

Multi-order Biometric Score Analysis Framework and its Application to Designing and Evaluating Biometric Systems for Access and Border Control

Dmitry O. Gorodnichy
Member, IEEE

Abstract—Traditionally, automated access and border control biometric systems are thought of and designed as verification 1-to-1 systems, where a single comparison between a probe and the claimed identity is examined to allow or disallow the entry to a person; and as such they have been evaluated to date – by using the error tradeoff statistics, which counts how many times a person was falsely accepted or rejected. Such a design however may soon become obsolete due to the recent shift towards applying biometrics to free-flow surveillance-like environments and also in the light of recent findings showing that performance of many verification systems can be improved through the use of several 1-to-N scores, instead of relying on a single 1-to-1 score only. As the framework for designing biometric-enabled access and border control systems changes, so has to change the methodology for the evaluation of such systems. This paper addresses this problem by establishing the multi-order biometric score analysis framework. The framework incorporates latest innovations and recommendations related to the comprehensive evaluation of biometric systems, including subject-based analysis, calibrated score analysis, and two new performance metrics: threshold-validated recognition ranking and non-confident decisions due to multiple threshold-validated scores. The framework is implemented in the Comprehensive Biometrics Evaluation Toolkit (C-BET) and has been applied for the evaluation of several biometric modalities, in particular, those that are frequently contemplated for the use in unconstrained access-border control applications, such as face, voice and iris. The results of the iris modality evaluation are presented in this paper.

I. INTRODUCTION

TRADITIONALLY, automated access and border control biometric systems are thought of and designed as verification 1-to-1 systems, where a single comparison between a probe and the claimed identity is examined to allow or disallow

Dmitry O. Gorodnichy is a Senior Research Scientist with the Video Surveillance and Biometrics Section of the Science and Engineering Directorate of the Canada Border Services Agency (CBSA-S&E), 14 Colonnade Road, Ottawa, Ontario, Canada, K2E 7M6.

Citation: Dmitry O. Gorodnichy, “Multi-order Biometric Score Analysis Framework and its Application to Designing and Evaluating Biometric Systems for Access and Border Control”, 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.

Crown Copyright 2011. Permission is granted to quote short excerpts and to reproduce figures and tables from this report, provided that the source of such material is fully acknowledged.

the entry to a person; and as such they have been evaluated to date — by using the error tradeoff statistics, which counts how many times a person was falsely accepted or rejected [1]-[11]. Such a design however may soon become obsolete due to the recent shift towards applying biometrics to free-flow surveillance-like environments, where person’s identity is retrieved and verified without person’s providing his/her name to the system [1]; and also in the light of recent findings showing that performance of many 1-to-1 verification systems can be improved by performing post-processing recalibration of 1-to-N comparison scores instead of using a single 1-to-1 comparison score [13], [14].

As the framework for designing biometric-enabled access and border control systems changes, so has to change the methodology for the evaluation of such systems.

This paper expands upon our previous efforts in addressing this problem and further extends the multi-order biometric analysis framework introduced in [15], [16]. The refinement of the framework is proposed to combine rank-based evaluation, which is traditionally used for evaluating investigation-type one-to-many identification systems, with threshold-based evaluation, which is done for biometric-enabled access/border control systems. As such it allows one to evaluate the applicability of the systems that have been traditionally used in semi-automated investigation mode for the applications where fully-automated biometrics-based decision is required.

The paper is organized in two parts. The first part presents two motivation factors for this work — one relating to the deficiency of the currently used designs for access / border control biometric systems, and the other relating to the limitations of the currently existing evaluation practices. It then presents the taxonomy for biometric systems, which categorizes biometric systems by the application mode in which they are used (fully-automated vs. semi-automated) rather than by design (1-to-1 vs. 1-to-N), and introduces the concept of 1-to-N verification. It is also in this part, where the C-BET evaluation paradigm and main objective are presented.

The second part of the paper is dedicated to the presentation of the best practices and recommendations for the comprehensive evaluation of biometric systems. It refines the multi-order score analysis terminology to include its application to the system design, and introduces threshold validated analysis and two new performance metrics: threshold-validated recognition

ranking and non-confident decisions due to multiple threshold-validated scores. It integrates the new concepts with subject-based performance analysis and other good evaluation practices and presents the results from C-BET evaluation of iris modality.

The paper concludes with the summary of innovations proposed in the paper and the discussion on the future work.

II. ONE-TO-MANY VERIFICATION

The concept of the *one-to-many verification* may sound contradicting to the very definition of *verification*, which is defined by many as a process involving a single one-to-one comparison between the unknown sample (probe) and the claimed identity sample (enrolled sample). One of the objectives of this paper is to break this stereotype and to show that verification system may and should be designed as 1-to-N systems, in particular, for access / border control applications.

The concepts of Access and Border Control are intentionally merged into a single concept of “access / border” or “entry” control. This is done to raise the awareness of the fact these are not two different applications, as currently treated by the industry [1], [2], [3], but is the same application driven by the same dual performance optimization objective, as further described below.

A. Two performance objectives of entry control systems

Table I summarizes the taxonomy and evolution of biometric system designs and evaluation methodologies. Following this table, access / border (entry) control systems are common in that both are designed with the same two concurrent performance optimization objectives in mind.

Definition: *Biometric-enabled access / border (entry) control systems are fully-automated systems designed to achieve the following two objectives:*

Objective 1: *to make access easier for registers users.*

Objective 2: *to make access harder for non-registered users.*

While biometrics has been applied for virtual and physical access control systems for several decades, its application for border control has started relatively recently and is growing fast. — With thousands of people crossing country borders every minute and with the ever increasing need to make cross-border travel both secure and efficient, Biometric-Enabled Automated Border Control (BE-ABC) is now seen as one of the most promising applications of biometrics.

Following the definition, a BE-ABC system has to achieve two main entry control objectives.

