Proc. IEEE Symposium on Computation Intelligence for Security and Defence Applications (CISDA'09)

Evolution and evaluation of biometric systems

Dmitry O. Gorodnichy

***Abstract*— Biometric systems have evolved significantly over the past years: from single-sample fully-controlled verification matchers to a wide range of multi-sample multi-modal fully-automated person recognition systems working in a diverse range of unconstrained environments and behaviors. The methodology for biometric system evaluation however has remained practically unchanged, still being largely limited to reporting false match and non-match rates only and the trade-off curves based thereon. Such methodology may no longer be sufficient and appropriate for investigating the performance of state-of-the-art systems. This paper addresses this gap by establishing taxonomy of biometric systems and proposing a baseline methodology that can be applied to the majority of contemporary biometric systems to obtain an all-inclusive description of their performance. In doing that, a novel concept of multi-order performance analysis is introduced and the results obtained from a large-scale iris biometric system examination are presented.**

I. Introduction

An organization that deploys or plans to deploy a biometric system needs to know how well the system performs and what factors affect its performance so that proper system selection or setup adjustments can be made. The only way to acquire such knowledge is through *evaluation*, which is *the procedure that involves testing of a system on a database and/or in a specific setup for the purpose of obtaining measurable statistics* that can be used to compare systems or setups to one another.

Biometric systems have evolved significantly over the years and are now applied in a wide variety of applications and scenarios. It is therefore understood that what is good for one application or scenario may not be as good for another, and, as a consequence, the evaluation procedure may have to be different for different applications and scenarios. In this paper, such differences are examined through establishing a taxonomy of biometric systems, including the definition of key concepts related to biometric system performance (Section II) and tracing the evolution of biometric systems (Sections III). The limitations of the conventional biometrics evaluation methodologies are then examined (Sections IV) and a new all-inclusive evaluation framework is proposed (Section V), followed by the presentation of a novel multi-order performance analysis approach, which is the main contribution of the proposed framework (Section VI).

Dmitry O. Gorodnichy is Senior Research Scientist with the Video Surveillance and Biometric Technologies group of the Laboratory and Scientific Services Directorate of the Canada Border Services Agency, 79 Bentley, Ottawa, ON, K2E 6T7, Canada (Tel: (613) 954-3785, e-mail: dmitry.gorodnichy@cbsa-asfc.gc.ca).

II. Terminology and critical concepts

Although there have been many books and recently several standards written defining key biometrics concepts, below we cite and redefine those of them that are most important in the context of the current presentation. Several new definitions are also introduced.

A. Biometrics as Image Recognition

We start from the definition of a biometric system that will help us to define the taxonomy quantifiers for biometric systems and to appreciate the fact, which should be always kept in mind while conducting an evaluation of a biometric system, that biometric solutions are derived from two main research areas: 1) Image Processing (IP), which is a part of computer science that deals with the extraction of numerals from imagery data, and 2) Pattern Recognition (PR), which is a part of statistical machine learning theory that can match numerals to one another.

Definition: ***Biometrics*** *is an automated technique of measuring a physical characteristic (**biometric data**) of a person for the purpose of recognizing him/her.*

The importance is given to the word "*automated*", which implies that all steps involved in the recognition process are done by a computer, and to the fact that the word "*recognition*" is used in general terms here.

This definition also defines two components that make a biometric system: *Capture component*, where "*measuring*" of a trait is done through an image/video/signal capture device, and *Recognition component*, which is a recognition software that performs analysis and matching of measurements.

In order to not confuse biometric raw data, which in the case of image-based biometrics are raw images, with biometric templates derived from the raw images by means of IP techniques and to highlight the two *stages of biometric deployment* we also use the following definitions:

Definition: ***Enrolled data*** *are biometric images that are stored in the system at the **Enrollment stage** for the purpose of being matched upon later. **Passage (or Test) data** are new biometric images that are presented to the system at the **Recognition stage** for the purpose of being recognized. A single piece of data is referred to as a **sample** or image.*

Note that Enrolled data are often of better quality than Passage data, due to the fact that enrollment happens only once and is therefore a well guided and controlled process.

B. Operational Biometric Recognition tasks

From the operational point of view, we can see that there are five operational recognition tasks for which a biometric system can be applied within an organization. These tasks vary significantly in their biometric data acquisition procedures, error costs and error mitigation strategies, as summarized below:

1. ***Verification***, also referred to as *authentication* or *1 to 1 recognition*, as when verifying ATM clients or Restricted Access Area officers using a bank or Access Card.
2. ***Identification***, or *1 to N recognition* (N is often large or can grow), or *positive (or "White list") identification*, as when identifying a pre-registered individual from a watch list, where a test sample is compared against all individuals in a database and the best match (or the best k matches) are selected to identify a person.
3. ***Screening***, or *negative (or "Black list") identification*, which is a special case of *1 to M recognition* (M is normally not large and fixed), as when monitoring traffic of people for the purpose of identifying criminals in it.
4. ***Classification***, or *categorization*, is a special case of *1 to K recognition* (K is small and fixed), where a person is recognized as belonging to a) one of the limited number of classes such as person's gender, race or various medical genetic condition, which can be used as *soft biometrics*, or b) one of limited number of identities as used in automated annotation (tagging) of people in teleconferences or video stream(s).
5. ***Similarity quantifier***, which is a special case of verification used in Forensic document investigation, in which both (or more) images to be compared are presented to a system at the same time and/or in which a biometric system is used to provide the comparative measurements rather than a final recognition decision so that a human analyst will make the final recognition decision himself.

