



Canada Border
Services Agency

Agence des services
frontaliers du Canada

How To Conduct All-Inclusive Performance Evaluation of Your Biometric System

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

The wordmark for Canada, featuring the word "Canada" in a serif font with a stylized maple leaf above the letter 'a'.

Canada

Outline



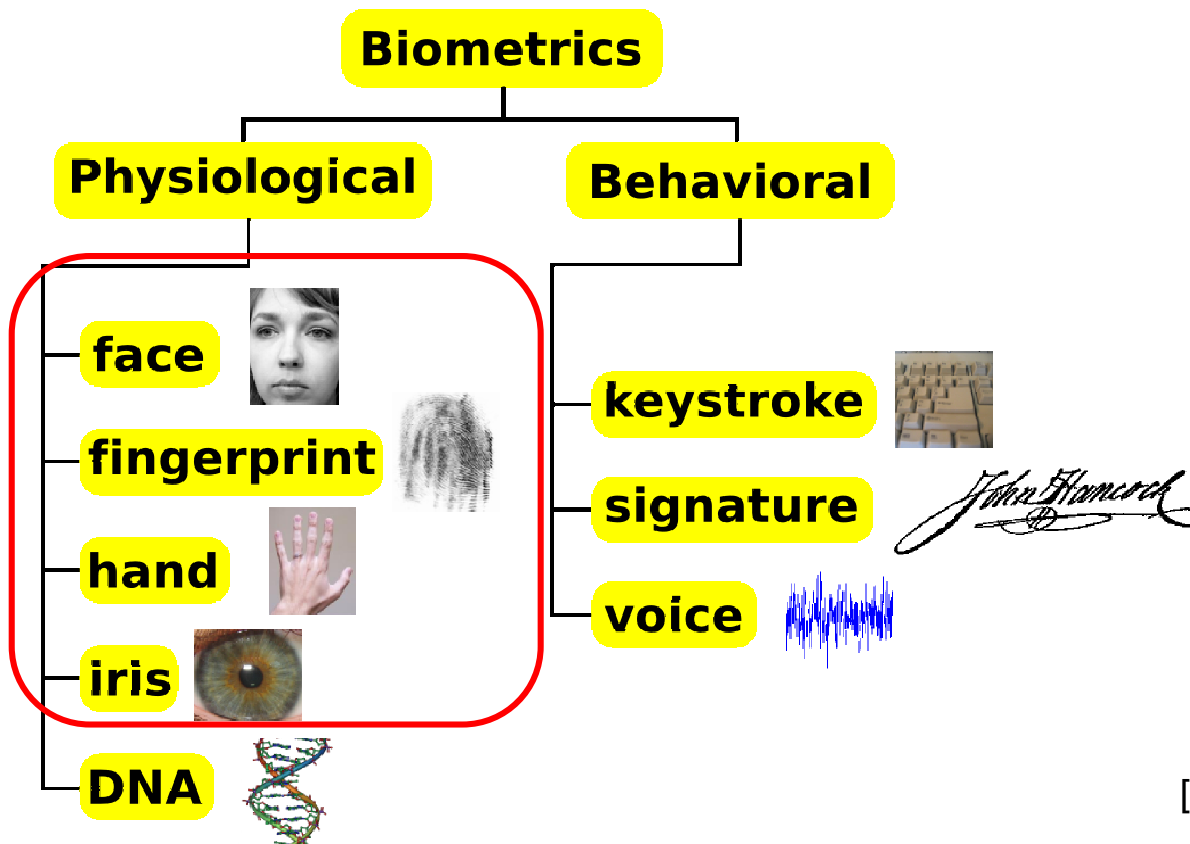
1. What's Biometrics? – to the user
 1. What is it used for?
 2. Biometric systems evolution and taxonomy
 - CBSA case studies: Iris, faces (ePassport, in video)
2. What's Biometrics? – to the developer
 1. Why Biometrics may fail? What are the factors?
 2. What is biometrics Performance Evaluation?
3. Conducting comprehensive Performance Evaluation
 1. General workflow
 2. Basic metrics: counting False Matches vs. Non-Matches
 3. Multi-order analysis: Taking into account all scores
 4. Examples – what to look for.
4. Lessons learnt

1. Biometric needs



What is Biometrics?