On one hand (Objective 1), the system should handle quickly and in a friendly manner a large volume of travelers, most of whom are bona-fide citizens, possibly tired or stressed due to travel constrains and duration, and possibly coming from various language and culture backgrounds. For these travelers, a system should provide an easy biometrics sample

scanning mechanism, which ideally would not require much (or any, if possible) interaction with the system. In another words:

For the end-user, a 1-to-N design, in which a system automatically retrieves a traveler's identity from a list of pre-enrolled N travelers, can be more advantageous to the 1-to-1 design, which requires a traveler to carry an additional card or interact with the system.

On the other hand (Objective 2), the system should be designed so that to minimize to the possible lowest level the risk of an Impostor to get through the system. To achieve this the environmental and procedural conditions and constraints would be normally carefully chosen and tuned to optimize the system performance, where the performance is measured in terms of False Match / Non-Match Rates (FMR/FNMR) and Detection Error Tradeoff (DET) curves. In doing so, however it is often forgotten that the system design itself may not be the most optimal. Specifically, as proposed by Gorodnichy & Hoshino [13], [14], the performance of 1-to-1 systems can be improved by performing 1-to-N comparisons instead of a single 1-to-1 comparison through the use of the post-comparison score calibration. This is demonstrated in Figures 1.b and 1.d that show FMR/FNMR statistics and DET curves for one of the best commercial off-the-shelf iris biometric products, the verification performance of which is improved using the techniques proposed in [13], [14]. In another words:

For the developer, a 1-to-N design, in which a system looks at all N scores of enrolled travelers, can be more advantageous to the 1-to-1 design, in which a system makes the verification decision based on a single comparison only.

B. Relationship to 1-to-N screening

One of the advantages of the 1-to-N design for a biometrics system is also seen in the fact that it allows one to apply the same system to “Watch List” screening tasks. For example, an “iris on the move” stand-off iris recognition system that is designed for expedited process of pre-registered travelers can also be used to screen the traffic of people against the wanted individuals. Similarly, voice or face biometric engines that have been mostly used for 1-to-N investigation purposes may be applied for fully-automated biometric-enabled access/border control.

III. TWO BIOMETRIC SYSTEM APPLICATION MODES AND THE NEED FOR BETTER EVALUATION PRACTICES

As summarized in Table I, regardless of biometric modality or biometric recognition task, there are essentially two application modes, in either of which a biometric system can be used:

Mode 1: Semi-automated or investigation mode, *where a system provides a number of options for a human*

TABLE I
TAXONOMY AND EVOLUTION OF BIOMETRIC SYSTEM DESIGNS AND EVALUATION METHODOLOGIES.

Application mode	Fully-automated / Instant Recognition (Mode 2) Based solely on captured biometric data				Semi-automated (Mode 1) Based on all available (not necessarily biometric) data about the subject	
	more constrained		→		less constrained	
Domain	Access / Border (Entry) Control Restricted access areas, Member / Pre-approved Traveller Programs Biometric-enabled Automated Border Control (BE-ABC)			Public Surveillance	Forensic / Investigation	
High-level objective	Recognize "Good" people			Recognize "Bad" people	Assist analyst to recognize a person	
Possible Recognition task(s)	Authentication (Verification)			Classification / Categorization Tagging / Tracking, Screening / Identification	Similarity Quantifier Positive and Negative Identification	
Performance objectives	1: Maximize convenience for members 2: Minimize the risk of spoofing			Maximize amount of gathered Intelligence		
System design	Card-based (Biometric stored on user card)	Biometric stored on server				
Level of user interaction		Input-based (user identifies him/her-self)	Input-less "Stand-by" (user presents h/s)	Input-less "Stand-off" in constrained settings	Input-less "Stand-off" in unconstrained settings	With input from analyst or Input-less
User participation	Cooperative			Non- & Cooperative	Un- & non-cooperative	
Recognition time	Instant	Real-time				Sufficiently long
Biometric modalities	fingerprint, iris	Single iris	Dual Iris, Iris + face/voice	(Dual) Iris (+ face) (+ voice)	(Dual) Iris (+ face) (+ voice)	face voice
Number of biometric samples & modalities	Single sample single modality	Single sample single modality	Multiple sample single modality	Multiple samples multiple modalities		Single/multiple modality single/multiple sample
Conventional design and evaluation						
Matching Design	1 to 1	1 to 1	1 to N (1 to First)	1 to N (1 to First)	1 to K, 1 to M (K<M<N)	1 to N
Decision Principle	Threshold	Single score + Threshold				Ranked scores + Manual
Evaluation Metrics	DET curves	???	DET curves	???	FMR / FNMR	???
Multi-order score analysis based design and evaluation						
Matching Design	Order-1	Order-1 (1 to 1) + Order-2 (1 to N) + Order-3 (confidences)				More comprehensive
Decision Principle	Threshold	scores & confidences + Threshold & AI				Rank scores & confidences+ Manual
Evaluation Metrics	Order-1	Order-1 (DET+ more) Order-2 (CMC + more) Order-3 (FCR + more)				More comprehensive

Biometric-enabled entry control systems have been evolving from card-based verification systems to input-less systems working in surveillance-like environments. However, their design and evaluation methodologies remained practically the same – largely limited to single score analysis. This paper establishes new biometric design and evaluation methodology based on the multi-order score analysis. This allows biometrics designers and users to optimize the performance of their biometric-enabled access and border control systems, by addressing the two key performance objectives of the systems, and adopting advanced Artificial Intelligence techniques for improved automated decisions.

analyst, who makes the final decision within a sufficiently long period of time based on available data about the subject.

Mode 2: Fully-automated or access/border (entry) control mode, where the final decision is made by a system instantly (or in real-time) based solely on the biometrics sample(s) collected from the subject.

