Exploring the upper bound performance limit of iris biometrics

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 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 in particular of the new large-scale iris dataset created for the purpose at the CBSA. The evaluation metrics and methodologies used in such evaluations are 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 [3], UK Home Office [4] and other European governments [2], [5], [6], [7] who use iris systems in Pre-approved Traveller Programs, the answer to which influences 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.

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 10 different iris products were tested using 3 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 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 those results: specifically, whether the tuning of systems prior to testing has been allowed for the dataset, which is the case with IREX evaluation and which produces better performance numbers, or whether the default system settings have been used with no code/setting adjustment performed, which is how the systems have been 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

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.

calibration and fusion techniques and show the recognition results obtained using these two techniques (Sections IV and V). The paper concludes with a discussion on future work.

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 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 average iris diameter is 275. Iris diamiter distribution is bimodal, with the average iris diameter equal 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 image have been collected using regular a 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 the commercial systems and may produce lower results due to the presence of iris images that do not confirm 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 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.

TABLE II

IRIS PERFORMANCE RESULTS OBTAINED ON DIFFERENT DATASETS.

a)

FMR	ICE	2	3	4	OPS	2	3	4	BATH	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	0.0045	0.0051	0.007	0.004	0.007	0.013	0.025

b)

FMR	G-500	2	3	4	+Calib	2	3	4	+Fused(all)	+Fused(two)
0.0001%	0.093	0.272	0.178	0.112	0.053	0.241	0.999	0.059	0.282	0.106
0.001%	0.062	0.157	0.174	0.074	0.041	0.097	0.999	0.055	0.074	0.082
0.01%	0.045	0.102	0.172	0.055	0.029	0.067	0.166	0.049	0.022	0.064
0.1%	0.035	0.065	0.166	0.048	0.022	0.042	0.162	0.045	0.006	0.008
FTA.E	0.008	0.132	0	0	-	-	-	-	-	-
FTA.P	0.011	0.233	0.001	0.012	-	-	-	-	-	-

The table shows FNMR at fixed FMR for the best four performing systems:
a) from IREX evaluation (obtained approximately from published DET curves shown in Figure 1), 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-pocessing using score calibration(+Calib) and using fusion (+Fused). The table also shows Failure to Acquire (FTA) of the systems tested on G-500 dataset (FTA.E is FTA for Enrolled images, FTA.P is FTA for Passage images. FTA = FTA.E + FTA.P).

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 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 is likely to be better than for a collection that did no such thing.

Because 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	

FTA numbers indicate the number of comparisons not performed due to Failure To Acquire.

III. THE CBSA "G-500" IRIS DATASET

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.

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 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