A match obtained in verification and positive identification tasks may be no longer questioned. On the other hand, the match result obtained in Screening or Classification would normally be further processed or investigated and, in many cases, also combined with other recognition data available about the person.

It is also understood that for verification and positive identification tasks a false non-match has much less negative impact/cost ("inconvenience") than a false match ("security breach"), whereas for negative identification tasks this is the opposite.

C. Operational Biometric modality characteristics

Based on the type of the operational recognition task, an organization may impose certain requirements on the biometric modality used by a biometric system, in particular with respect to the following modality characteristics [5]:

1. *Universality*: each person should have the trait.
2. *Uniqueness*: how well biometrics separates individuals from one another.
3. *Permanence*: measures how well a biometric resists aging, fatigue etc.
4. *Performance*: accuracy, speed, and robustness of technology used.
5. *Collectability*: ease of acquisition for measurement.
6. *Acceptability*: degree of public approval of technology.

A trade-off between Performance and Acceptability is normally observed as illustrated in Figure 1 (from [6]). - Well performing biometrics systems use biometric data that are very personal and may therefore be less accepted by the public or harder to collect. Such biometrics may require person's permission and/or cooperation, which is the case with verification or "White list" identification. On the other hand, "Black list" identification will likely rely on biometric data that can be easily collectable from people without their cooperation, but which, as a result, is less discriminating.

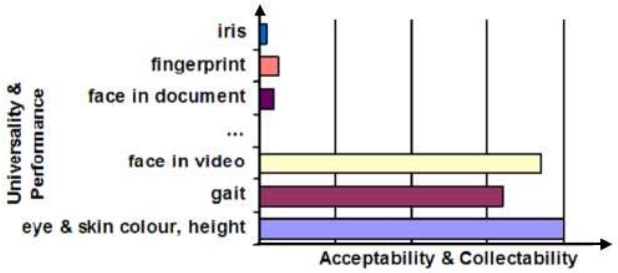


Fig. 1. Performance of different image-based biometric modalities with respect to different operational modality characteristics.

D. Operational conditions

Based on the recognition task and scenario, several operational conditions may need to be imposed and/or expected, such as

1. *Overt* vs. *Covert* image capture
2. *Cooperative* vs. *Non-cooperative* participant
3. *Structured* vs. *Non-structured* (constrained vs. non-constrained) environment – *environment-wise* (eg. lighting condition)
4. *Structured* vs. *Non-structured* (constrained vs. non-constrained) environment – *procedure-wise*.
5. *Size of the database*: Large vs. small
6. *Local* vs. *Centralized* data storage
7. Relative *Impact (Cost)* of False Match vs. those of False Non-Match

Note that Conditions 1-4 for the Enrollment stage may be different from those observed at the Recognition stage.

E. Recognition steps and bottlenecks

In order to understand why biometric recognition may fail and how to conduct the evaluation, one needs to know how biometric recognition works. Figure 2 illustrates the processing steps performed in face recognition (from [6]) and iris recognition systems, which are applicable to most image-based systems. These steps are:

1. *Capture* of image(s)
2. Best image(s) selection and *enhancement* (*preprocessing*)
3. Biometric region extraction (*segmentation*)
4. *Feature detection and selection*: minutia, colour, edges…
5. *Computation of template:* set of *L* numbers (*0<Xi<MAXi, i=1...L*) corresponding to feature attributes (angles, RGB values, wavelet coefficients …)
6. Computation of *match scores* (*similarity distances*): *Sk*
7. *Recognition decision*: based on a statistical rule, the simplest and most commonly used of which is binary comparison to a fixed threshold, optionally followed by its integration / fusion with other data (*post-processing*).

Error in *any* of these steps may drastically affect the final recognition decision of the system. The examples shown in Figure 2, taken from face and iris recognition systems, illustrate these steps and some of problems that could occur. It should be appreciated that solutions to these problems rely on the techniques from both Image Processing (Steps 1-5) and Pattern Recognition (Steps 5-7) research.