Traditionally, when the performance of a biometric system needs to be evaluated, the evaluation metrics are chosen to match the anticipated application mode of the system. In particular, if the system's intended use is to assist a human analyst to recognize a person (Mode 1), then the system is called a 1-to-N identification system and identification rates and the corresponding *Cumulative Match Characteristic (CMC) curves* are computed.

On the other hand, if the system is developed for access control (Mode 2), then the system is called a 1-to-1 verification system and verification single score based statistics such as FNMR, FMR and DET/ROC curves are computed.

The questions arise (refer to Table I): - What if the application mode of the system is not known? - Or if a modality that has been conventionally used in investigation (non-automated) mode is examined for its suitability in fully automated applications, which is the case with Face and

Voice modalities that are now contemplated for Access/Border Control and Public Surveillance tasks? - Or vice versa, when a modality that has been traditionally used in constrained cooperative environments is examined for its suitability in unconstrained (or much less constrained), un-cooperative (or non-cooperative) environments, which is the case with Iris modality and "Iris on the move" and other stand-off biometric technologies? Which Figures of Merit and statistics should be used then?

In response to these questions and driven by the operational need to better understand the technology that could be potentially deployed in the field, the Comprehensive Biometric Evaluation Toolkit (C-BET) framework has been developed. The framework is designed to supplement the results that would normally be reported elsewhere (eg. NIST), but so that to provide a deeper understanding or a "better feel" of the "Black Box", which the biometric system is, through the investigation of all higher detail information that could be possibly extracted and inducted through the analysis of the system performance.

In the conception of the C-BET framework, another important observation related to the biometric evaluation process was instrumental.

Large scale evaluation of a biometric system can be considered as a three-stage process. The amount of time and

effort required to prepare the testing datasets that contain large number of Genuine and Impostor sample pairs (Stage 1) and the amount of time and effort required to learn a biometric product and to have it run on the entire dataset to obtain all Genuine and Impostor comparison scores (Stage 2) is normally much more significant than that of the final task of processing all computed scores and reporting the obtained performance statistics and graphs (Stage 3).

It appears therefore unfortunate that after an immense effort invested in the first two stages of the evaluation process leading to the computation of all scores, it is only a fraction of the statistical analysis, which could be potentially obtained from all computed scores, that is reported.

In many cases, once an evaluation report is published, the score data that has been used to generate the report will be discarded, and neither the user nor the developer of the system will ever know “the rest of the story” about the performance of a biometric “black box”!

The C-BET evaluation methodology and the toolkit are developed to allow one to report “the entire story” about a biometric system’s performance.

IV. MULTI-ORDER SCORE ANALYSIS

Multi-order score analysis is introduced in [15], [16] as an important biometric performance methodology that facilitates the investigation into the risks and risk mitigation factors related to having non-confidence outputs in fully-automated biometrics systems. It was inspired by and originally applied to the evaluation of commercial iris biometrics systems such as those that can be potentially used for the CBSA-operated NEXUS traveler program [22].

The multi-order terminology for the analysis comes from the analogy with multi-order statistics, in which order-0 statistics signifies using the value itself, order-1 statistics signifies computing the average of several values, and order-2 and order-3 statistics signify computing the deviation (variance) and high-order statistical moments. Similar to statistics, the scores of a biometric system can be analyzed at several levels (or orders) of detail to provide incrementally more information for better decision making.

A. Multi-order score analysis for better system design

As shown in [13], [14], the multi-order score analysis can be used to improve the overall performance of a biometric system, when applied as a post-processing score recalibration filter. The usage of the multi-order score analysis terminology for biometric system design is illustrated by the following example.

Consider an iris recognition system with the matching threshold set at $T = 0.33$. When a probe iris image is compared against the images of five (different) people in an enrollment database, five matching scores are obtained (0.45, 0.32, 0.47, 0.34, 0.31). The Order-1 system, which makes the decision based on the assumption that being below a threshold is sufficient for the recognition decision, finds the first score below the threshold (0.32) and reports match for the 2nd

person. The Order-2 system finds the smallest score below threshold (0.31) and reports the match for the 5th person. The Order-3 system however will not simply report the match for the 5th person, but would also assign a confidence value to this match based on all score information available, which in this case will be low, since the score of the 4th person (0.34) is also close to the threshold.

B. Multi-order score analysis for better system evaluation

When used within a comprehensive performance evaluation procedure, the multi-order score analysis is shown to provide additional insights on the system performance and reliability, and to expose the risks due to non-confident recognition decisions [18], [20].

In the following, we present the definitions related to the multi-order score analysis and the C-BET comprehensive biometric performance evaluation based thereon.

This analysis/evaluation has been used in evaluating the applicability of iris, voice and face modalities for fully-automated access/border entry control applications. The results from the C-BET evaluation of iris systems are presented in this paper and are used to illustrate the new concepts described in the paper. The evaluation has been conducted with the CBSA-developed G-500 iris dataset, the description of which is given in [19]. The results from the C-BET evaluation of voice biometrics are presented in [20].

V. DEFINITIONS AND GOOD EVALUATION PRACTICES

A. Order-0 and Order-1 analysis

Definition: Order-0 score analysis is defined as the statistics and visual analysis of the probability distributions of Genuine and Impostor scores.

Order-0 analysis is the visual analysis and it does not produce a performance metric in itself, yet it is found very useful to provide insights on how a system performs and where the performance bottlenecks could be.

Order-0 visual score analysis is shown in Figure 1.c. Such visual analysis about an unknown “Black Box” biometric system should always be obtained first, prior to further examination of the system, because it reveals the inner properties of the system. Particularly, it can be used to obtain the a-priori probabilistic distributions of Genuine and Impostor matching scores, which can then be used to maximize the probability of more reliable decisions through post-processing score calibration proposed by Gorodnichy & Hoshino in [13], [14]. It also summarizes the properties of the dataset, such as the number of genuine and impostor comparisons used in the evaluation, which can be used to obtain the FMR/FNMR confidence bounds [12].