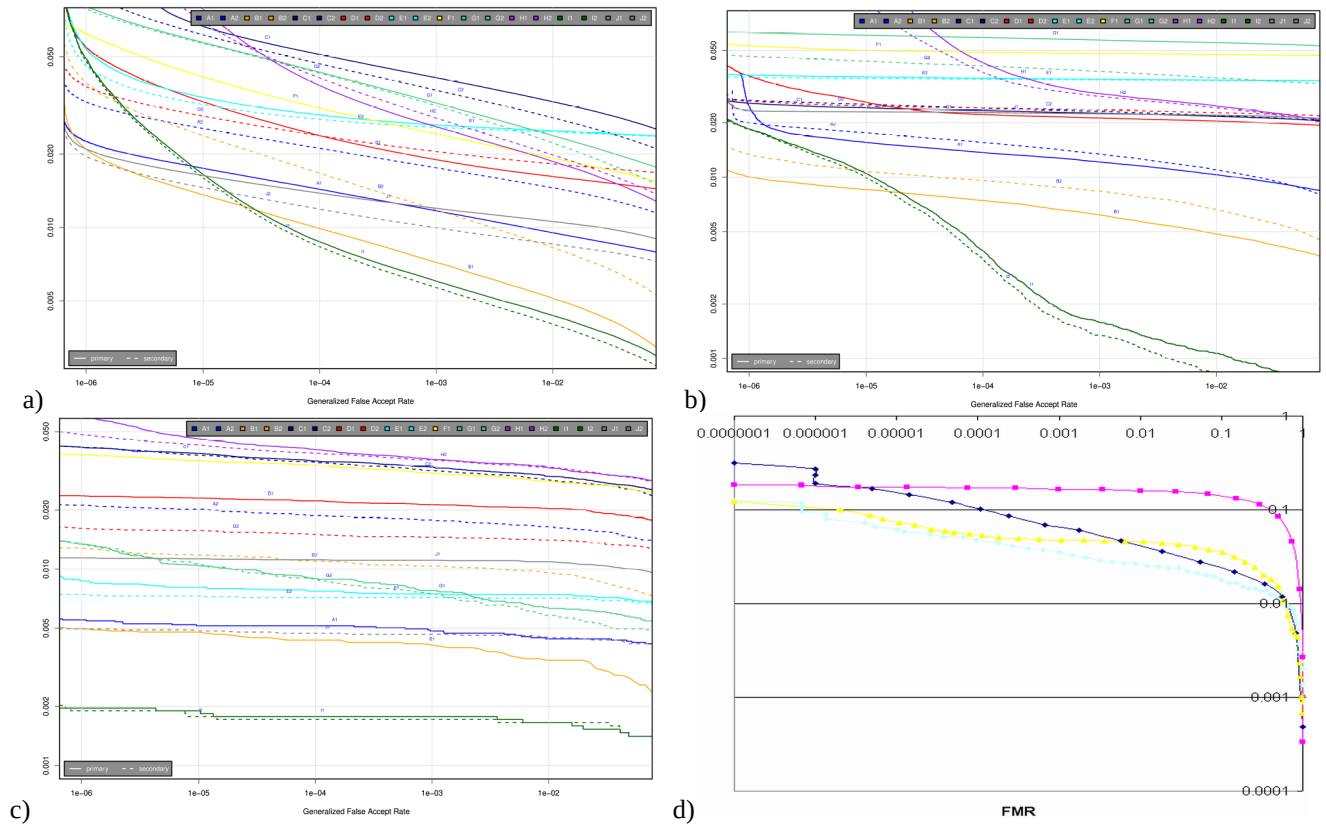

Fig. 1. *DET* curves obtained on four different datasets: ICE (a), BATH (b)and OPS (c) used by NIST (taken from[8] for uncompressed images), and G-500 (d) used by CBSA.

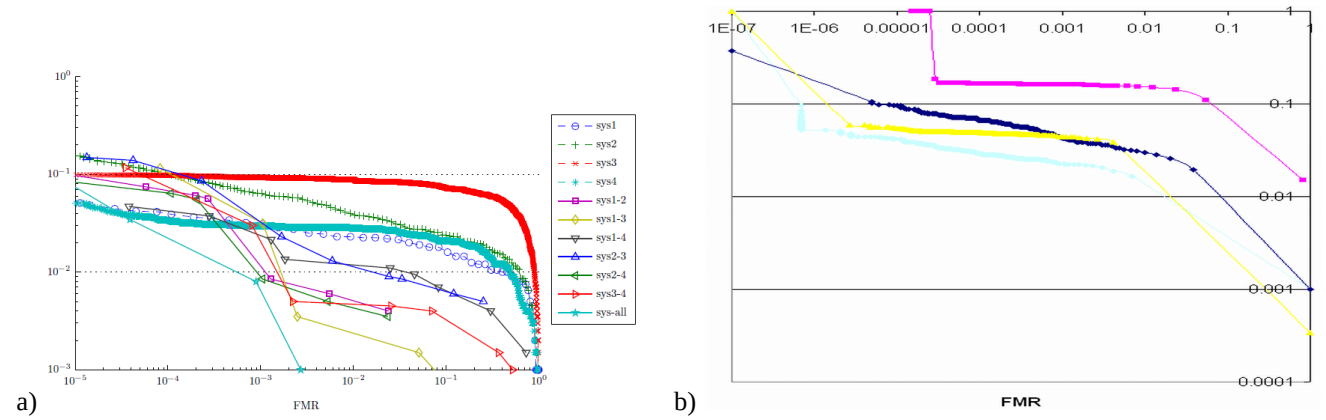

Fig. 2. *DET* curves obtained on the G-500 dataset with fusion (a) and score calibration (b). - Compare to Figure 1.d.

