



Canada Border
Services Agency

Agence des services
frontaliers du Canada

Multi-order biometric performance analysis

C-BET: Comprehensive Biometric Evaluation Toolkit

SPIE Conference on Defense, Security, and Sensing.
DS108: Biometric Technology for Human Identification track,
Orlando, 5 - 9 April 2010

Dr. Dmitry O. Gorodnichy
Video Surveillance & Biometrics Section
Science and Engineering Directorate

Canada

Outline



1. Who we are (CBSA-S&E VSB) and What we do
 - Why we have to do it (Biometric Evaluation) ?
 - From applications to the needs

2. Conducting Comprehensive Biometrics Performance Evaluation.
 - How we do it (Biometric Evaluation) ?
 - Multi-order analysis & C-BET

3. Next steps...



NRC-IIT Video Recognition research aims to advance the newly-established science of Video Recognition (through tutorials and international workshops) and address the needs of Canadian companies that deal with video data. The team conducts research in all of the aforementioned research areas and develops generic and custom-tailored computer vision systems that perform video recognition tasks. The team develops Video Recognition Systems called Perceptual Vision Systems (in order to differentiate them from ordinary Computer Vision Systems), along three application directions:

- Security, Surveillance and Monitoring
- Visually-enabled computer-human interaction
- Intelligent video communication and processing

For additional information, please contact:

Dr. Dmitry Gorodnichy

Phone | 613 998-5298

Email | Dmitry.Gorodnichy@nrc-cnrc.gc.ca

NRC Institute for Information Technology

iit-iti.nrc-cnrc.gc.ca

"It [an application developed by D. Gorodnichy] is a convincing demonstration of the potential uses of cameras as natural interfaces."

The Industrial Physicist ("Recent advances in computer vision"), February 2003.

"Using a computer will soon be a lot easier for disabled people, thanks to a hands-free device created by Canadian researchers." CNN, September 2004.

"Dr. Gorodnichy's work on visual recognition of body motion goes back to his days working on upgrading the robotic lifting arm used in the space shuttle." New York Times, October 2004.



Who we are

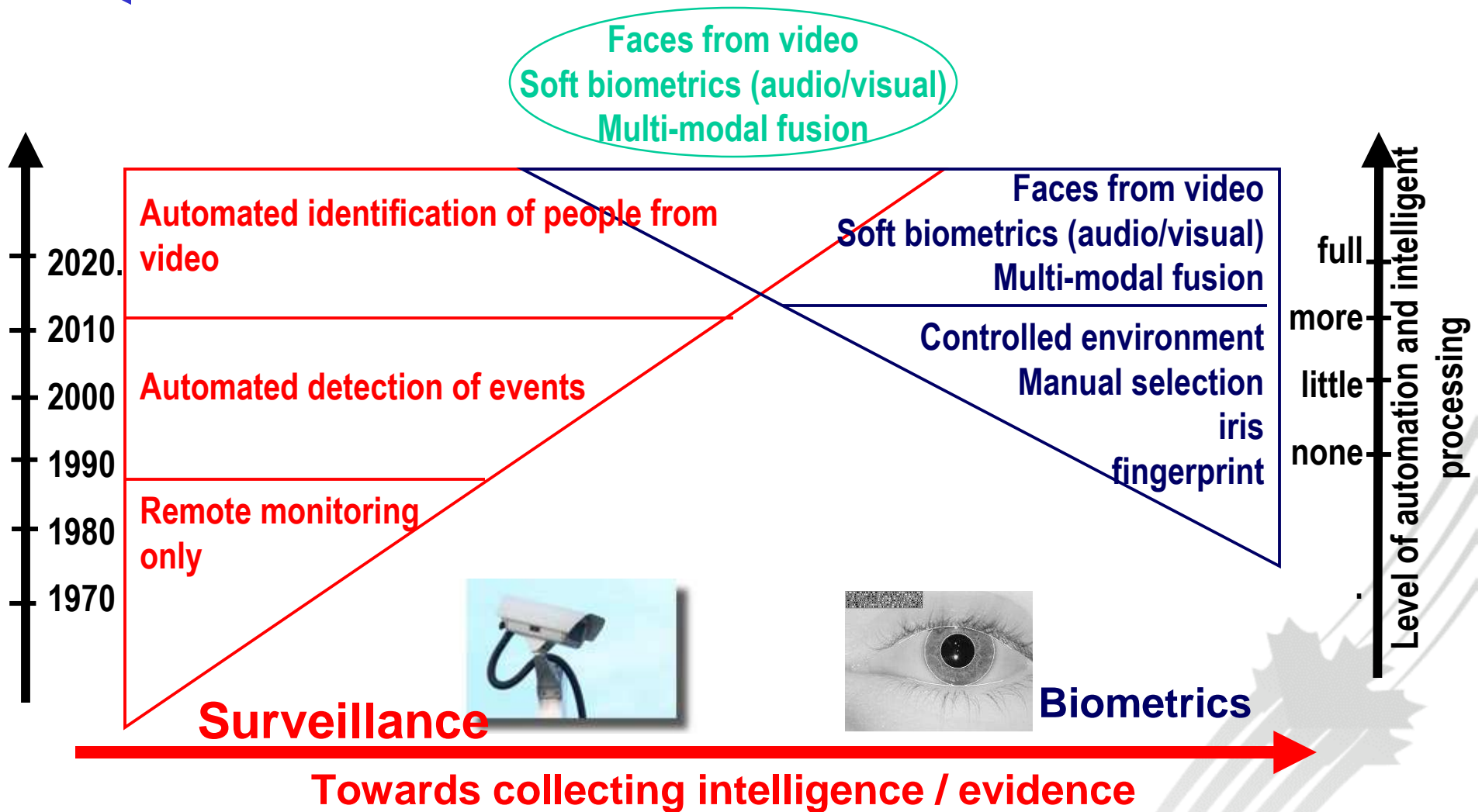


Video Surveillance and Biometrics (VSB) Section

- Based on NRC Video Recognition Systems expertise (2000-2008)
 - www.videorecognition.com
- Following “Border Science: 5-Year Vision/Strategy” (2008)
 - “Make decisions based on sound evidence”
- Created in CBSA-S&E Directorate (January 2009)
 - R&D capacity is achieved - by consolidating expertise in **Image Analysis & Pattern Recognition**
- To support agency’s Portfolios in **Video Surveillance** & **Biometrics**
- To become the prime R&D center for GoC in the areas of **Video Analytics** and **Biometrics**
 - In cooperation with DRDC-CSS: (Defence R&D Canada, Center for Security Science)

Evolution of Video Surveillance & Biometrics

← Towards more collectable, unconstrained environments



Three foci of our R&D work:



Our objective: **To find what is possible and the best**

- **in Video Analytics, Biometrics, Face Recognition**
- **for LAND and AIRPORT Points of Entry (POE)**

to be in a position to build solutions to CBSA & OGD.

- Focus 1: Evaluation of Market Solutions
- Focus 2: In-house R&D
- Focus 3: Live Tests/Pilots in the Field

See also:

- “VAP / VAT: Video Analytics Platform and Testbed for testing and deploying Video Analytics” - in Proc. SPIE “Defense, Security, and Sensing” Conference (Track on Visual Analytics for Homeland Defense and Security) 5 - 9 April 2010, Orlando

NEXUS Iris Recognition



Cross often? Make it simple, use NEXUS.

NEXUS is designed to expedite the border clearance process for low-risk, pre-approved travellers into Canada and the United States.