Definition: Order-1 score analysis is based on computing and analyzing single matching score statistics, as in fully-automated 1-to-1 verification systems and when plotting DET/ROC curves.

In traditional terminology, Order-1 analysis can be viewed and referred to as the “verification analysis”, which is conventionally performed for automated access/border control systems.

Figures 1.d,e,f show Order-1 score analysis results. When plotting DET / ROC curves, all measured points should be explicitly shown on the curve. Showing only the extrapolated curves may mislead people into believing that certain rates are achievable by a system, when they are not. Showing the measured points can also serve to validate the appropriateness of the threshold increments used in conducting the evaluation.

Additionally, to avoid misinterpretation, for plots drawn using logarithmic scales, it is recommended to mark points corresponding to FMR / FNMR equal to zero as **Virtual 0**, as shown in Figure 1.d.

B. Order-2, Order-3, and Threshold-Validated analysis

Definition: Order-2 score analysis is based on computing multiple matching scores and analyzing the score ranking statistics (or best K scores), as in 1-to- N comparisons used in investigative-mode recognition and when plotting the CMC curves.

In traditional terminology, Order-2 analysis can be viewed and referred to as the “identification analysis”, which is conventionally performed for 1-to- N investigative systems.

Figures 1.i-j show Order-2 score analysis, which plots the number of instances when the genuine match was the best, second best etc. These curves can be seen as the derivative of the traditional CMC curves, which are used for evaluation of biometric identification systems for forensic purposes. The CMC curves show the integral value of the identification rank, indicating that the genuine score was among the best K scores without specifying whether it was the K th-best, $K-1$ th best or the best score.

The reason for plotting Order-2 score curves as shown in Figures 1.i-j and not as traditional CMC curves is to provide more information about the system.

It also allows us to apply the *Threshold Validated* terminology described below, according to which each matching score is labeled either as *Threshold Validated (TV)* or *non-Threshold Validated (non-TV)* depending on whether it passed a comparison to the threshold or not, i.e. whether it is smaller (or higher, depending on system design) than the threshold.

Definition: When a biometric comparison score passes the system matching threshold it is called **Threshold Validated**.

Such definition is introduced to avoid referring to the scores as “Matched” or “Accepted” (as they are traditionally called in 1-to-1 verification systems), because the final “Match” / “Accept” system decision with the high-order score analysis may not only be based on the score comparison to the threshold, but other factors such as confidence.

In particular, the concept of Threshold Validated biometric identification becomes vital when designing card-less / input-less biometric-enabled access and border control systems, in which a person’s identity is recognized from a list of pre-approved enrolled individuals (see Table I).

It also becomes very important when evaluating the applicability of biometric systems that have been conventionally used for manual investigation recognition tasks (such as face or voice recognition systems used by police) to the tasks where instant fully-automated recognition by a computer is required. These include the traditional biometric recognition applications such as “Black List” screening and identification as well as the extended biometric recognition applications such as triaging, classification, categorization, tagging, and tracking, which have become of high interest to many agencies, because of their applicability to video surveillance applications [15].

For an open dataset, Rank 0 signifies that a probe sample is not in the enrolled dataset, which is the case in Figures 1.i-j. In this case, Threshold Validated analysis provides the information on the likelihood that a random person can pass through the system.

The Threshold Validated terminology becomes also very useful in defining and applying Order-3 score analysis, which looks at all the relationship among the scores and their relationship with a threshold.

Definition: Order-3 score analysis is based on computing / analyzing the relationship between the match scores, as when finding the difference between the best and second-best match scores, finding all threshold-validated scores, or when applying the post-processing recalibration of the scores based on 1-to- N comparisons.

Order-3 score analysis results are shown in Figure 1.b and Figures 1.f-g, j-k, l-m. Figure 1.b presents the Performance Report Summary table, which shows, in addition to FMR and FNMR, the Failure of Confidence Rate (FCR) for each threshold. FCR, introduced in [17], is defined as the number of instances when there is more than one Threshold-Validated match for a probe.

Figure 1.g-h shows the rate of recognition confidence measured in terms of the normalized distance from the best score to the second best score - for genuine best matches (Figure 1.g) and for Impostor best matches (Figure 1.h). Ideally, one would like to have low confidence for best score if it belongs to an Impostor, and high confidence if it belongs to a genuine comparison.

Figure 1.k-l shows statistics on the number of Threshold Validated matches. Additionally, the number of those Threshold Validated matches that scored the best are marked for each genuine and Impostor match. Ideally, for a fully automated system, one would like to have only one Threshold Validated match which is the best and corresponds to the genuine comparison (Yellow colour). If it corresponds to the Impostor comparison (Blue colour), then such system cannot be used for automated access/border control. At the same time, even

if there is more than one Threshold Validated match, but the genuine score is the best (Blue colour), then such a system has a good potential to be used for automated decision making, provided that it is designed to maximize the confidence of its decisions through the use of Order-3 score analysis.

Finally, Figure 1.m-n shows the threshold-based analysis summary introduced in [20], further described below.

C. Threshold-based analysis summary

The threshold-based analysis summary presents the statistics with respect to the six possible recognition outcomes of the system, as shown in Table II.

TABLE II
SIX OUTCOMES OF THRESHOLD-BASED ANALYSIS

G (bt) T (bt) I	G(bt)I(bt)T	I(bt)G(bt)T	I(bt)T(bt)G	T(bt)G(bt)I	T(bt)I(bt)G
BEST	OK to Accept *	OK to Reject*	BAD	OK**	OK**

G signifies the score of the Genuine match, *I* signifies the score of the Impostor match, *T* stands for Threshold, *(bt)* stands for “better than” and is either $<$ or $>$ depending on the system design. (*) indicates that more processing is required for the recognition decision to be done. (**) indicates that data may be of insufficient quality and another sample may need to be taken.