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



[CBSA NEXUS Iris Recognition system]

NB: Most biometric modalities are image-based!

Biometric system taxonomy



- By recognition task
- By operational considerations
- By environmental conditions

- By modality
- By modality characteristics

- By recognition performance requirements
- By decision making



Four “recognition” tasks



1. **Verification / Authentication: 1 to 1** (eg. Access Card)
 - Instant decision making, facilitated by pre-stored biometric data
2. **“White List” identification: 1 to N**
 - a) N is small/limited (eg. laptop users, secret facility personnel)
 - b) N is large and growing (bank clients, trusted travelers)
 - Instant decision making, facilitated by Cooperation from users
3. **“Black List” identification / Screening: 1 to M, M not large**
 - Time-weighted decision made by an Trained Analysts
 - With intelligence coming from difference sources
4. **Classification / Categorization: 1 to K, K – small**
 - What is his type, eg. Gender, Age, race ? (soft biometrics)
 - Whom of K people he resembles most ? (eg. tracking in video)

Operational Considerations



- Recognition task: verification vs. **identification (harder)**
- Overt vs **covert** image capture
- Cooperative vs **non-cooperative** participant
- Structured (constrained) vs **non-structured** environment
- Small database vs. **large database**
- Lighting conditions at enrollment and passage
- Relative impact (Cost) of False Match vs False Non-Match

Other:

- Training of staff etc.

Decision making



- Final decision
vs. list of options/candidates for further investigation
- Based only biometric data provided
vs. with additional data
- Instantly made – in live mode
vs. time-weighted decision – by go
- Fully automated (no human involved)
vs. made by Forensic Analyst

Example: Face Recognition

used for helping police/immigration to capture criminals
vs.

used for Access Control with ePassports and Smartgates

CBSA case study: Why iris?



Must be easily accepted by public:
not intrusive

Must not have a bad association

- To take a fingerprint image – be treated like a criminal
- To take a face (eyes) photo – be treated like a traveler

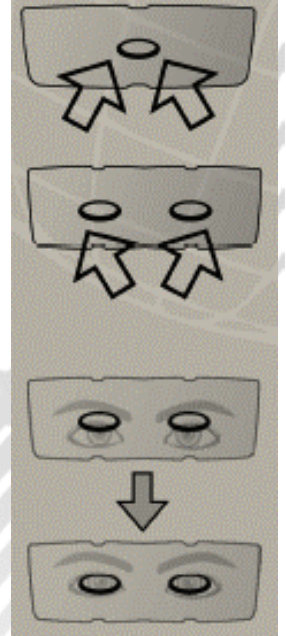
... so that people would voluntarily
participate (and cooperate)

Yet, be well-performing...



NEXUS as example of “White List” application

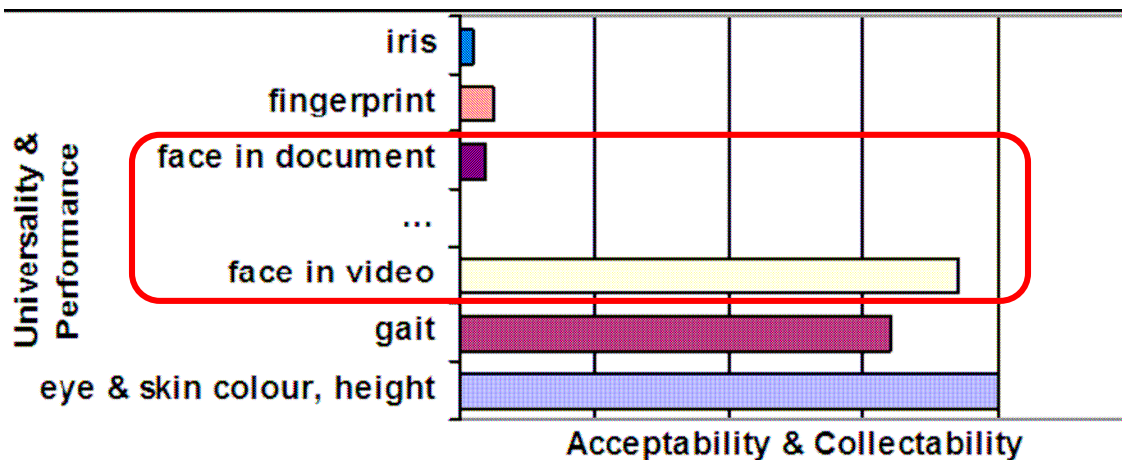
- Cost of False Match = security breach
 - False Match – not allowed, smallest possible FMR required
 - Result of spoofing attack (intentional defeat of a system)
- Cost of False Non-Match = inconvenience / frustration
 - The smaller number of FM, the larger number of FNM...
 - But members cooperate well to help the system