At LAND Point of Entry (POE)



- “Watch List” / PDP (previously deported persons)



- Also: Voice biometrics - possible



At AIR Point of Entry (POE)



- “Watch List” / PDP (previously deported persons)



- 1-to-1 verification for e-Passport (will be needed soon)

What we have done:

- Iris Biometrics Large-Scale Comprehensive Examination (RFI)
 - Identified Evaluation Standards Gap, Recommendations to ISO-SC 37
 - Proposed Multi-order score analysis
 - D. Gorodnichy. **Evolution and Evaluation of Biometric Systems**, Proceedings of Second IEEE Symposium on Computational Intelligence for Security and Defense Applications. Ottawa, Canada, 9-10 July 2009

Supported by DRDC-CSS:

- **C-BET (Comprehensive Biometrics Evaluation Toolkit):**
 - developed by CBSA S&E Directorate
 - for Community of Practice (CoP) in Biometrics in the Gov't of Canada
 - for selecting new and tuning existing biometric systems
- PSTP Study PSTP08-0110BIO: “Biometric Border Security”.
Lead: CBSA-S&E, Contractor: IBG. Delivery date: 31 March 2010
- PSTP Study PSTP08-0109BIO: “Stand-off Biometrics Evaluation”.
Co-lead with RCMP



Biometric Border Security Evaluation Framework

PSTP 08-0110BIOMT

Public Security Technical Program



PSTP Mission Area

Surveillance, Intelligence and Interdiction, Biometrics

Partners

Lead Federal Department: Canada Border Services Agency (CBSA)

Additional Partners

Royal Canadian Mounted Police, Department of Foreign Affairs and International Trade, Defence Research and Development Canada (DRDC) - Toronto, Office of the Information and Privacy Commissioner of Ontario, University of Toronto, IBG-Canada

Objective

To evaluate biometrics used to identify and verify persons of interest seeking entrance to Canada through various border environments, while allowing the efficient and seamless passage of people and goods across borders, consistent with the Government of Canada's dual prosperity and security mandates.

Results

- The Study established a Strategic Plan and a Roadmap for the use of biometrics in border

- The Study included a detailed performance evaluation of commercial face recognition systems, measuring identification rates for video images under various controlled and uncontrolled "watchlist" conditions
- Results from the face recognition evaluation will facilitate decisions on development, deployment, and optimization of current and future surveillance systems.
- The Study also evaluated the state of the art of speaker identification technology.
- The Study includes an assessment of legal, ethical, cultural, and privacy aspects of border security applications, focusing on risks associated with biometric databases.



Contacts

Portfolio Manager: Pierre Meunier
Surveillance, Intelligence and Interdiction
DRDC Centre for Security Science
613-943-2499 / pierre.meunier@drdc-rddc.gc.ca



Iris Biometrics examination



- Given > ¼ million of enrollees iris images
- Each having 1-100 passage images
- Analyzed by image quality and match score
- Representative sample datasets created: 100, 500, 1000, 4000
 - Each person: 1 enrolled image + 6 passage images

- Several IRIS matcher products examined
- Over 50.000.000 comparisons done / score obtained
 - Over 6 months with 4 full-time employees

- Task: to get to know the State-of-Art: what's possible & gaps
- ... and in doing so, to better understand our own data/system
 - Risks? Factors? Risk minimizing strategies / recommendations

Why to conduct evaluation ?



Because ...

- **Biometric system is not a “magic box”, but a statistics-based tool, and it is not error-free (and never will !)**

And because you want ...

- **To select the best system for your needs**
- **Or, if you already got one, to make it perform better!**

Main motivation for deploying biometrics

“Even though no biometric modality is error-free, with proper system tuning and setup adjustment, critical errors of the biometric systems can be minimized to the level allowed for the operational use”.

And it is only through comprehensive performance evaluation that

- biometric systems errors, and
- factors / parameters that affect the recognition performance can be discovered and properly taken into account!

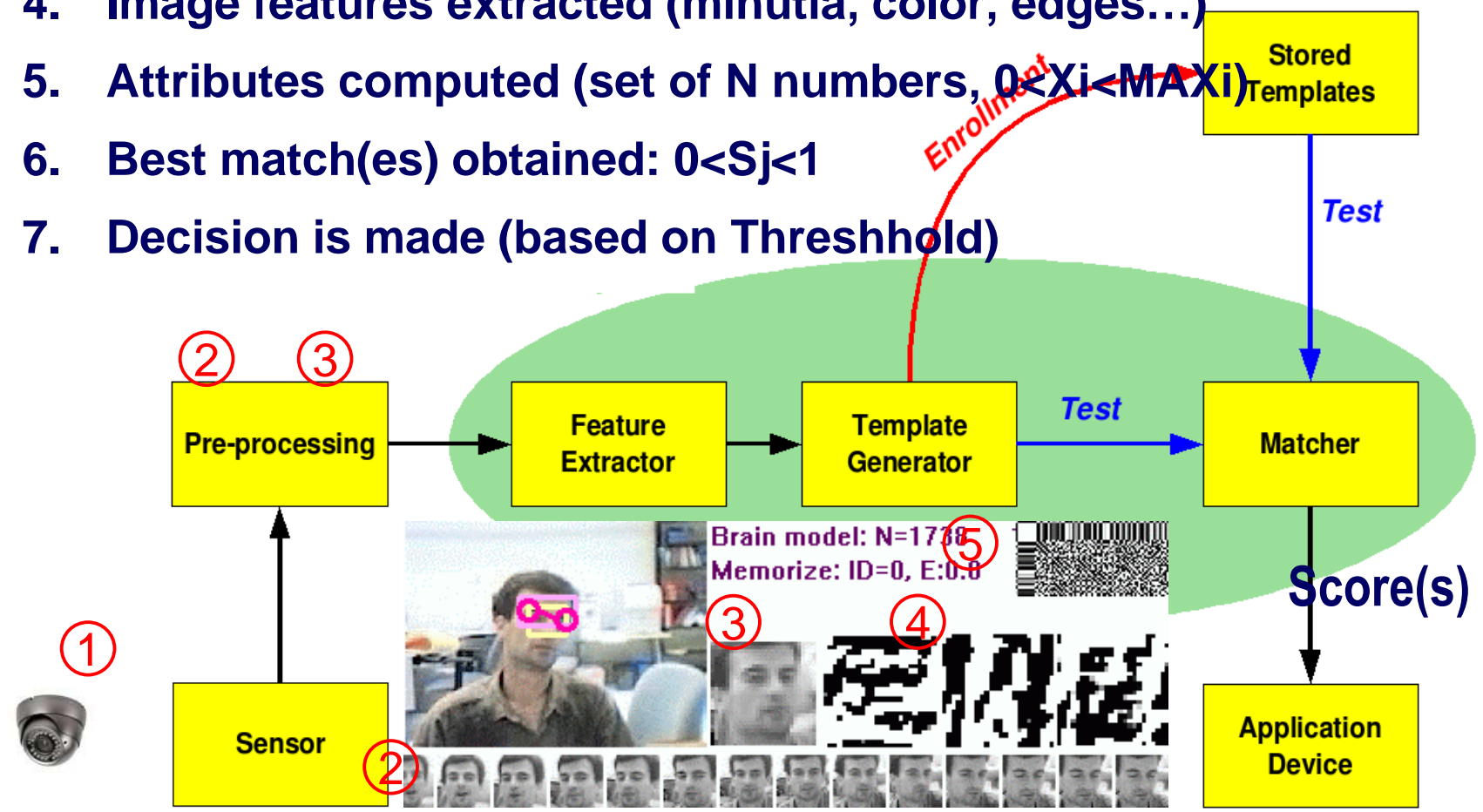
Why Biometrics may fail?