The six possible outcomes of the threshold-based statistics are sorted according to their meaning for the system performance description. Outcome 1 is an ideal outcome for a fully-automated system. Outcome 4 is the worst outcome. Outcomes 2 and 3 indicate that additional processing is required for the system to be operational - either done by a human analyst, or done by a computer through the higher order score analysis. Outcomes 5 and 6 are indicative of the fact that either the data is not reliable or the biometric modality is not sufficiently constrained.

The threshold-based analysis summary can be obtained directly from the Threshold Validated Order-2 and Order-3 analysis described above (Figure 1.i-j and 1.k-l). As presented in [20], it provides a succinct way to describe the performance of the Mode-2 (fully-automated) systems.

D. Calibrated score analysis

As discussed above, the performance of certain systems can be improved through the post-processing score calibration, based on Order 3 score analysis with the technique proposed by Gorodnichy & Hoshino in [13], [14]. Therefore, FMR/FMNR and DET curves can be computed using the original system scores and also using the calibrated scores, as shown in Figures 1.b and 1.d.

The fact that the performance of a closed-box commercial off-the-shelf biometric product can be further improved using post-processing may become as a surprise to biometric users and illustrates that biometric technology still have room for improvement, in particular, by using better design principles such as those described in this paper.

E. Subject-based analysis

In order to further investigate the performance of a modality or a system, a subject-based performance evaluation, known as biometric menagerie or Doddington’s zoo analysis [23], [24], should also be conducted. Rephrasing the Doddington’s

(“sheep-lamp-wolf-goat”) zoo terminology into a biometric-enabled access/border control context, the biometric system performance may vary substantially for different types of users. In particular, four types of users 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 of several iris systems using the G-500 dataset. In this analysis, FNMR statistics is computed by counting the number of all enrolled users who are falsely rejected by the system, instead of counting the number of mismatched comparison transactions. The obtained results, shown Figure 1.b, are very revealing (the results from other iris systems are given in [19]).

In particular, *the subject-based analysis have shown that the percentage of enrolled members who would experience a false rejection problem (i.e. subject-based FNMR) is higher than the conventionally reported FNMR computed by averaging over all performed matches.*

This does not come as a surprise though, because it is understood that the transaction-based FNMR equals subject-based FNMR only when the mismatched transactions are evenly distributed over all subjects, which is rarely the case even in a well balanced dataset such as the G-500 dataset.¹

The situation however is even worse when FNMR numbers are measured in a live operational context. In real-life (pilot) testing of a biometric system performance, the transaction-based results will show even more skewed results, because normally the subjects who have more biometrics transactions are those who have less problem using it. *The subjects who have higher false rejects rates may have much less transactions recorded, than those who do not experience false rejects.*

To illustrate the point, refer to the results of a recently conducted live pilot-based evaluation of a stand-off iris system. Based on 1694 transactions from 49 volunteers a low FNMR of less than 2% is reported. However, the report did not indicate how many of these 49 volunteers have contributed to this FNMR. It is therefore impossible to conclude whether the system performance is good or bad. — It could have been only one person who was falsely rejected several times, or it could have been ten people who were falsely rejected, but the number of their transactions was outweighed by those who were never falsely rejected.

It is therefore strongly recommended, especially when the dataset is not large or when evaluation is done in a live pilot, to report subject-based performance measurements, instead of (or in addition to) transaction-based measurements.

¹The CBSA G-500 dataset has exactly 6 genuine and 499 impostor matches performed for each of 500 enrollees.

F. Evaluation summary report and the C-BET software

“Making a sense of it all” is a big challenge in any application where a significant number of data is generated and need to be analyzed and understood. This has recently become the driving force for the creation of a new and very demanded Science and Technology area called “Visual Analytics”.

Once the recommendations for proper and all-inclusive evaluation of biometrics systems are provided, there still remains a question on how to efficiently implement these recommendations and how to efficiently report the obtained findings.

An effort therefore has been made to develop the convenient report layout for presenting and summarizing C-BET results, which provides all-inclusive, self-explanatory, succinct and efficient, description of systems performance - in Visual Analytics sense. Figure 1 shows such a layout. In a two-page briefing note style, the report graphically shows the main results of the multi-order score analysis of the system performance.

Following the establishment of the new theoretical foundation for proper all-inclusive evaluation of biometrics systems, the CBSA-S&E has also developed the C-BET visual analytics software that can automatically generate the C-BET-defined performance graphs and metrics discussed in this paper for any large-scale evaluation of any biometric modality or product [18]. It is implemented as a JAVA program that takes all scores obtained through the evaluation and instantaneously generates multi-sheet MS EXCEL files containing easy to browse and analyze graphs and metrics related to the system performance. As such, the program allows one to quickly compare biometrics systems to one another and to efficiently select optimal threshold and other system parameters by visually comparing images to one another.

VI. CONCLUSIONS

By definition, biometrics is an automated technique or system used for person recognition based on their biometric traits. With respect to this definition, this paper introduced a number of innovative concepts and methodologies for designing and evaluating biometric systems, which are summarized below.

→ Instead of the traditional taxonomization of biometric systems as being either a 1-to-1 verification or a 1-to-N identification system, the paper proposes a novel taxonomy according to the mode of operation as a fully-automatic (or access/border control, or Mode-2) system or semi-automated (or investigation, or Mode-1) system.

→ The concept of 1-to-N verification is introduced within the taxonomy and evolution of fully-automated biometric-enabled solutions. This concept allows us to redefine the way the design and evaluation of access / border control systems is done.

→ The development of the Comprehensive Biometric Evaluation Toolkit (C-BET) framework is further motivated – as a tool to produce “the entire story” about biometric “black boxes”.