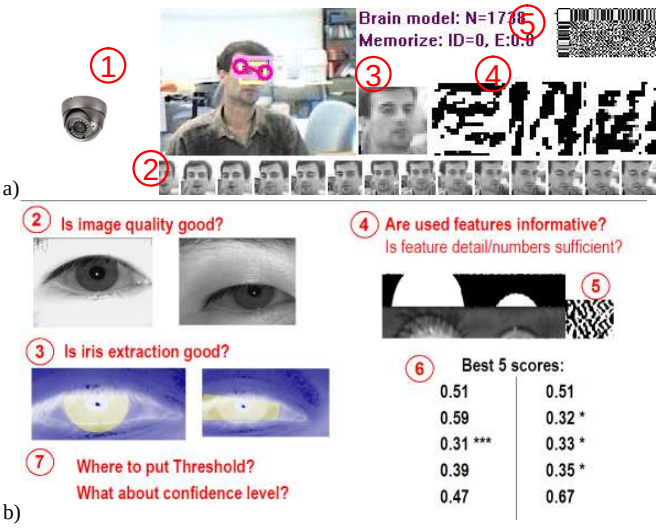


Fig.2. Recognition steps performed in face biometrics (a) and iris biometrics (b), and the associated problems that may occur at each step.

III. Biometrics Evolution

A. Evolution towards surveillance

As one examines the evolution of biometrics, one can see that over the years, as computers become faster and more automated intelligent processing is done, biometric systems are increasingly applied to less intrusive, less constrained, free-flow surveillance-like environments, where biometric data can be acquired at a distance and possibly in inconspicuous (covert) manner. As a result, for such systems to achieve reliable performance, the *recognition results may need to be integrated or fused over time and/or with results obtained from other biometric systems.*

Of a particular interest is the phenomenon of merging Biometrics and Video Surveillance, illustrated in Figure 2, and the arrival of such biometric technologies as *Biometric Surveillance*, *Soft Biometrics* and *Stand-off Biometrics*, also identified as *Biometrics at a Distance*, *Remote Biometrics*, *Biometrics on the Move* or *Biometrics on the Go*, and an increased demand for *Face Recognition from Video*, which is where Biometrics meets Video Surveillance and which is seen as a golden solution to many operational needs.

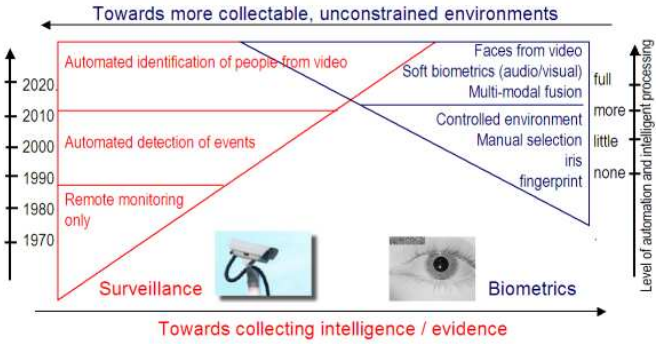


Fig. 3. Evolution of Biometric and Video Surveillance systems: towards each other, with overlap in Face Recognition.

B. Special Interest: Face Recognition

While for humans recognizing a face in a photograph or in video is natural and easy, computerized face recognition is very challenging. In fact, automated recognition of faces is more difficult than recognition of other imagery data such as iris, vein, or fingerprint images due to the fact that the human face is a *non-rigid* 3D object which can be *observed at different angles* and which may also be *partially occluded*. It is important therefore for an organization interested in using face recognition systems to know what is possible and what is not in the area of automated facial recognition as well as to know how to evaluate such systems.

Based on prior work [6-8], we summarize in Table I the readiness level of those face recognition technologies that are most the closest for deployment, and also highlight the fact that we are far away from the general face recognition as performed by humans.

TABLE I
READINESS LEVEL OF FACE RECOGNITION TECHNOLOGIES
5 - ready for deployment, 4 -needs minor R&D, 3 – needs some R&D, …
0 - not ready at all

RL=5	Human-assisted Recognition From Video (not biometrics per se), where face is automatically extracted from video, e.g. to be linked with boarding pass or vehicle plate number or matched with passport photo.
RL=4	Face image and geometry automatically extracted from video is used together with other modality (eg. Iris) recognition.
RL=3	Automated Recognition from ICAO-conformed passport photographs - as good as finger or iris recognition.
RL=3	Automated Recognition From Video only – is possible, if procedural constraints are imposed (to make video snapshot image quality closer to that of passport image).
RL=3	Identification in small-size database, as in monitoring access-restricted areas applications.
RL=0.1	General unconstrained automated face recognition.

In order to know how to conduct evaluation of Face Recognition systems, one needs to know what makes such stand-off biometrics so different from other biometrics.

C. Special Interest: Stand-off biometrics

As opposed to other biometrics, in which a person intentionally comes in contact with a biometric sensor, *stand-off biometrics* is applied to a person without his/her direct engagement with the sensor. In many cases, a person would not even know where a capture device is located or whether his/her biometric trait is being captured. As a result, a single biometric measurement or output of a stand-off biometrics system is normally much less identifying than that of other biometrics system. This means two things.

First, it is common for a stand-off biometric system to have more than one match below the matching threshold, or to have two or more matches having very close matching scores.

Second, the final recognition decision of a stand-off biometric system is not based on a single measurement or output, but rather on a number of biometric measurements taken from the same or different sensor, combined together using some data fusion technique.

This leads us to reconsidering the way the performance evaluation of biometric systems is done.