- 1. Image(s) captured
- 2. Best image(s) selected and enhanced - preprocessing
- 3. Biometric region extracted - segmentation
- 4. Image features extracted (minutia, color, edges...)
- 5. Attributes computed (set of N numbers, $0 < X_i < MAX_i$)
- 6. Best match(es) obtained: $0 < S_j < 1$
- 7. Decision is made (based on Threshold)

IP

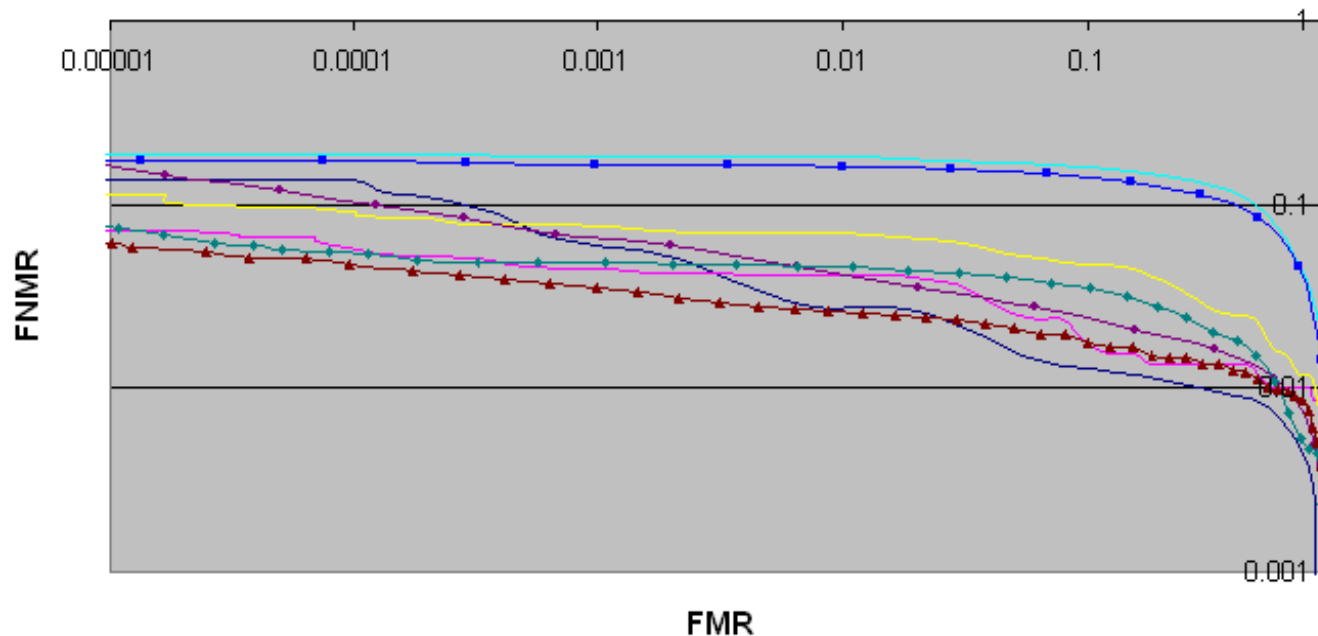
PR



Status-Quo methodology



- **False Match Rate (FMR)**
(False Accept, False Positive, False Hit, Type 1 Error)
- **False Non-Match Rate (FNMR)**
(False Reject, False Negative, False Miss, Type 2 Error)
- **Detection Error Trade-off (DET) curves** - the graph of FMR vs FNMR, which is obtained by varying the system parameters such as **match threshold**.



Limitations of basic metrics



What if [Wayman,...]:

1. there is more than one match below the threshold ?
2. there are two or more very close matching scores ?

A:

0.61

0.59

0.36 ***

0.49

0.57

B:

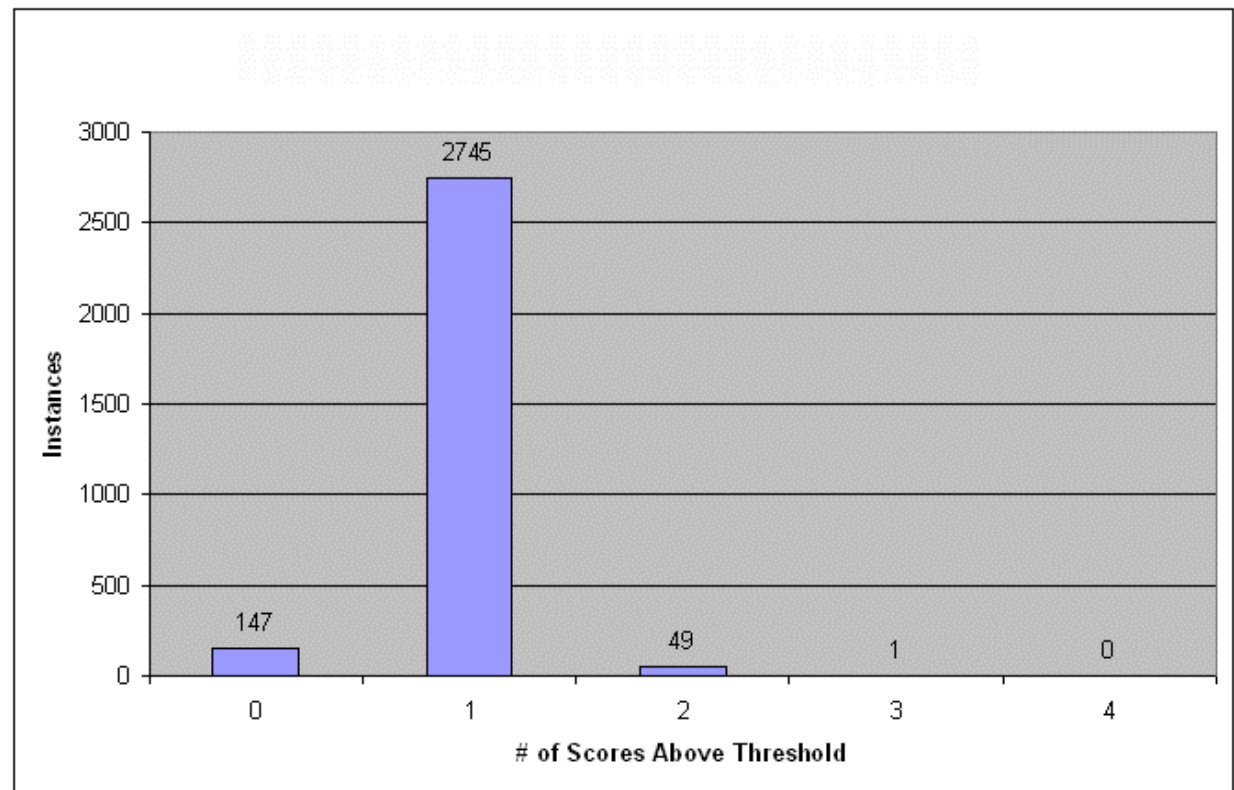
0.51

0.38 *

0.39 *

0.41 *

0.67



Canadian Contribution to ISO-SC 37 WG 5

There is currently no evaluation standard / methodology in industry (ISO SC-37, IBG) that is sufficient for operational use (eg. CBSA needs). - We had to develop it!...

➤ For ISO meeting in Moscow (July 2009):

“There is a need for a comprehensive biometrics performance evaluation standard that would take into account not only the best matching scores, but also the "runner-up" matching scores.”