→ The multi-order biometric score analysis framework is

further refined to include its application to biometric system design. Additionally, Order-2 and Order-3 score analysis is extended to include two new performance metrics: threshold-validated recognition ranking and non-confident decisions due to multiple threshold-validated scores. This makes it possible to combine rank-based evaluation, which is traditionally used for evaluating investigation-type one-to-many identification systems, with threshold-based evaluation, which is done for biometric-enabled access/border control systems. As such it allows one to evaluate the applicability of the systems that have been traditionally used in semi-automated investigation mode for applications where fully-automated biometrics-based decision is required.

→ Best practices for reporting traditional single-score-based metrics, referred to as Order-0, Order-1 and Order-2 score analysis, are summarized.

→ Subject-based score analysis, the importance of which is shown using a real-life iris data example, is made part of the comprehensive evaluation, as is the analysis of the calibrated scores using the Gorodnichy-Hoshino postprocessing calibration.

→ The results presented in this paper along with other recommendations related to the all-inclusive evaluation of biometric systems are made available for the Government of Canada’s Biometric Community of Practice users and partners, at the dedicated C-BET portal [21] maintained in partnership of the CBSA-S&E and Defence Research and Development Canada’s Center for Security Sciences (DRDC-CSS).

→ The C-BET methodology and toolkit have been used in a number of evaluations conducted by the CBSA-S&E. The results from the evaluation of iris modality are presented in this paper. More iris performance results are presented in [19]. The results from the evaluation of a voice biometrics are presented in [20].

What is not covered by multi-order score analysis

The multi-order score analysis allows one to evaluate the performance of a system based on the collected biometric data. In real life biometrics deployment or pilot, the collected biometric data unfortunately are not always fully representative of the actual performance of the system. There can be many instances when a system did not capture or store a biometric data or transaction, due to its internal implementation. For example, many systems do not store “bad” (failed to acquire) images or images that did not match anyone in a database, or they store only the best image from a number of captured images. As a result, the evaluator “does not know what he does not know” and so s/he cannot evaluate what the system did not acquire.

For a live pilot-based evaluation of a biometric system it is therefore recommended that additional performance auditing tools are developed to collect the data which otherwise are not captured or stored by the biometric system. For iris or face biometric systems used for access / border control systems, such tools can be developed using Video Analytics techniques, which log the images of all individuals who approach the cam-

era. The development of such Video Analytic tools is one of the current activities of the CBSA-S&E's Video Surveillance and Biometrics Section [25].

ACKNOWLEDGMENT

The valuable feedback from the colleagues from other directorates, in particular, Michael Chumakov, is gratefully acknowledged, as is the help of S&E colleagues, especially, Elan Dubrofsky, and many students, who have contributed to testing different biometric products/modalities using the C-BET and developing the C-BET software.

The Comprehensive Biometric Evaluation Toolkit (C-BET) is developed under the partial funding from the Defence Research and Development Canada's Center for Security Science (DRDC-CSS) and has been refined and tested in the DRDC-CSS sponsored PSTP projects PSTP08-0109BIO ("Stand-off Biometrics Evaluation") and PSTP08-0110BIO ("Biometric Border Security Evaluation Framework").

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

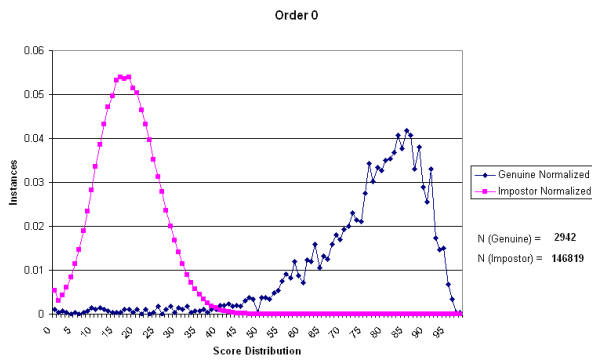
- [1] ISO SC 37 WD 29195, Technical Report on passenger processes for biometric recognition in automated border crossing systems, Last edition: 2010-08-12
- [2] ISO/IEC SC 37 19795-2:2007 Biometric performance testing and reporting - Part 2: Testing methodologies for technology and scenario evaluation. Also, ANSI INCITS 409.3-2005 Biometric Performance Testing and Reporting - Part 3: Scenario Testing and Reporting
- [3] ISO/IEC SC 37 FCD 19795-5, Information Technology - Biometric Performance Testing and Reporting - Part 5: Grading scheme for Access Control Scenario Evaluation
- [4] International Biometric Group. Biometric Performance Certification and test plan - http://www.biometricgroup.com/testing_and_evaluation.html
- [5] Proceedings of NIST International Biometric Performance Conference (IBPC 2010), NIST Gaithersburg, 2010.
- [6] NIST Multiple Biometrics Evaluation (MBE): <http://www.nist.gov/itl/iad/ig/mbe.cfm>. Also NIST IREX evaluation: <http://iris.nist.gov/irex/>
- [7] NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook>
- [8] Anil K. Jain, Ruud Bolle, Sharath Pankanti, Biometrics: personal identification in networked society - Technology & Engineering 1999
- [9] Li, S., (2009). Encyclopedia of Biometrics. Elsevier.
- [10] Mansfield, N., Wayman, J.L. (2002). U.K. biometric working group best practices document. Teddington, UK. National Physical Laboratory.
- [11] Wayman, J.L., Jain, I.K., Maltoni, D., Maio, D. (2005). Biometric Systems: Technology, Design and Performance Evaluation. Springer.
- [12] Atkinson, T.J, Schuckers, M.E, Approximate Confidence Intervals for Estimation of Matching Error Rates of Biometric Identification Devices, <http://myslu.stlawu.edu/~msch/biometrics/projects.htm>
- [13] D.O. Gorodnichy, R. Hoshino. Calibrated confidence scoring for biometric identification. NIST International Biometric Performance Conference (IBPC 2010), March 2-4, 2010
- [14] 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
- [15] 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.
- [16] 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
- [17] D. O. Gorodnichy. Exploring the upper bound performance limit of iris biometrics using score calibration and fusion. In Proc. of IEEE SSCI Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM), Paris, April, 2011.
- [18] Dmitry O. Gorodnichy, Dave Bissessar, Elan Dubrofsky, Jonathon Lee. Analyzing the performance and risks of biometrics systems using Comprehensive Biometrics Evaluation Toolkit (C-BET), Justice Institute of British Columbia and the U.S. DHS's Center of Excellence VACCINE Workshop on "Visual Analytics for Public Safety Professionals", Sept. 20 -21, New Westminster, BC, 2010
- [19] Dmitry O. Gorodnichy*, Elan Dubrofsky, Richard Hoshino, Wael Khreich, Eric Granger, Robert Sabourin. Exploring the upper bound performance limit of iris biometrics using score calibration and fusion. In Proc. of IEEE SSCI Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM), Paris, April 11-15, 2011.
- [20] Dmitry O. Gorodnichy, Michael Thieme, Dave Bissessar, Jessica Chung, Elan Dubrofsky, Jonathon Lee . C-BET evaluation of voice biometrics, SPIE Defence, Security & Sensing Conference, Special Track on Biometric Technology for Human Identification VIII (DS108), Orlando, April 25-29, 2011
- [21] Comprehensive Biometric Evaluation Toolkit (C-BETT) Portal: [https://partners.drdc-rddc.gc.ca/css/Portfolios/Biometrics \(Human ID Systems\)/C-BET](https://partners.drdc-rddc.gc.ca/css/Portfolios/Biometrics%20(Human%20ID%20Systems)/C-BET)
- [22] www.Nexus.gc.ca
- [23] G. Doddington, W. Liggett, I. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance. In Proc. Fifth Int'l Conf. Spoken Language Processing (ICSLP), pages 1351-1354, 1998
- [24] 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/>.
- [25] D. O. Gorodnichy. VAP/VAT: Video Analytics Platform and Testbed for testing. SPIE Conference on Defense, Security, and Sensing, DS226: Visual Analytics for Homeland Defense and Security track. Orlando, 5 - 9 April, 2010.