IV. Biometrics Evaluation

It is well accepted nowadays that biometrics, especially image-based, will never produce error-free recognition results. However and most importantly, it is also appreciated now that, *with proper system tuning and setup adjustment, critical errors of the biometric systems can be minimized to the level allowed for the operational use.*

The insights on system tuning and setup adjustment, as well as on the selection of the system and risk mitigation procedures that best suit the operational needs, can only be obtained through system performance evaluation. However, the performance evaluation protocols and metric should be appropriate for the task and scenario to which the systems are applied.

A. From Door opening to Intelligence gathering

Fostered by end-users' perception of access control biometric systems and by the way biometric systems are marketed by industry, there has been a widespread stereotype created about biometric systems that they are tools to open a "door" - either a physical door (to enter a plane or restricted access area) or a virtual "door" (as in a laptop or a cell phone). This stereotype creates a simplistic understanding of how biometric system results are obtained, used and judged upon. In particular (see Table II), it could be seen that it creates parallels between two very different technologies: an intelligence gathering device, which a Biometric System is, and a Proximity Sensor that is used to open a door (or valve) in a presence of person.

The most striking similarity between the two technologies is seen in the way both technologies are evaluated. Indeed, as one examines current biometric evaluation standards [1,2] and evaluation reports of various biometric technologies [3,4], one can find that the way biometric recognition performance is evaluated and reported is still primarily based on counting the number of times a "door" has opened correctly and incorrectly, i.e. using the False Match and False Non-Match Rates (FMR and FNMR) and the trade-off curves built thereon.

In the light of the biometrics evolution and its current applications, which is highlighted in the previous sections, such evaluation framework may no longer be found sufficient and/or appropriate. Instead, a new evaluation framework needs to be developed that allows one to obtain the *all-inclusive* description of the performance of a biometric system based on its place in biometric taxonomy and *all* data measured during the run of the system.

TABLE II
BIOMETRIC SYSTEMS VS. PROXIMITY SENSORS

	Biometric systems (for access control)	**Proximity sensors**
Application Task	Open the "door" for ***the*** person	Open the "door" for ***a*** person
Measurements taken	Similarity distance (match score): S	Distance to a person: D
Tasks achieved	when S < T	when D < T
Calibration done by	computing similarity distances of genuine and imposter data	measuring distances at different ranges
Performance metric	FMR, FNMR (ROC / DET curves)	FMR, FNMR (ROC / DET curves)

B. Conventional performance evaluation metrics

According to conventional methodology, the following two binary errors that a system can exhibit are counted:

- *False Accept* (FA) also known as *False Match* (FM), false hit, *false positive* or *Type I error*; and
- *False Reject* (FR) also known as *False Non-Match* (FNM), false miss, *false negative* or *Type 2 error*.

By applying a biometric system on a significantly large data set, the total number of FA and FR is counted to compute the cumulative measurements:

- *False Accept Rate* (FAR)
- *False Reject Rate* (FRR) or *True Acceptance Rate* (TAR = 1 - FRR), also known as *Hit Rate*,

at fixed rates of another or as functions of *match threshold*.

The *trade-off curves*, also called *Figures of Merit*, are also computed such as:

- *Detection Error Trade-off* (DET) curve, which is the graph of FAR vs. FRR, obtained by varying the system *match threshold*, or
- *Receiver Operator Characteristic* (ROC) curve, which is similar to DET curve, but plots TAR against FAR.

It is important to note that when counting the number of matches and non-matches, *verification match and identification match are defined differently***.** In verification, an image is matched if its matching score is less (or larger) than a threshold, whereas in identification an image is matched if its score is the smallest (or largest).

Two additional metrics/curves have been specifically proposed for Identification systems to address the issue:

- *Rank-k identification rate* (*Rk*) - the number of times the correct identity is in the top k most likely candidates.
- *Cumulative Match Characteristic (CMC) curve*, which

plots the rank-k identification rate against *k*.

These rates/curves still do not offer a complete picture about the system performance, as they do not provide any metric to estimate the confidence of the system in its recognition decision (Step 7 in Figure 2, Section II.E). Nor can they be used to distinguish False Reject Rate from true "not-in-the-list" detection rate, if applied to an *open dataset.*

Additionally, besides recognition measurements, other system usability factors also have to be evaluated, in order to see if the conditions/requirements imposed on the systems operation (Section II.C) are met and to insure that it can be further customized and upgraded.

We therefore propose a new all-inclusive evaluation methodology that would allow one to investigate most of the issues related to the performance of a state-of-the-art system.

V. TOWARDS ALL- INCLUSIVE EVALUATION

A. Hierarchy for generic biometrics evaluation

Table III shows the hierarchy of steps for a general all-inclusive evaluation of a biometric system, which takes into account modality suitability, cost, factors and the performance criteria. Normally, the suitability of the modality should be evaluated first and prior to making the decision on a particular biometric solution or product.