➤ Added to ISO SC-37 WG 5 Roadmap: Biometrics Evaluations Gaps and Future Needs

General evaluation process



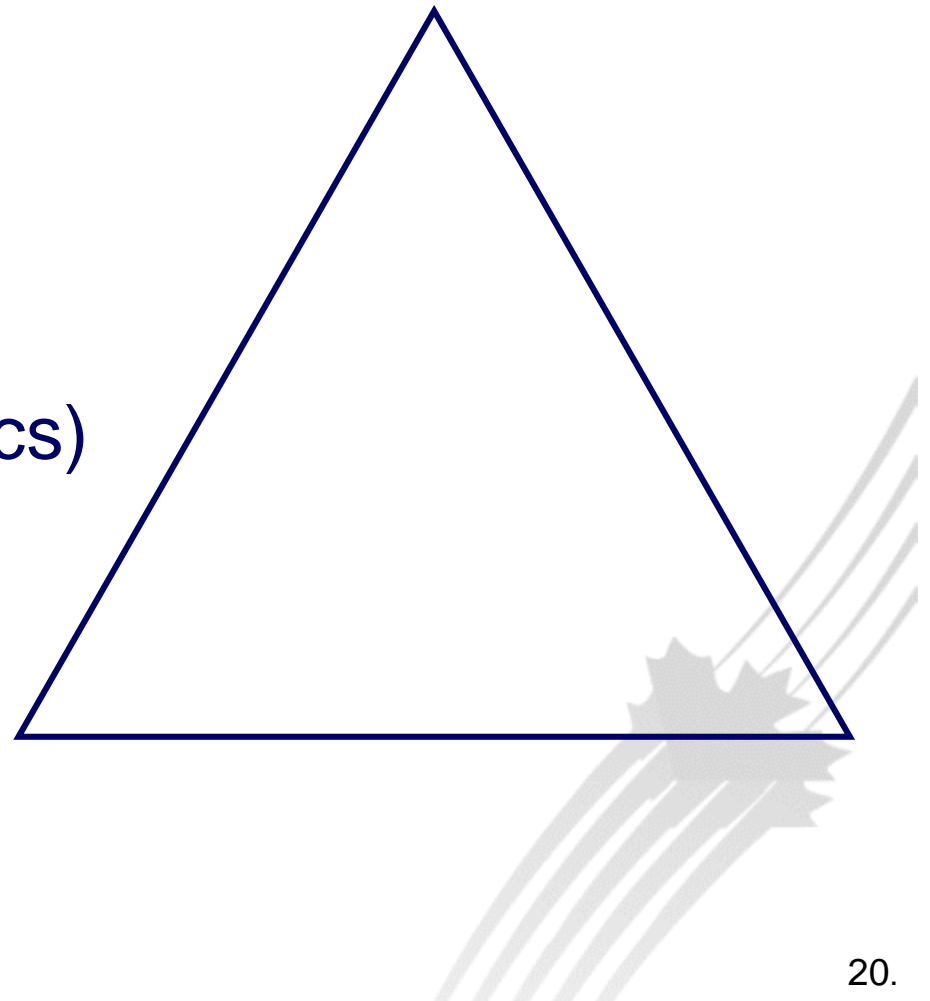
1. Determine suitability of modality (-ies)
2. Determine costs/impact of FM and FNM
3. Determine all factors affecting performance
4. Measure performance
 1. wrt all factors
 1. On large-scale database (>1000)
 2. On Pilot project (in real environment)
5. Evaluate the capability to be integrated / customized
 1. Wrt input parameters (pre-processing)
 2. Wrt output parameters (post-processing)

Know your Factors !



➤ THREE sources of problem:

1. Capture device
2. User
3. Light condition
(for image-based biometrics)



General protocol



Step 0: Data preparation

- Analyze and select Enrolled and Passage datasets:
 - of several sizes (N): 100, 500, 1000, 5000
 - corresponding to different factors/setup

Step 1: Encode ALL images (get binary templates)

- Record Failure to Acquire (FTA)

Step 2: Get ALL Scores for ALL image PAIRS

- A) For Enrolled – Imposters only
- B) For Passage – Imposters and Genuine

Step 3: Analyze ALL obtained scores (many,many...)

- Using multi-order analysis



Multi-order biometric performance analysis

Order 0: (Visualization only)

- Visualization of ALLs scores distributions

Order 1: (at Score-level) - Traditional

- Single-score statistics (FMR/FNMR) and trade-off curves

Order 2: (at Decision level)

- Examination of all scores and finding best (smallest) score:
“Does it belong to the genuine?”

Order 3: (at Confidence of Decision level)

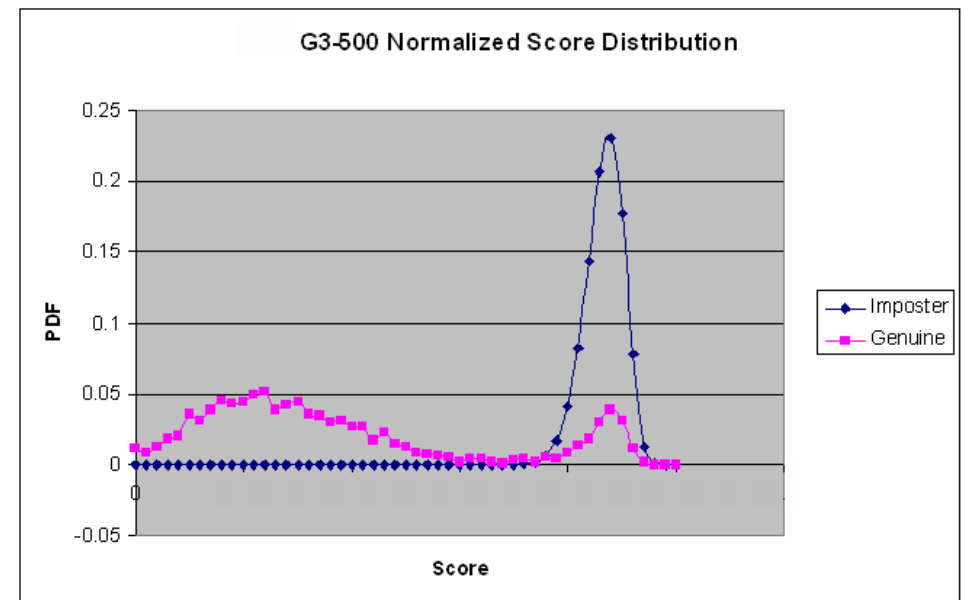
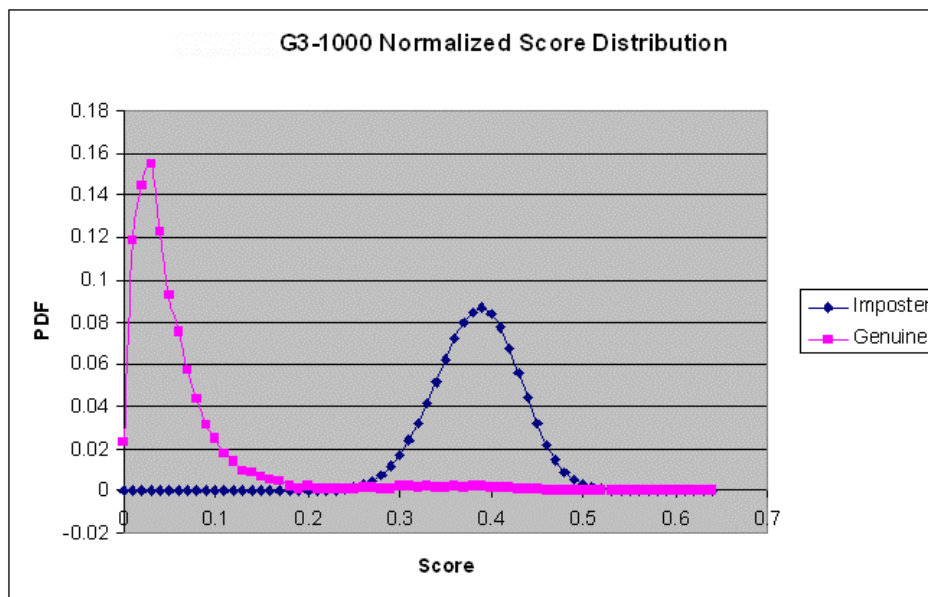
- Examine relationship between the scores:
 - See difference between best and second best scores,
 - See ALL scores below threshold

Order-0 analysis



Visualizing the score only:

- Just by looking at the score distribution (Order-0 Analysis), one may spot a problem or a deficiency of the system