(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 G-500 dataset.

Issues, values and C-BET evaluation

While further analysis on the effect of image compression in the CBSA datasets and the fact that all their passage images have been already matched by the system need to be further explored, 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 the 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 [25], which provided the scientific and biometric user communities with guidelines on all-inclusive reporting of biometric system performances. 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 (fully automated access/border control vs. manual forensic investigation) and any capture environment (constrained and cooperative vs. unconstrained and covert).

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] a post-processing score calibration algorithm is proposed to improve the performance of biometric systems. The improvement is achieved by making the use of the a-priori knowledge of the genuine and impostor score distributions of 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 the $\hat{m}$ 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 $1 \leq i \leq n$, define $s_i = s(X, x_i)$ 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.

This SCF function replaces matching scores with meaningful confidence scores that are perfectly calibrated and normalized. This algorithm is 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 system with the G-500 dataset is shown in Figure 2.b (Compare to Figure 1.d).

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 long score normalization.

A ROC-based iterative BC (IBC) technique has been proposed in [14], [15] for efficient fusion of responses from multiple soft, crisp, or hybrid classifiers. IBC has been applied to the combination of responses from a multiple-HMM system for host-based intrusion detection. However, it represents a versatile information fusion technique for biometrics, where the ROC curves may result from a wide range of biometric systems designed with different traits, sensors, feature sets, classifiers, training data and/or user-defined parameters. This decision-level combination technique combines the ROC curve produced with a set of classifiers exploiting all Boolean functions, prior to applying the MRROC, and does not require any prior assumption regarding the independence of classifiers and the convexity of ROC curves. IBC has been shown to provide a significantly higher level of accuracy than related techniques in literature, especially when classifiers are trained with limited data, with a time complexity is linear with respect to the number of classifiers.

The main steps of BC_{ALL} are presented in Algorithm 1 shown in Figure 3. The BC_{ALL} technique inputs a pair of ROC curves defined by their decision thresholds, T_a and T_b, and the labels for the validation set. Using each of the ten Boolean functions, BC_{ALL} combines the responses of

each threshold from the first curve (R_{a_i}) with the responses of each threshold from the second (R_{b_i}). Responses of the fused thresholds are then mapped to points (fpr,tpr) in the ROC space. The thresholds of points that exceeded the original ROCCH of original curves are then stored along with their corresponding Boolean functions. The ROCCH is then updated to include the new emerging points. When the algorithm stops, the final ROCCH is the new MRROC in the Newman-Pearson sense. The outputs are the vertices of the final ROCCH, where each point is the results of two thresholds from the ROC curves fused with the corresponding Boolean function. These thresholds and Boolean functions form the elements of s^*_{global}, and are stored and applied during operations.

Algorithm 1: $BC_{ALL}(T_a, T_b, labels)$: Boolean combination of two ROC curves

Input: Thresholds of ROC curves, T_a and T_b, and *labels* (of validation set)
Output: ROCCH and fused responses (*Rab*) of combined curves