TABLE III
ALL-INCLUSIVE BIOMETRICS EVALUATION

1. Determine suitability of modality (-ies)
2. Determine costs/impact of FM and FNM
3. **Determine all factors affecting performance**
4. **Evaluate performance of market solutions *:**
   - I. wrt all factors that affect the performance
     - a. On large-scale database (>1000)
     - b. On Pilot project (in real environment)
   - II. wrt capability to be integrated / customized
     - **c.** Wrt input parameters (pre-processing)
     - **d.** Wrt output parameters (post-processing)

B. Factor-driven datasets

There are several datasets publicly available for many image-based biometrics. Such datasets would be of great value for any biometric system. It is recommended however that data presented in those datasets be first analyzed for the variability of factors in them that may affect the recognition performance. In many cases such factors are listed along with dataset description, as it is for face databases. In particular, a summary of facial dataset sorted out according to the factors that affect face recognition performance is prepared in [10].

If the information of dataset images factors is not available, such information can be obtained through preprocessing of images with image quality analysis tools, which are often supplied with biometric systems.

C. Matching vs. Capture evaluation

A single provider may not be the best in the market in both the capture component of the biometric system and in the matching component of it. It is therefore recommended that evaluation be done independently for the capture components of the biometric system and the matching components, and that an organization imposes an open-architecture constraint on systems to be deployed - in order to insure that they provide access to as many parameters as possible and allow their integration with other sensors or system components.

D. Evaluation criteria types (for Matching and Capture)

Evaluation criteria for Matching components are divided into three types:

- Type M0: General questions. These questions, usually graded **Y**es/**N**o or **U**nsure, relate to the abilities and functionality of the program, rather than to evaluating its recognition performance.
- Type M1: Recognition performance tested on large-scale production factor-agnostic dataset(s).
- Type M2: Recognition performance tested on factor-specific dataset(s).

Evaluation criteria for Capture components are divided into two types:

- Type C0: General questions, related to functionality, convenience and ease of use of the Capture module, and
- Type C1: Capture performance tested on factor-specific dataset(s).

Example of C1 criteria questions that identify factors that affect iris recognition performance is given in Table IV.

TABLE IV
CAPTURE CRITERIA C1: ROBUSTNESS TO FACTORS (FOR IRIS RECOGNITION)

ID #	Performance with respect to the following factors:
C1.1a	Orientation – Iris
C1.1b	Orientation – Camera
C1.2a	Iris resolution – in pixels
C1.2b	Iris resolution – distance to camera
C1.3	Occlusion
C1.4	Image quality: focus, motion blur
C1.5a	Illumination: Light source location (Front, back, side)
C1.5b	Illumination: specular reflection (from LED or Lamps)
C1.5c	Illumination: brightness / contrast

E. Data preparation, collection and analysis

The core of any matching evaluation is obtaining and analyzing the recognition matching scores produced by the system. For comprehensive performance evaluation, the procedure described in Table V is proposed. This procedure employs a novel multi-order performance analysis approach, which is described in more detail in the next section.

The procedure commences from a small-size dataset with a goal of obtaining a “bird’s-eye view” of the system’s functionality and to obtain the estimates of the speed and level of programming effort that is required for each of the steps defined in the protocol.

The most time consuming step in the procedure is the computation of all-to-all match scores (Step 2). If for a given dataset size (N) a system permits computing such scores within a reasonable amount of time, then the multi-order analysis of the system performance for this size is performed. For a reference, Table VI shows the estimate time needed to perform Encoding and Matching steps for different dataset sizes, based on testing several iris biometric systems.

TABLE V

PROTOCOL FOR COMPREHENSIVE PERFORMANCE EVALUATION OF A BIOMETRIC SYSTEM

Step 0. Data preparation

Select Enrolled and Passage (possibly of lower quality) datasets:

- of several sizes (N), eg. 100, 500, 1000, 5000
- with K passage images per each enrolled image,
- (if possible) corresponding to different factors that affect the performance

Apply one set at a time for each system (or parameter, or factor), starting from a smaller set, and measure the time needed for each of the following steps. Don't proceed to a larger set, if the estimated time is over the limit.

Step 1. Encoding (of all images in a Enrolled and Passage sets)

Measure:

- Failure to Acquire for Enrolled images (FTA.E)
- Failure to Acquire for Passage images (FTA.P)
- Image quality numbers

Step 2. Matching (Obtaining scores for ALL available data):

i) using default settings/threshold,
ii) using other possible settings/thresholds

Step 2a. Get match scores for Enrolled set - Imposter tests only

- Measure: FAR = #FalseAccepts/(N-FTA.E)

Step 2b.1. Get match scores for Passage set – Genuine tests only

- Measure: FRR = #FalseRejects/(N-FTA.P)

Step 2b.2. Get match scores for Passage set – Imposter tests only

- Measure: FAR = #FalseAccepts/(N-FTA.P)

Step 3. Multi-order analysis (of ALL obtained scores)

Step 3.a. Order-0 (no Analysis, Visualization only):

- Plot Probability Distribution Functions PDF(S) of genuine and imposter scores (at different increments to highlight trade-off zone)

Step 3.b. Order-1 (conventional) analysis:

- Compute/Plot verification rates and curves, where match is defined when a score is below a threshold:
  - FMR, FNMR, DET

Step 3.c. Order-2 analysis:

- Compute/Plot Rank-1 identification rates, where match is defined when it is a Minimal score:
  - FMR, FNMR, DET
  - distribution of best scores values (optional)

Step 3.d. Order 3 analysis:

- Compute /Plot Rank-k (k=2,3,4,>5) identification rates and distribution of Confidences, defined as below:
  1: PDF(S2-S1) of second best score minus best score
  2: PDF (N(S<T)) of number of scores less than a threshold
  3: PDF(Rk) of identification rank

(Steps 3.c and 3.d can be performed in a single procedure).