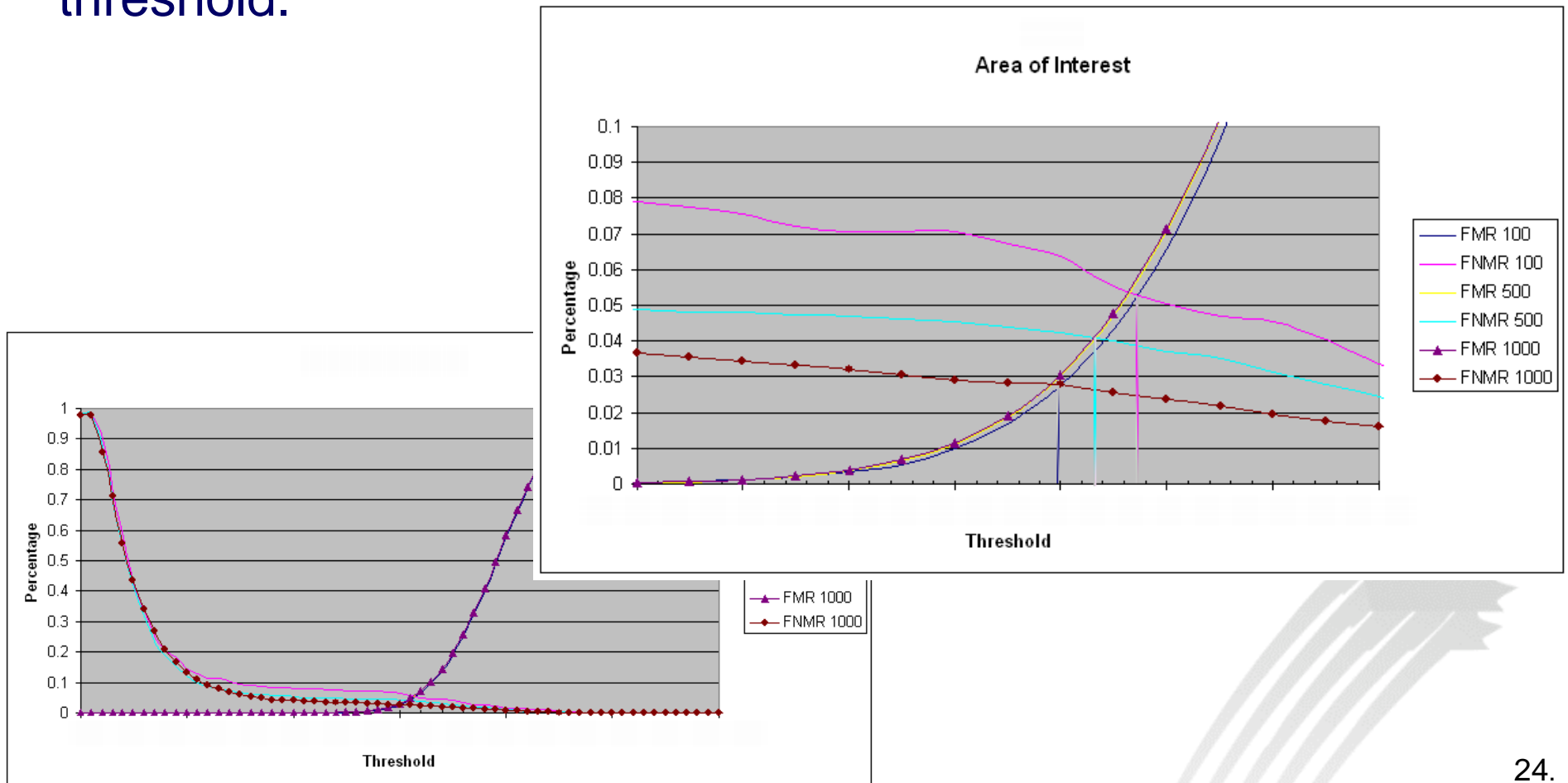


One system is (likely) NOT robust wrt to one (or more) factors present in the enrolled images.

→ Modify your setup or buy another system!

Order-1 analysis: FMR / FMNR curves

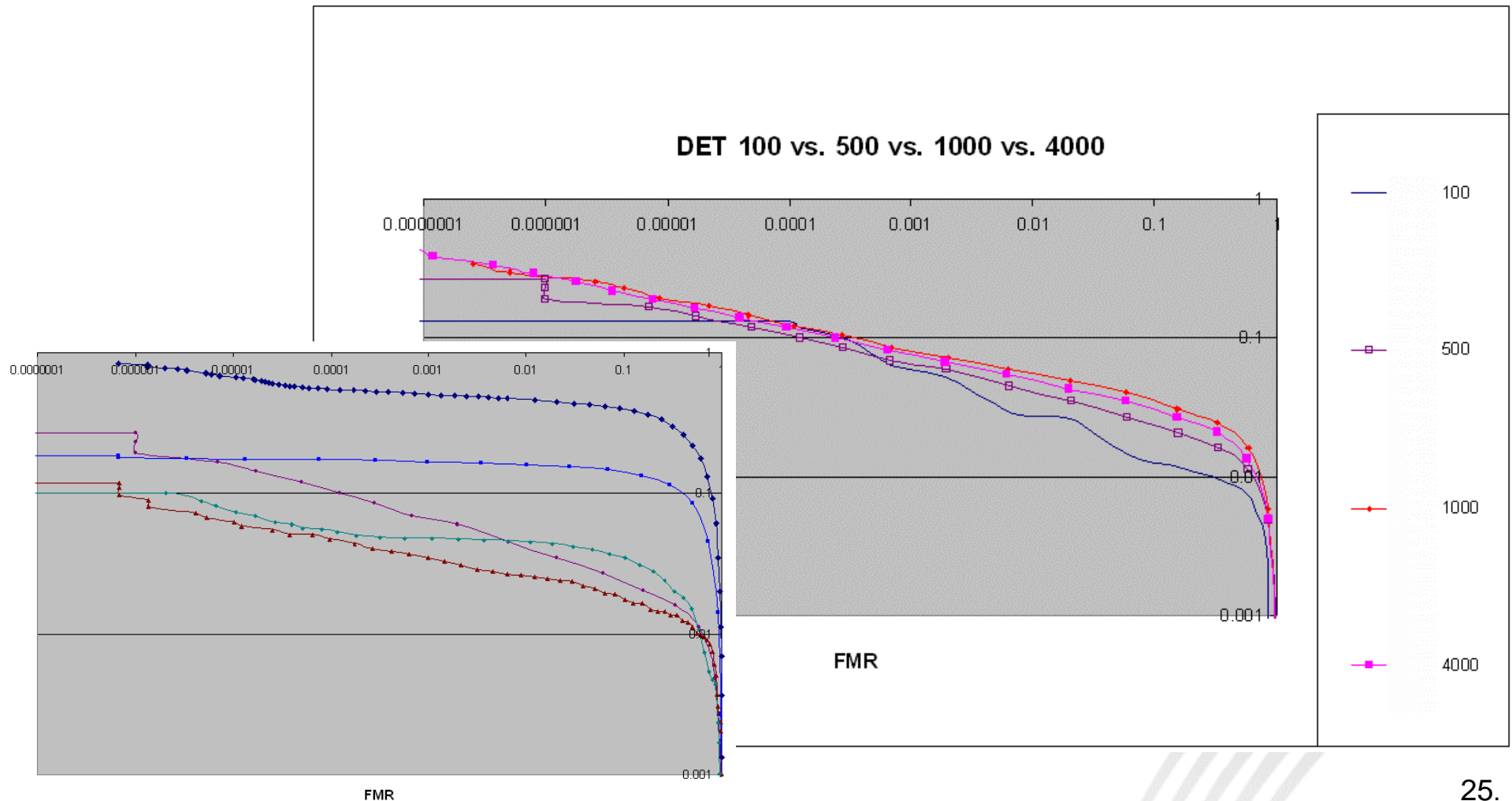
- By plotting FMR/FNMR as function of threshold for different data-set sizes, one may see how to optimally adjust the threshold.