Biometric Modality: Iris
Biometric systems: #11
Data Source: "G-500" CBSA iris dataset of anonymized iris images of people enrolled to NEXUS program, property of CBSA

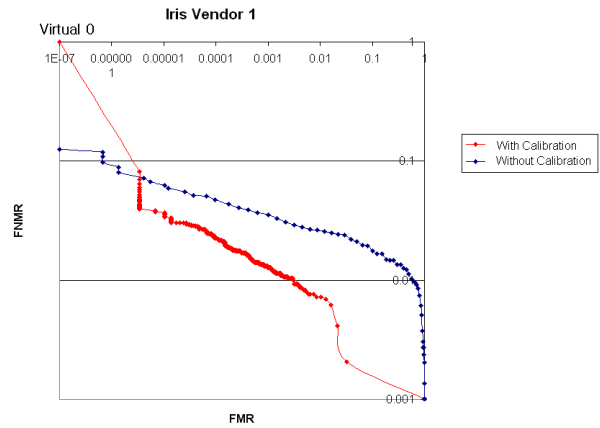
Failure to Acquire (Enrolled & Passage Images): FT Ae = 0.8%, FT Ap=1.1%					
False Match (Accept) Rate: FMR (%)	Failure of Confidence Rate: FCR (%)	False Non-Match (Reject) Rate: FNMR (%)			
		Using system original scores		Using Gorodnichy-Hoshino score calibration	
		Transaction based	Subject based	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

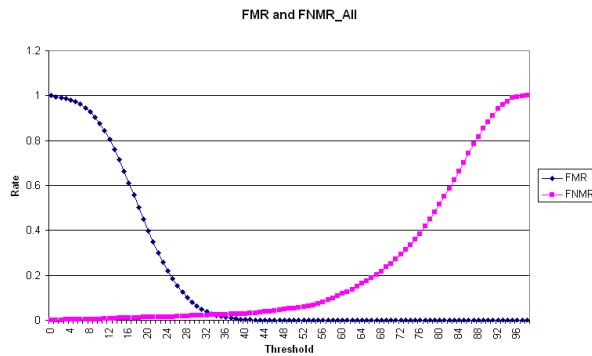
a)



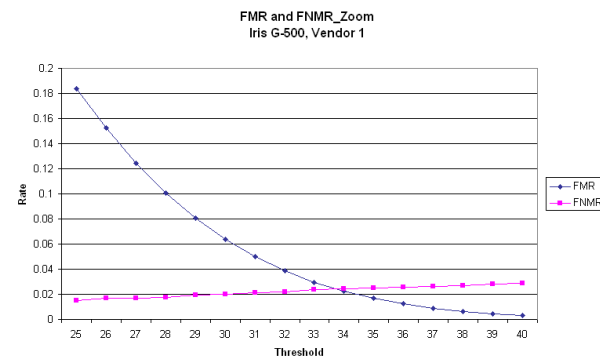
b)



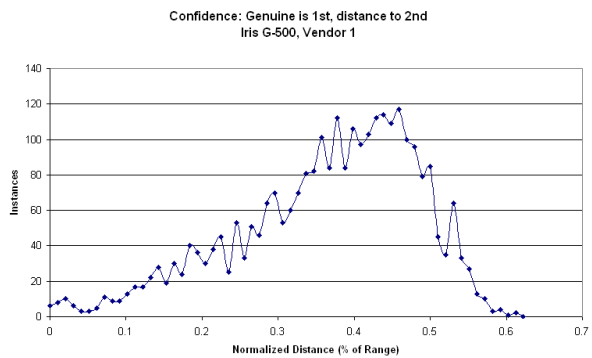
c)