Trade-off curves obtained on sets of different sizes are plotted on the same graph to highlight the tolerance to scalability, with all output dots visible.

TABLE VI

TIME REQUIRED TO ENCODE AND MATCH DATASETS OF DIFFERENT SIZES

N	100	500	1000	5000	10K	20K	50K
Step 1	5'	30'	1h	6h	12h	1d	3d
Step2a	.5'-20'	1'-6h	5'-1d	2h-1w	4h-1m	8h/4m	3d-1y+
Step2b	10'-3h	30'-3d	1h-2w	8h-50w	17h-4y	1.5d-16y	5d-100y

VI. MULTI-ORDER PERFORMANCE ANALYSIS

The multi-order terminology for the proposed innovative analysis comes from the analogy with multi-order statistics terminology, 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.

Similarly, the multi-order biometric performance analysis framework is defined as an approach that examines the evaluation of the system at several levels (or orders) of detail[1]. This framework defines the conventionally used performance metrics, such as summarized in Section IV, as the Order-1 analysis and introduces the concepts of Order-2 and Order-3 analysis defined as follows.

Definition: *The **Order-1 analysis** of the biometric system performance is the analysis that is based on a single number output (score) of the system, as when computing verification match/non-match rates and the error trade-off curves based on a binary comparison of a single 1-to-1 match score to a threshold.*

Definition: *The **Order-2 analysis** of performance is the analysis that is based on all scores that can be obtained by the system for a sample, as when finding the best match score in 1-to-N identification.*

Definition: *The **Order-3 analysis** of the biometric system performance is based on the relationship between the match scores obtained by the system for a sample, as when finding the difference between the best and second-best match scores or all scores that are lower than a threshold.*

Additionally, all statistics and graphical visualization related to score distributions obtained by the system is referred to as the ***Order-0 analysis***. Such analysis does not produce a metric that can be used to quantify the quality of the system performance. Nevertheless, as demonstrated in Figure 4, it provides very important insights on how a system performs and where the performance bottlenecks could be.

The results obtained from the Order-1 analysis are shown in Figure 5. These are the results that would normally be found in evaluation reports published to date or that would be obtained for a product with existing evaluation standards. It should emphasized that when plotting the Order-1 tradeoff curves, it is important that points that are used to extrapolate the curves be shown too. The reason is that a system may never attain certain low levels of FMR or FNMR that are shown on the curve. This is why it is also very useful to report the FMR and FNMR curves (as functions of threshold) in addition to the DET or ROC curves.

[1] Strictly speaking, to follow the analogy with statistics, we should have called the conventional single-number-based evaluation as the Order-0 analysis, with Order-1 and Order-2 analysis corresponding to Order-1 and Order-2 statistics. The shift in numbering is due to the introduction of the Order-0 analysis, which strictly speaking is not an analysis but a visualization of the inner properties of a biometric system.

By highlighting the area of error trade-off and plotting the curves obtained for different dataset sizes on one graph, one can investigate the issues related to the scalability of the system such as an increased number of false rejects and/or the necessity to modify the match threshold (see Figure 5.b).

Very useful and informative curves of Order-1 could be, they still do not provide a complete answer on what system is the best. In particular, *a system that has a higher FNMR (for a fixed FMR) can still be preferable to a system that has a lower FNMR, if it has better mechanisms to report and deal with non-confident recognition decisions.*

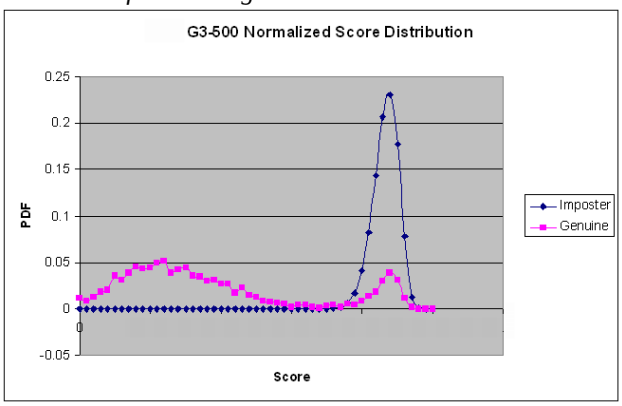


Fig.4. Order-0 analysis visualizes Probability Distribution Functions for genuine and imposter scores and allows one to spot some problems with the data or the matching algorithm.

As for Order-2 analysis, it is by definition routinely performed for identification applications, which require examination of scores for everyone in a database. It is however rarely performed for verification applications, where we believe it could be found very useful too, for example to insure that 1-to-1 match is indeed the best and only match in the entire dataset.