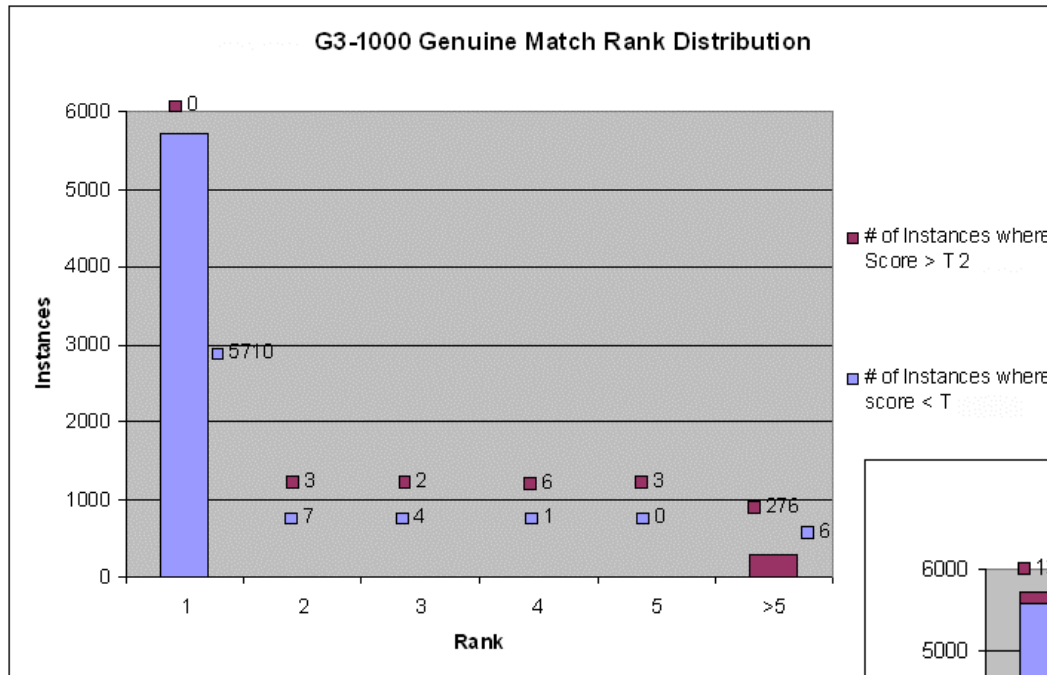
Order-1 analysis: DET curves



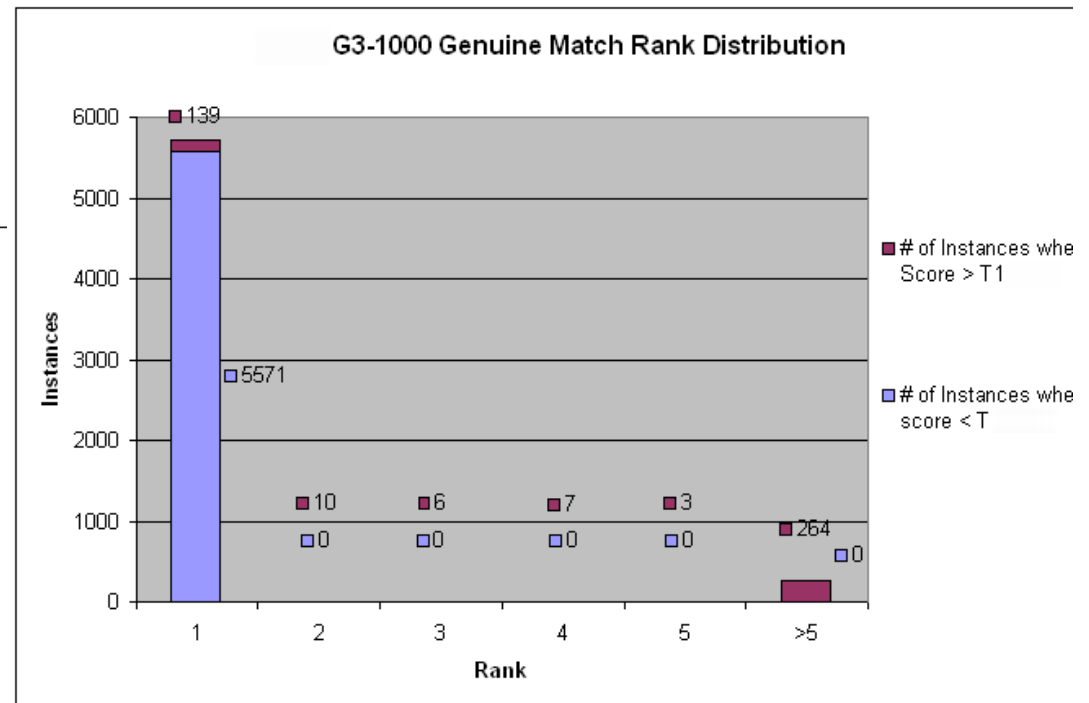
- Measured points must be shown, not only extrapolated lines!
 - Especially in the area of prime interest



Order 2. Do Genuine data have best scores ?



Which system/setup is better?...

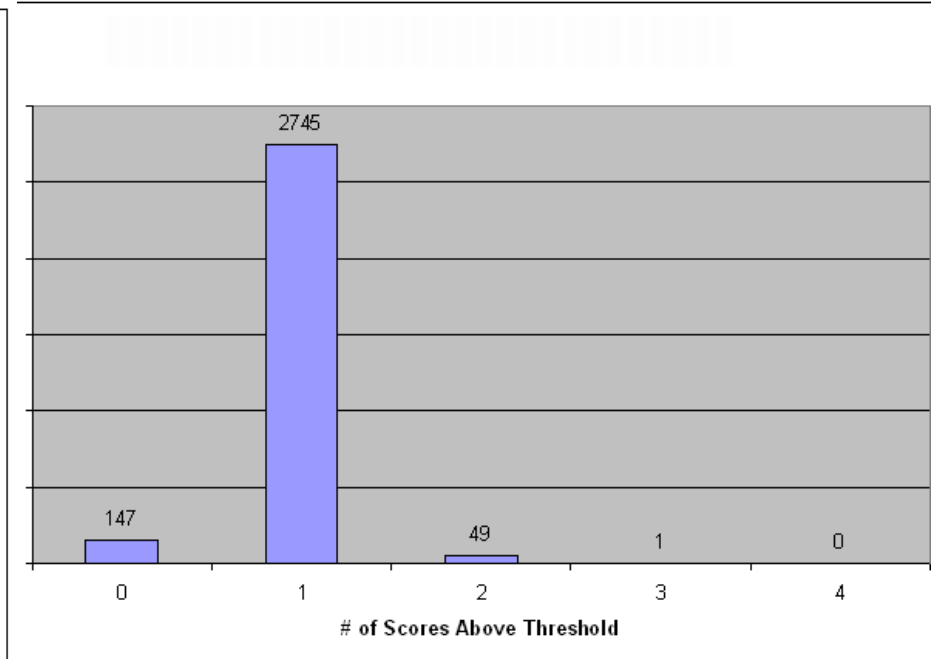
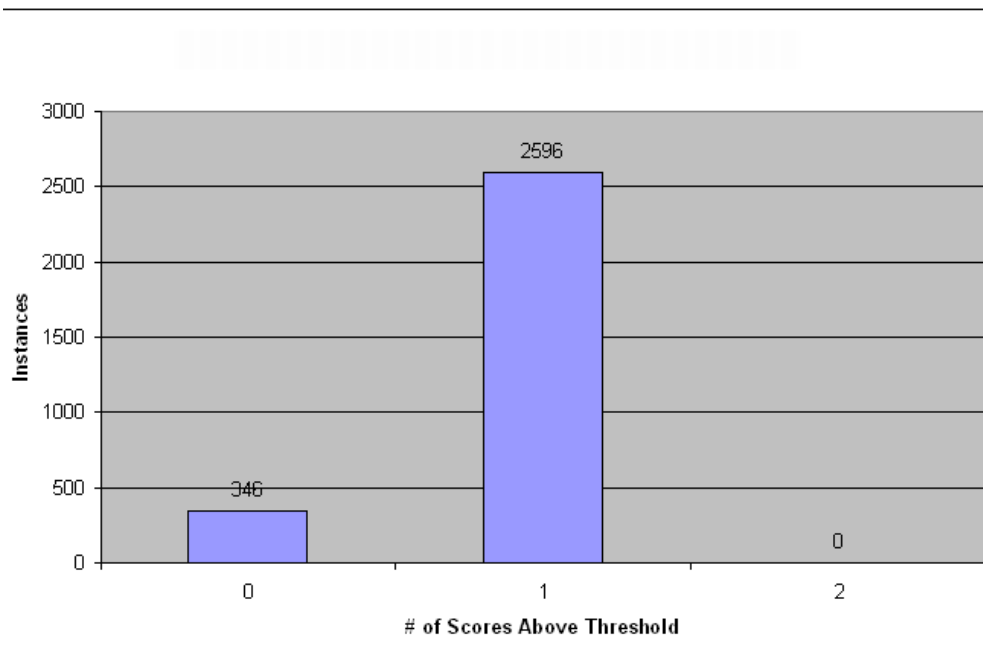


➤ How many times genuine was the 1st, 2nd, 3rd best score?