let $m \leftarrow$ number of distinct thresholds in T_a
let $n \leftarrow$ number of distinct thresholds in T_b
Allocate F an array of size: $[2, m \times n]$
BooleanFunctions $\leftarrow \{a \wedge b, \neg a \wedge b, a \wedge \neg b, \neg(a \wedge b), a \vee b, \neg a \vee b, a \vee \neg b, \neg(a \vee b), a \oplus b, a \equiv b\}$
Compute $ROCCH_{old}$ of the original curves
foreach $bf \in BooleanFunctions$ **do**
  for $i = 1, \ldots, m$ **do**
    $R_a \leftarrow (T_a \geq T_{a_i})$
    for $j = 1, \ldots, n$ **do**
      $R_b \leftarrow (T_b \geq T_{b_j})$
      $R_c \leftarrow bf(R_a, R_b)$
      Compute (tpr, fpr) using R_c and *labels*
      Push (tpr, fpr) onto F
  Compute $ROCCH_{new}$ of F
  Store thresholds and corresponding Boolean functions that exceeded the $ROCCH_{old}$, $s^*_{global} \leftarrow (T_{a_i}, T_{b_j}, bf)$
  Store the responses of these emerging points into R
  $ROCCH_{new} \leftarrow ROCCH_{old}$
Return $ROCCH_{new}, R, s^*_{global}$

Algorithm 2: $BCM_{ALL}([T_1, \ldots, T_K], labels)$: Cummulative combination of multiple ROC curves based on BC_{ALL}

Input: Thresholds of K ROC curves $[T_1, \ldots, T_K]$ and *labels*
Output: ROCCH of combined curves

$[ROCCH_1, R_1] = BC_{ALL}(T_1, T_2, labels)$
for $k = 3, \ldots, K$ **do**
  $[ROCCH_{k-1}, R_{k-1}] = BC_{ALL}(R_{k-2}, T_k, labels)$
Return $ROCCH_{K-1}, R_{K-1}$ and the stored tree of the selected responses/thresholds fusions along with their corresponding fusion functions

Fig. 3. IBC technique for fusion of system responses.

The BC_{ALL} technique makes no assumptions regarding the independence of the detectors. This techniques directly fuses the responses of each decision threshold, accounting for both independent and dependent cases. In fact, by applying all Boolean functions to combine the responses for each threshold, it implicitly accounts for the effects of correlation. In the worst-case scenario, when the responses of detectors provide no diversity of information, or when the shape of the ROC curve on the validation set differs significantly from that of the test set, the BC_{ALL} is lower bounded by the MRROC of the original curves.

Exploiting all Boolean functions accommodates for the concavities in the curves. The AND and OR rules will not provide improvements for the inferior points that correspond to concavities and make for an improper ROC curve, or points that are close to the diagonal line in the ROC space. Other Boolean functions, for instance those that exploit negations of responses, may however emerge. The BC_{ALL} technique can therefore be applied even when training and validation data are limited and heavily imbalanced, to combine the decisions of any soft, crisp, or hybrid detectors in the ROC space.

As described in Algorithm 1, the thresholds (T_1 and T_2) of first two ROC curves are initially combined with the BC_{ALL} technique. Then, their combined responses (R_1) are directly input into line 8 of Algorithm 1 and combined with the thresholds of the third ROC curve (T_3).

Further improvements in performance may be achieved by re-combining the output responses of combinations resulting from the BC_{ALL} (or BCM_{ALL}) with those of the original ROC curves over several iterations. A novel Iterative Boolean Combination (IBC_{ALL}) is presented in [15] and allows for combination that maximize the AUC of K ROC curves by re-combining the previously selected thresholds and fusion functions with those of the original ROC curves. During the first iteration, the ROC curves of two or more detectors are combined using the BC_{ALL} or BCM_{ALL}. This defines a potential direction for further improvement in performance within the combination space. Then, the IBC_{ALL} proceeds in this direction by re-considering information from the original curves over several iterations. 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. The iterative procedure stops when there are no further improvements to the AUC or when a maximum number of iterations are performed.

BC_{ALL} is efficient in scenarios with limited and imbalanced data because the number of distinct thresholds is typically small. Assume a pair of detectors, C_a and C_b, having respectively n_a and n_b distinct thresholds on their ROC curves. During the design process, the worst-case time complexity (required for computing all ten Boolean functions to combine thresholds) and memory complexity (required to store the temporary results (tpr, fpr) of each Boolean function) of IBC_{ALL} is $\mathcal{O}(n_a n_b)$. When the BCM_{ALL} is applied to combine the ROC curves of K detectors, the worst-case time can be roughly stated as K times that of the BC_{ALL} algorithm. However, after combining the first two ROC curves, the number of emerging responses on the ROCCH, is typically very small with respect to the number of thresholds on each ROC curve.

