5 - 9 April, 2010.

[11] D. O. Gorodnichy. Further refinement of multi-order biometric score analysis framework and its application to designing and evaluating biometric systems for access and border control. In Proc. of IEEE SSCI Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM), Paris, April 11-15, 2011.

[12] D.O. Gorodnichy, R. Hoshino. Calibrated confidence scoring for biometric identification. NIST International Biometric Performance Conference (IBPC 2010), March 2-4, 2010

[13] D.O. Gorodnichy, R. Hoshino. Score calibration for optimal biometric identification. Proceedings of the Canadian conference on Artificial Intelligence. Ottawa, May 31 - June 2, 2010

[14] W. Khreich, E. Granger, A. Miri, R. Sabourin, "Iterative Boolean Combination of Classifiers in the ROC space: An Application to Anomaly Detection with HMMs," Pattern Recognition, 2010, 43, 2732-2752

[15] W. Khreich, E. Granger, A. Miri, R. Sabourin,"Boolean Combination of Classifiers in the ROC Space," Intl Conf. on Pattern Recognition, Istanbul, Turkey, August 23-26, 2010.

[16] P. J. Phillips et al. Overview of the multiple biometrics grand challenge. Technical report, National Institute of Standards and Technology, www.nd.edu/ kwb/PhillipsEtAlICB 2009.pdf [on June 24, 2009], 2008.

[17] D. M. Monro. University of bath iris image database. Technical report, University of Bath, 2008. http://www.bath.ac.uk/elec-eng/research/sipg/irisweb/ [on June 22, 2009].

[18] ANSI INCITS 409.3-2005 Biometric Performance Testing and Reporting - Part 3: Scenario Testing and Reporting

[19] ISO/IEC SC 37 19795-2:2007 Biometric performance testing and reporting - Part 2: Testing methodologies for technology and scenario evaluation

[20] ISO/IEC SC 37 FCD 19795-5, Information Technology - Biometric Performance Testing and Reporting - Part 5: Grading scheme for Access Control Scenario Evaluation

[21] International Biometric Group. Biometric Performance Certification and test plan - http://www.biometricgroup.com/testing_and_evaluation.html

[22] Mansfield, A., Wayman, J. L. (2002). U.K. biometric working group best practice document. Teddington, UK: National Physical Laboratory.

[23] J. L. Wayman, A. K. Jain, D. Maltoni, and D. Maio, editors. Biometric Systems: Technology, Design and Performance Evaluation. Springer, New York, 2005.

[24] A. K. Jain, P. Flynn, A. Ross, "Handbook of Biometrics", Springer, 2007.
Stan Li (Editor), Encyclopedia of Biometrics, Elsevier Publisher, 2009.

[25] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance. In Proc. Fifth Intl Conf. Spoken Language Processing (ICSLP), pages 13511354, 1998

[26] Tabassi, E., Image specific error rate: A biometric performance metric, 20th International Conference on Pattern Recognition ICPR), August 22-26, 2010.