Order 3: Recognition confidence I



Many systems can improve the match/non-match tradeoff at the cost of allowing more than one scores below a threshold. (by raising the threshold) - Will you deploy it for Access Control ?!



➤ Number of scores below a threshold (for 3000 images).

Hits=2596, Misses=346

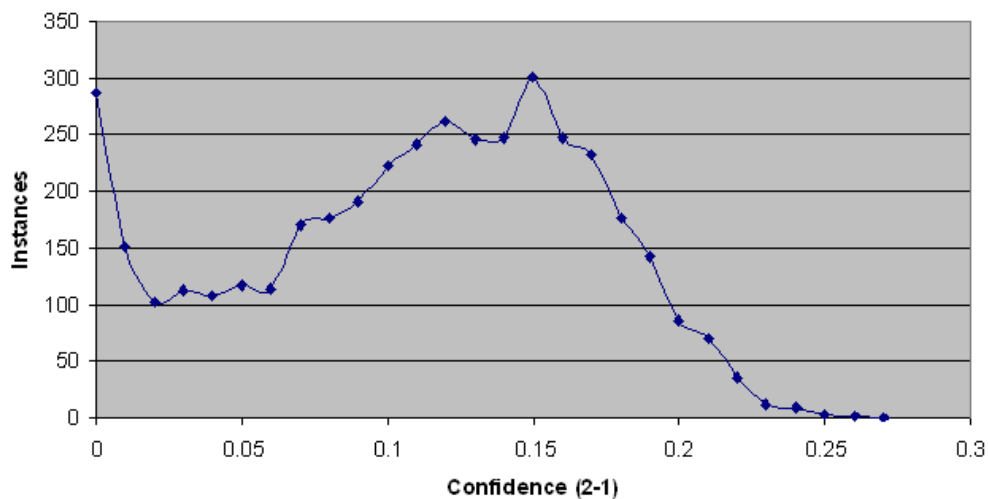
Hits=2745, Misses=147

Order 3: Recognition confidence II

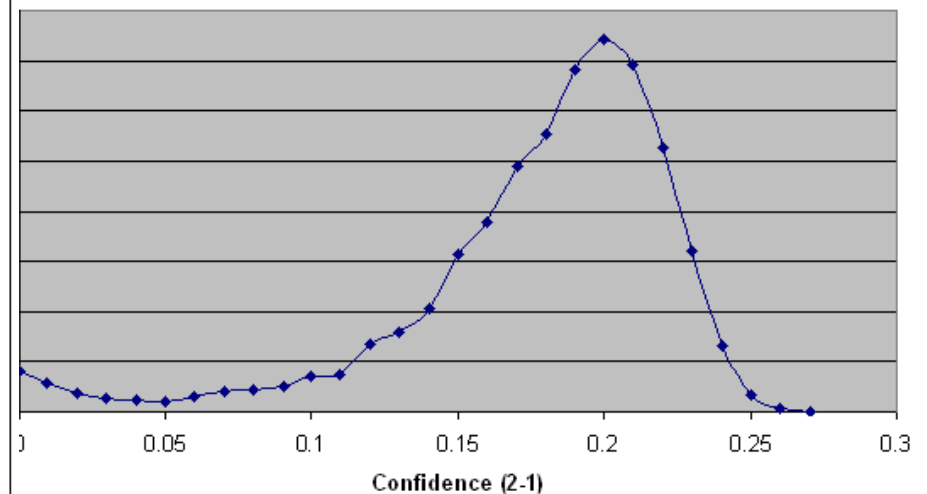


- Distance from “runner-up” and “winning” scores – Which do you prefer?

G3-1000 Confidence (2-1) T=0.38 Distribution for Rank 1 Scores Only



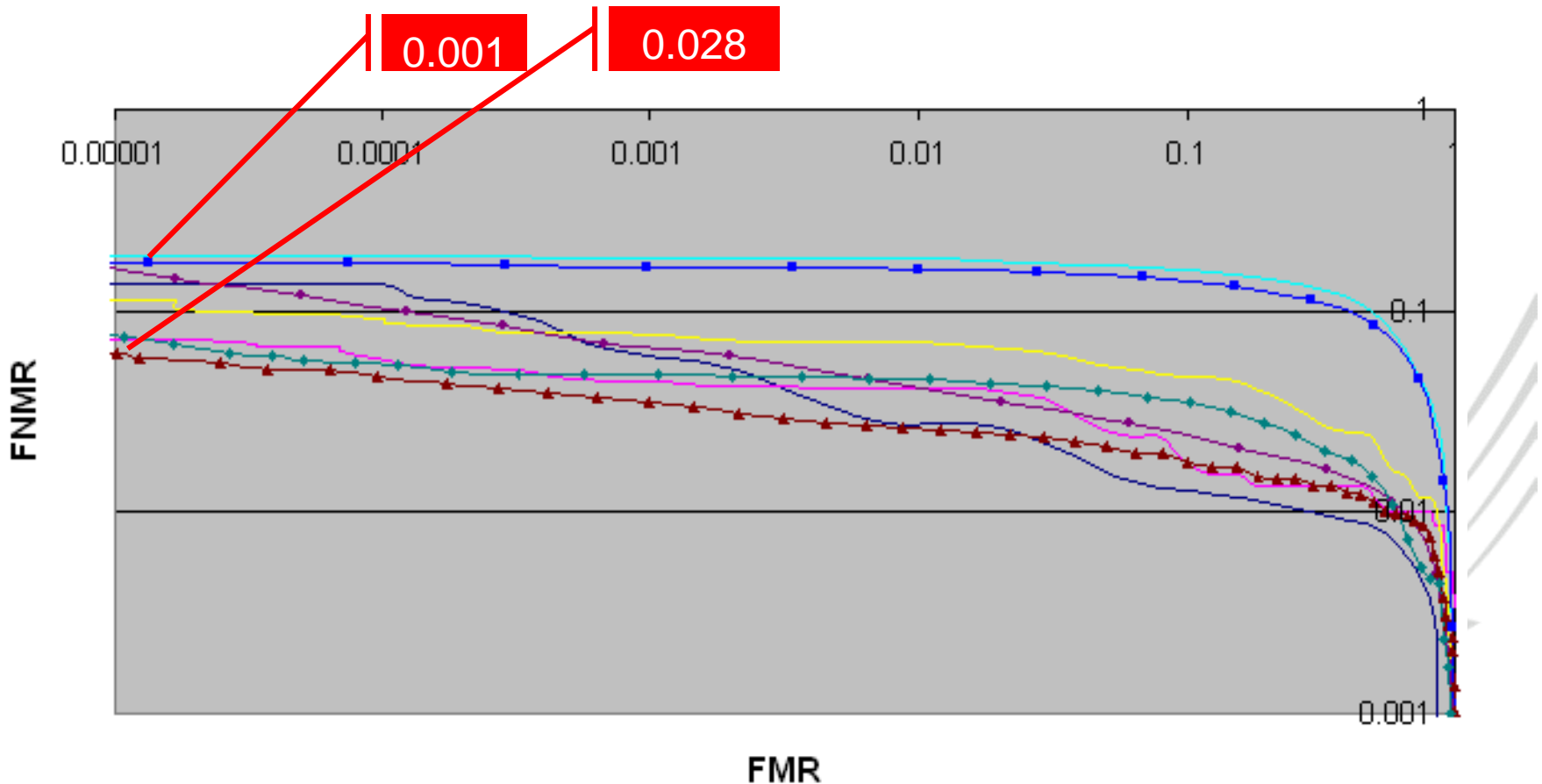
G3-1000 Confidence (2-1) Distribution for Rank 1 Scores Only



Trade-off Curves with FCR



DEFINITION: Failure of Confidence Rate (FCR) – the rate of incidences in which there are more than one match below threshold



In addition to Match/Non-Match Errors ...