A. Order-3 analysis and Recognition Decision Confidence

The limitation of the Order-1 analysis and the need for higher order analysis is best demonstrated by Figure 2.b (Step 6). The figure shows the best five matching scores obtained for two test images presented to a biometric system for the purpose of identification. As seen, the scores obtained for the test image in the left column provide a very confident winner - the minimal score, whereas the scores obtained for the test image in the right column are much less identifying, as there are several scores that are close to the minimum. Additionally, depending on where the match threshold is, there could be more than one scores below the threshold.

More comprehensive statistics on this phenomenon is shown in Figure 6, which shows the Order-3 performance analysis results obtained from several state-of-the-art systems.

The experiments were run following the evaluation protocol described in Section V (Table V), with datasets containing iris images from 100, 500, 1000, and 4000 individuals, each individual having one enrolled image and six passage images of the same (right) eye. The results reported in Figure 6 are from the 1000-identities passage dataset: ie. containing 6000 iris images.

- Figure 6.a shows the number of instances when there were 0 (ie false rejects), 1, 2, 3, and so on scores below a default threshold.
- Figure 6.b shows how close the second best score was to the best score, by showing the number of instances when the second best score was within 0.01, 0.02, and so on distance from the best score.
- Figure 6.c shows how many times the genuine person scored the best (Rank-1), second best (Rank-2), third best (Rank-3) and so on, of which the portion of scores that were above the default threshold is marked in dark red.

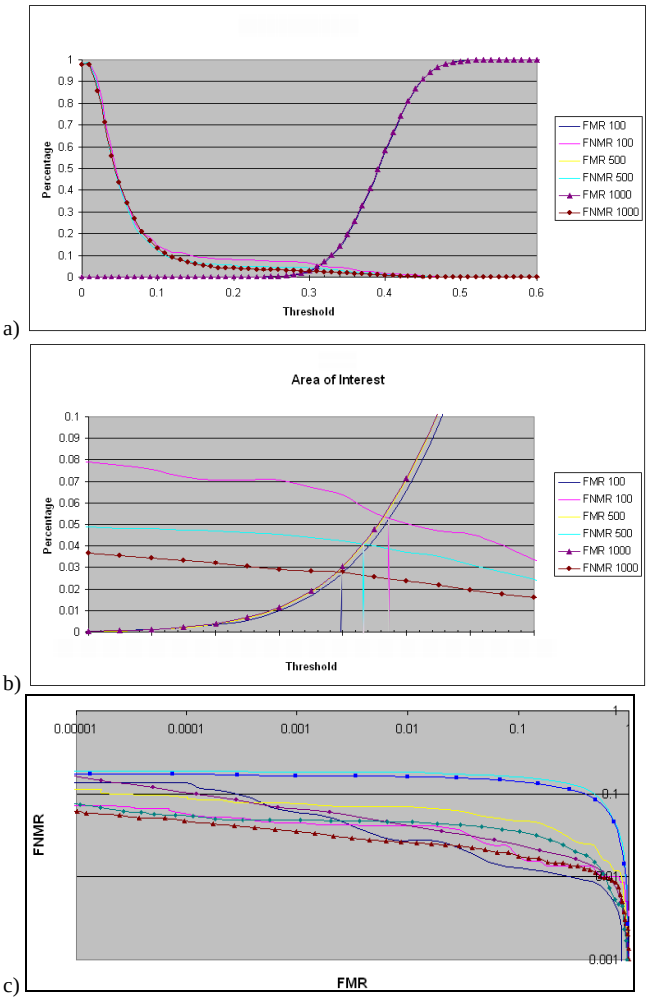
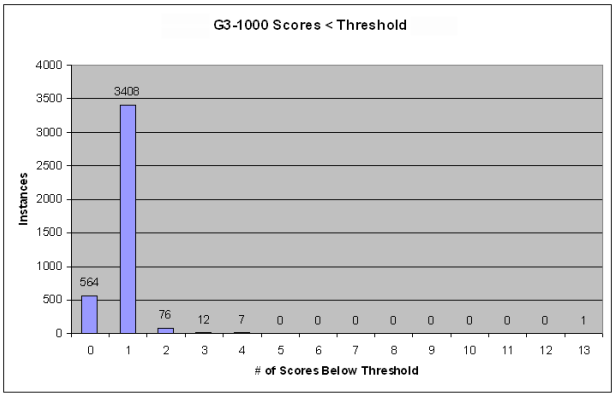


Fig.5. Order-1 analysis is the current performance evaluation standard and is based on computing verification-based (1-to-1) False Match and False Non-Match Rates and the associated error trade-off curves (c).