Figure 2.a displays the results achieved on the G-500 data by applying the IBC technique for various decision-level combination of responses from the four different iris systems. For an unbiased evaluation and combination of the systems, the failure to acquire images have been filtered out. The G-500 data set is reduced to $2,000$ Genuine and $975,077$ Impostor images.

VI. Discussion

This paper explored the performance limits of iris biometrics, but summarizing the results obtained to date, and

put together the best reported DET curves and FMR/FNMR results obtained for this modality. The results are obtained on 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 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. Next steps

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), which show DET curves for the same four products with and without post-processing score calibration. — A system with worse DET curve may be better performing than the system with better DET curve, because it may allow 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]. Therefore, to get "the entire story" about a biometric product or biometric modality it is recommended that the multi-order biometric score analysis developed in [9], [10], [11] be used.

Additionally, to further understand the limitations of the modality or system, subject-based performance evaluation, known as biometric menagarie or Doddington's zoo analysis [26], should also be conducted. Rephrasing the Doddington's zoo terminology into 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" ("sheep"), who rarely/never get False Match or Non-Match errors, 2) "happy but causing risk" users ("wolf"), who rarely/never experience False Match, but who often generate False Non-Match errors thus creating higher security risk for using the system, 3) "frustrated, but causing no risk" ("lamb"), who frequently experience False Match problem, but do rarely generate False Non-Match errors, and finally 4) "frustrated and unsafe" users ("goat"), who get frequently both False Match or Non-Match errors. To see how well iris system performs on different types of people, the subject-based analysis of iris scores should be conducted, as discussed in [8], [27]. This is topic of our currently research, the results of which will be published soon.

For the Government of Canada's Biometric Community of Practice users and partners, these and other recommendations related to all-inclusive evaluation of biometric systems are now summarized in the Comprehensive Biometric Evaluation Toolkit (CBET) posted on the dedicated CBET portal [25], maintained in partnership of the Canada Border Services Agency's Science and Engineering Directorate and the Defence Research and Development Canada's Center for Security Science.

ACKNOWLEDGMENT

The support of the Defence Research and Development Canada's Center for Security Science (DRDC-CSS) for the development of the C-BET, including the creation of the DRDC-CSS-hosted CBET portal for sharing C-BET methodology with the Government of Canada's Community of Practice, is acknowledged. The stimulating discussions with the colleagues from the Solutions directorate are also gratefully acknowledged, as is the help of many S&E COOP students, who have contributed to the conducting of iris product tests and the developing of the C-BET software.

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.schiphol.nl/Travellers/AtSchiphol/PriviumIrisscan/WhyPrivium/FastBorderPassageWithTheIrisScan.htm
[3] http://www.globalentry.gov/netherlands.html
[4] http://www.ukba.homeoffice.gov.uk/travellingtotheuk/Enteringtheuk/usingiris/
[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. Submitted to SSCI, CIBIM 2011
[12] Gorodnichy, D.O., Hoshino, R. Calibrated confidence scoring for biometric identification. NIST International Biometric Performance Conference (IBPC 2010), March 2-4, 2010 - on-line at http://www.nist.gov/itl/iad/ig/ibpc2010_presentations.cfm.

[13] Gorodnichy, D. O., Hoshino, R. 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, “Combining Hidden Markov Models for Anomaly Detection,” IEEE Intl Conf. on Communications, Dresden, Germany, June 14-18, 2009.

[15] W. Khreich, E. Granger, R. Sabourin, A. Miri, “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] CBET Portal : https://partners.drdc-rddc.gc.ca/css/Portfolios/Biometrics (Human ID Systems)/C-BET

[26] 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

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