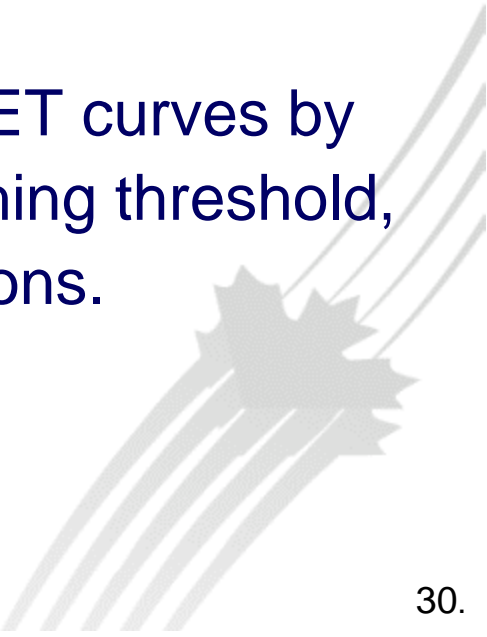
Report:

➤ FTA (Failure To Acquire):

because some systems may produce better DET curves by rejecting (i.e. failing to acquire) the images that are more difficult to recognize, eg. iris images that are occluded.

➤ FCR (Failure of Confidence Rate):

because some systems may produce better DET curves by allowing more matches below/above the matching threshold, ie by producing less reliable recognition decisions.



Performance Report Card




FTA=0.23	FMR	FNMR	FCR
	0.00067	0.0688	0.122
	0.00028	0.0854	0.059
	0.00012	0.1000	0.029
	0.000050	0.1195	0.013
	0.000017	0.1429	0.0048
	0.000007	0.1669	0.0008
	0.000001	0.1932	0.0004

Fig.8. All-inclusive biometric performance summary should report such information as FTA (Failure to Acquire rate) as well as FCR (Failure of Confidence Rates) in addition to commonly used False Match (FMR) / False Non-Match (FNMR) rates obtained by varying a match threshold.

Next Step: Threshold-Based Analysis

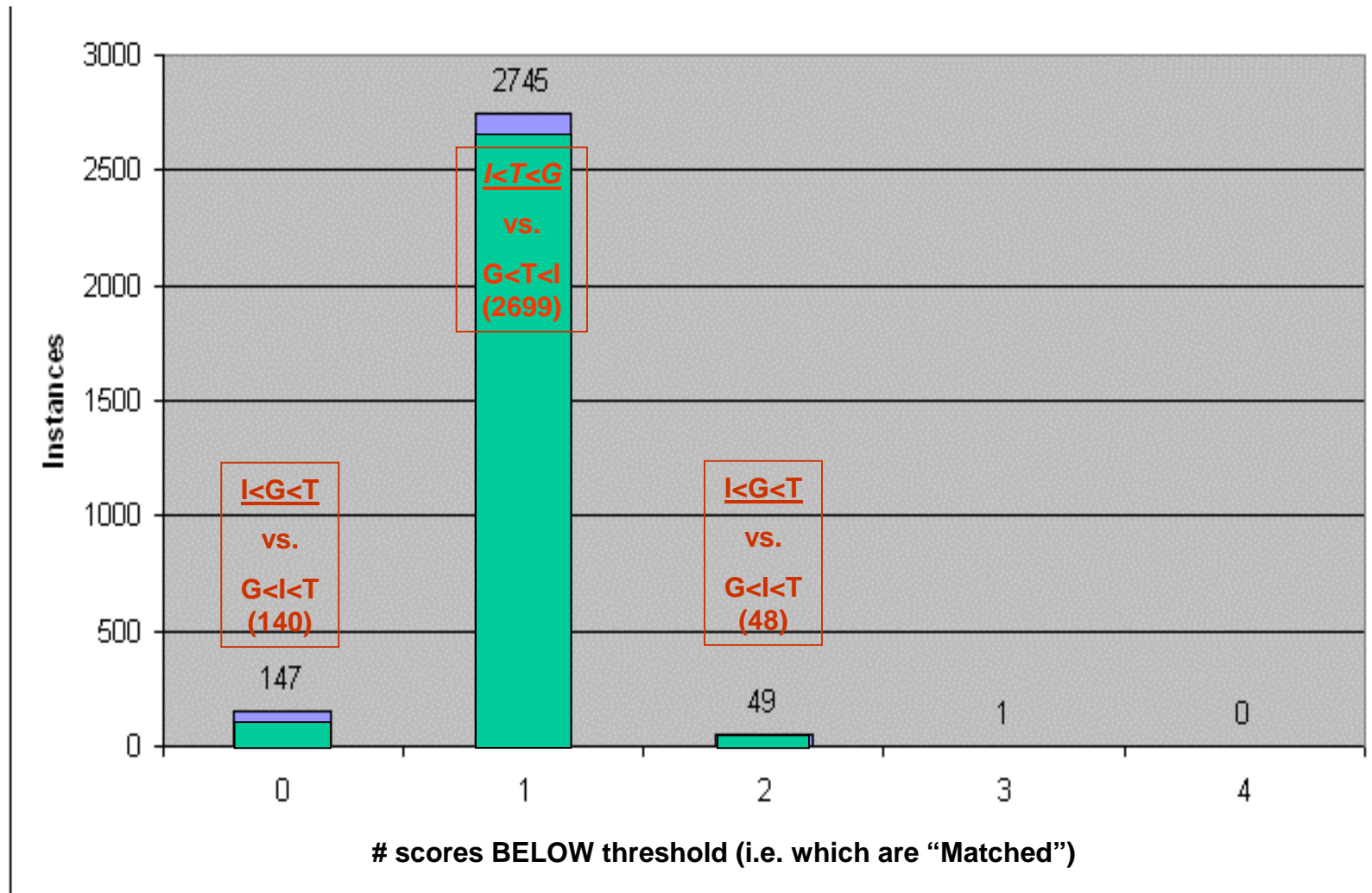


- Developed with IBG for PSTP08-0110BIO Study (2009-2010)
 - Each event falls into one of six categories. From most to least desirable:
 - **Genuine > Threshold > Impostor (G>T>I)**: highest genuine score exceeded threshold, highest impostor score lower than threshold
 - **Genuine > Impostor > Threshold (G>I>T)**: highest genuine and impostor scores each exceeded threshold, highest genuine score stronger than highest impostor score
 - **Threshold > Genuine > Impostor (T>G>I)**: no genuine or impostor scores exceeded threshold, highest genuine score stronger than highest impostor score
 - **Threshold > Impostor > Genuine (T>I>G)**: no genuine or impostor scores exceeded threshold, highest impostor score stronger than highest genuine score
 - **Impostor > Genuine > Threshold (I>G>T)**: highest genuine and impostor scores each exceeded threshold, highest impostor score stronger than highest genuine score
 - **Impostor > Threshold > Genuine (I>T>G)**: highest impostor score exceeded threshold, highest genuine score lower than threshold
- 

Order-3 Analysis: Threshold-Based Analysis

→ The distribution of the number of scores below a threshold (as in this paper)

+ The distribution of six possible {G, I, T} outcomes: $G < T < I$ (GOOD), ... , $I < T < G$ (BAD)



Future Work



➤ **C-BET (Comprehensive Biometrics Evaluation Toolkit):**

Under LoA with

DRDR-CSS (Defence R&D Canada, Center for Security Science)

➤ Apply to evaluation of new and traditional modalities:

- PSTP Study PSTP08-0110BIO: “Biometric Border Security”.
Lead: CBSA-S&E, Contractor: IBG. Delivery date: 31 March 2010

- PSTP Study PSTP08-0109BIO: “Stand-off Biometrics Evaluation”.
Co-lead with RCMP

➤ Next Improve Biometrics Performance by using Order-3 analysis: by introducing Confidence Scores based on thereon

- Gorodnichy, D.O., Hoshino, R. (2010). Calibrated confidence scoring for biometric identification. Proceedings of the NIST International Biometric Performance Conference.

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

➤ Contact: Dmitry.Gorodnichy@cbsa.gc.ca

THANK YOU!