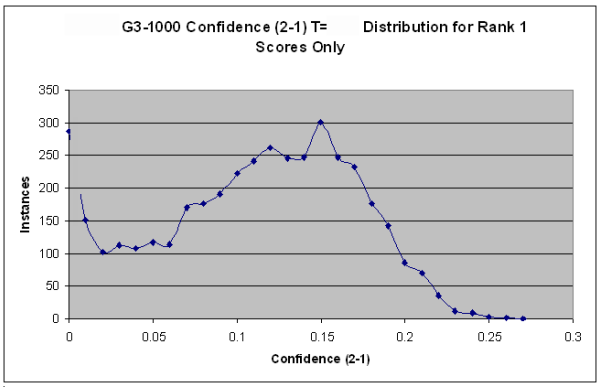
As can be seen, the information obtained with Order-3 analysis provides a sense of the reliability of the biometric recognition results for both verification and recognition, and can therefore be used as biometric recognition confidence metrics.

What is also important to indicate is that, as the presented results show, there are many instances when there is more than one match below the matching threshold, or when there are two or more matches having very close matching scores. With traditional status-quo evaluation methodologies this important information is lost. However, with the proposed multi-order methodology this information is not lost and can be used to fine-tune the system, as well as to develop the procedures to mitigate the risks associated with having non-confident recognition results.

a)



b)



c)

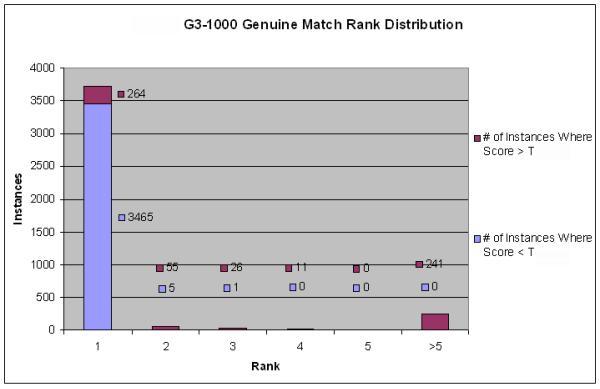


Fig.6. Order-3 analysis involves computing the rates of recognition confidences, computed as: a) the number of matches below a threshold, b) distance from best score to second best score, and c) recognition rank itself.

VII. Conclusions

Performance evaluation plays a critical role in biometric system deployment, due to the fact that biometric systems can produce errors. In-house technical evaluation allows one to insure that the quality of the software delivered by the vendor meets the operational requirements. It also allows one to build an operational and efficient system tailored to a specific need, by ensuring that a biometric system provides access to as many parameters of the system as possible and allows its integration with other sensors or system components.

It takes a good understanding of all technical problems and stages underlying the biometric process to conduct a comprehensive evaluation. All factors and system taxonomy differentiators have to be taken into account when evaluating a biometric system. The recognition performance needs to be understood, and *all performance changes* that are due to a change of a system or system parameters and not only the match/non-match errors have to be analyzed.

Even though no biometric modality, except DNA, is error-free, critical errors can be minimized to the level allowed for the operational use - with a proper performance evaluation and optimization strategy. Despite the fact that performance may also deteriorate over time, as the number of stored people increases and spoofing techniques become more sophisticated, there are also many ways to improve biometric system performance - by using more samples, modalities, and adding additional environmental and/or procedural constraints. For an organization that intensively relies on biometric technology for its day-to-day activities, it is therefore recommended that continuous performance monitoring, tuning and upgrading of its biometric systems be carried out, accompanied with a regular all-inclusive system performance evaluation. To conduct such an evaluation, the biometrics taxonomy accompanied by the multi-order performance analysis framework proposed in this paper can be used.

Disclaimer: The data and results presented in this paper are not associated with any production system or vendor product. They are obtained from lab environment experiments performed on a variety of state-of-the-art iris and face recognition biometric systems using real anonymized biometric data. They are chosen to be representative of many cases observed throughout the experiments and are used here solely for the purpose of illustrating the concepts presented in the paper.

Acknowledgement

This work has been done in part for a CBSA Iris Biometric Technology Examination, and in part for the PSTP projects on Stand-off Biometrics Evaluation (PSTP08-0109BIO) and Biometric Border Security Evaluation Framework (PSTP08-0110BIO). The author gratefully acknowledges the work of many CBSA colleagues who prepared the iris data and conducted the experiments reported in this paper.

References

[1] ISO/IEC 19795-1:2005 Biometric performance testing and reporting. Part 1: Principles and framework.

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

[3] In Face Recognition Vendor Test website, *http://www.frvt.org*.

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

[5] A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics, 14(1):4–20, January 2004.

[6] D. O. Gorodnichy. Video-based framework for face recognition in video. In Second Intern. Workshop on Face Processing in Video (FPiV'05), Proc. of Second Canadian Conference on Computer and Robot Vision (CRV'05), pp. 330-338, Victoria, BC, online http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48216.pdf, 2005.

[7] D. O. Gorodnichy. Seeing faces in video by computers (Editorial). Image and Video Computing, Special Issue on Face Processing in Video Sequences. (online at http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48295.pdf), 24(6):1–6, 2006.

[8] D. O. Gorodnichy. "Face databases and evaluation" chapter in Encyclopedia of Biometrics (Editor: Stan Li), 2009, Elsevier Publisher (on-line at http://www.videorecognition.com/doc)