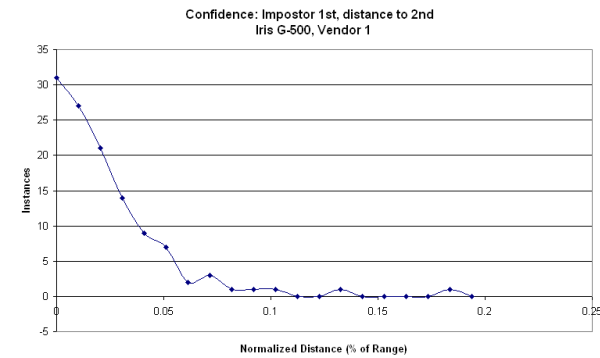
d)



e)



f)



g)

h)

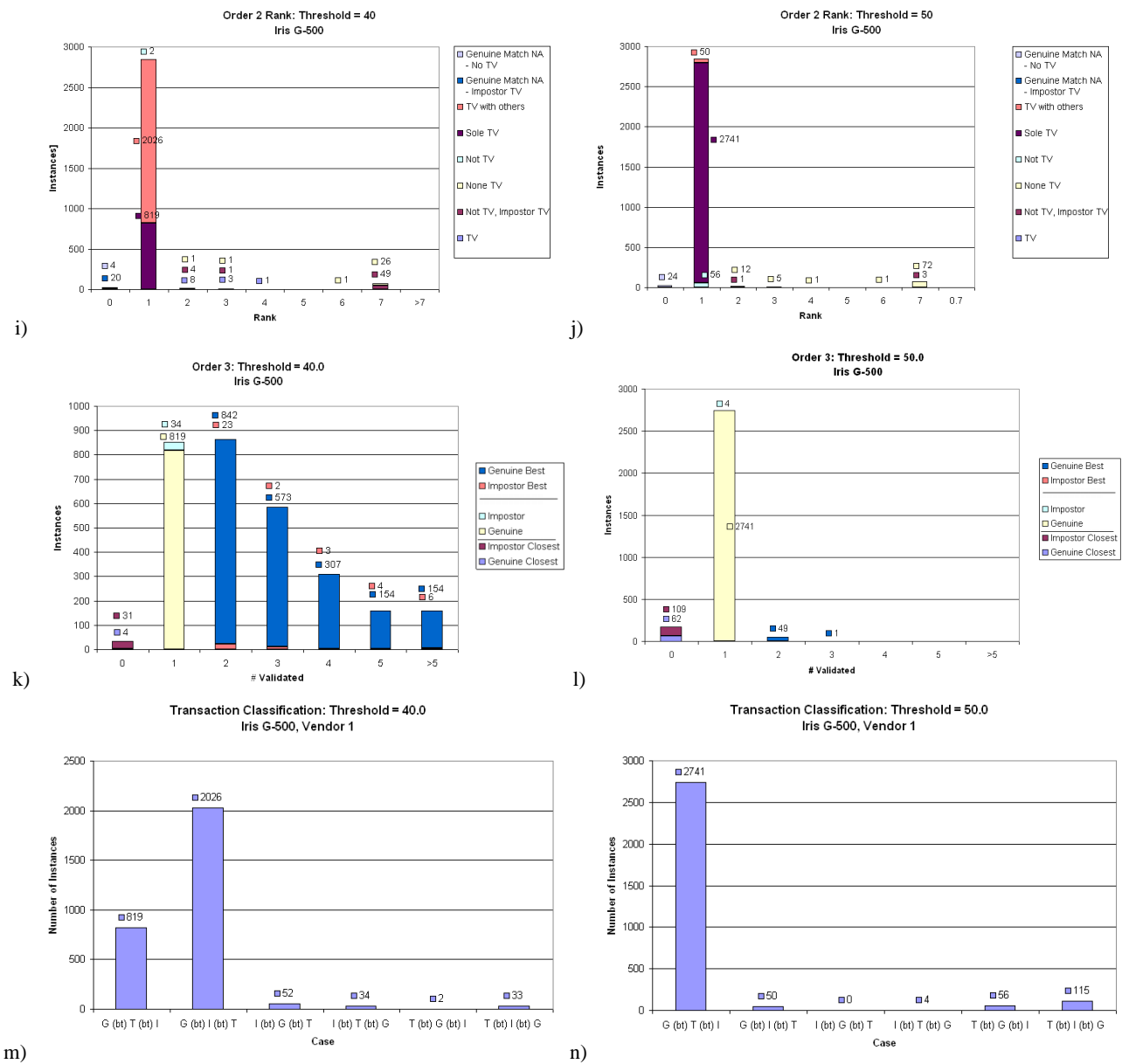


Fig. 1. Two-page summary report of the C-BET evaluation of an iris system (for more details see Section V). Page 1 shows the basic system performance results based on the the Order-0 and Order-1 score analysis: a) Description of the product and dataset used in the evaluation; b) Performance Summary table showing most important evaluation metrics (FTA, FCR, and FNMR at given FMR points – obtained with transaction-based and subject-based analysis, using the original system scores and the calibrated scores); c) Distributions of Genuine and Impostor matching scores obtained by the system (Order-0 analysis); d) DET curves with and without calibration (Order-1 analysis); e) FMR/FNMR distributions (Order-1 analysis); f) FMR/FNMR distributions zoomed on the area of highest importance; and g-h) the Order-3 recognition confidence statistics – for genuine best matches and for impostor best matches, where confidence is measured as the normalized distance from the best score to the second best score. Page 2 shows the threshold-validated Order-2 and Order-3 score analysis – for two different threshold values: i-j) The number of instances when the genuine match was the best, second best etc.; k-l) The statistics on the number of Threshold Validated matches and those of them that ranked the best; l-m) Six-outcomes of the summarized threshold-based analysis.