Biometric modality characteristics

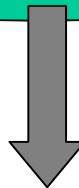
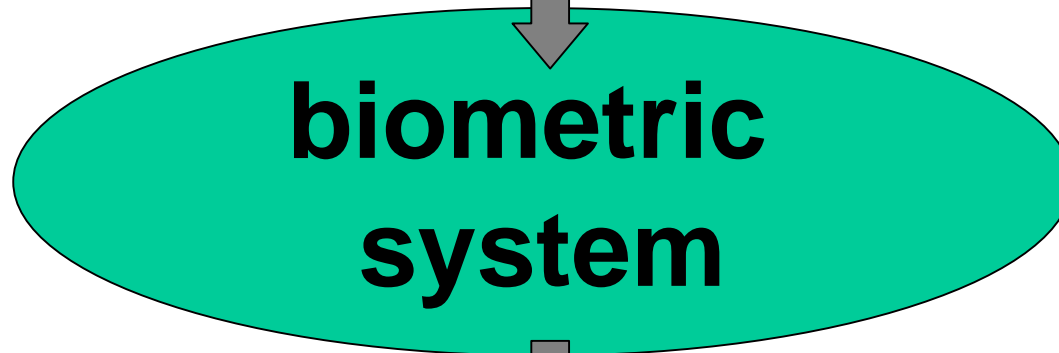
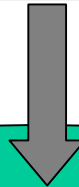
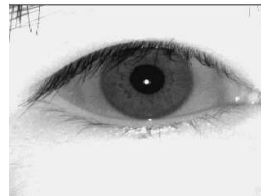


<u>Universality</u> : each person should have the characteristic	g
<u>Uniqueness</u> : how well separates individuals from one another.	g
<u>Permanence</u> : how well a biometric resists aging/fatigue etc	g
<u>Circumvention</u> : ease of use of a substitute.	? m
<u>Performance</u>: accuracy, robustness of technology used.	? m-g
<u>Collectability</u> : ease of acquisition for measurement.	? m-g
<u>Acceptability</u> : degree of approval of a technology	? m-g



2001 - 2009

2. What's inside that box ?



“Hello John Smith !”

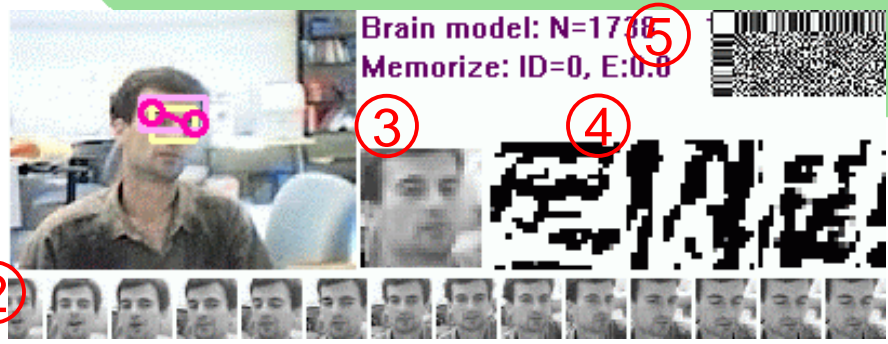
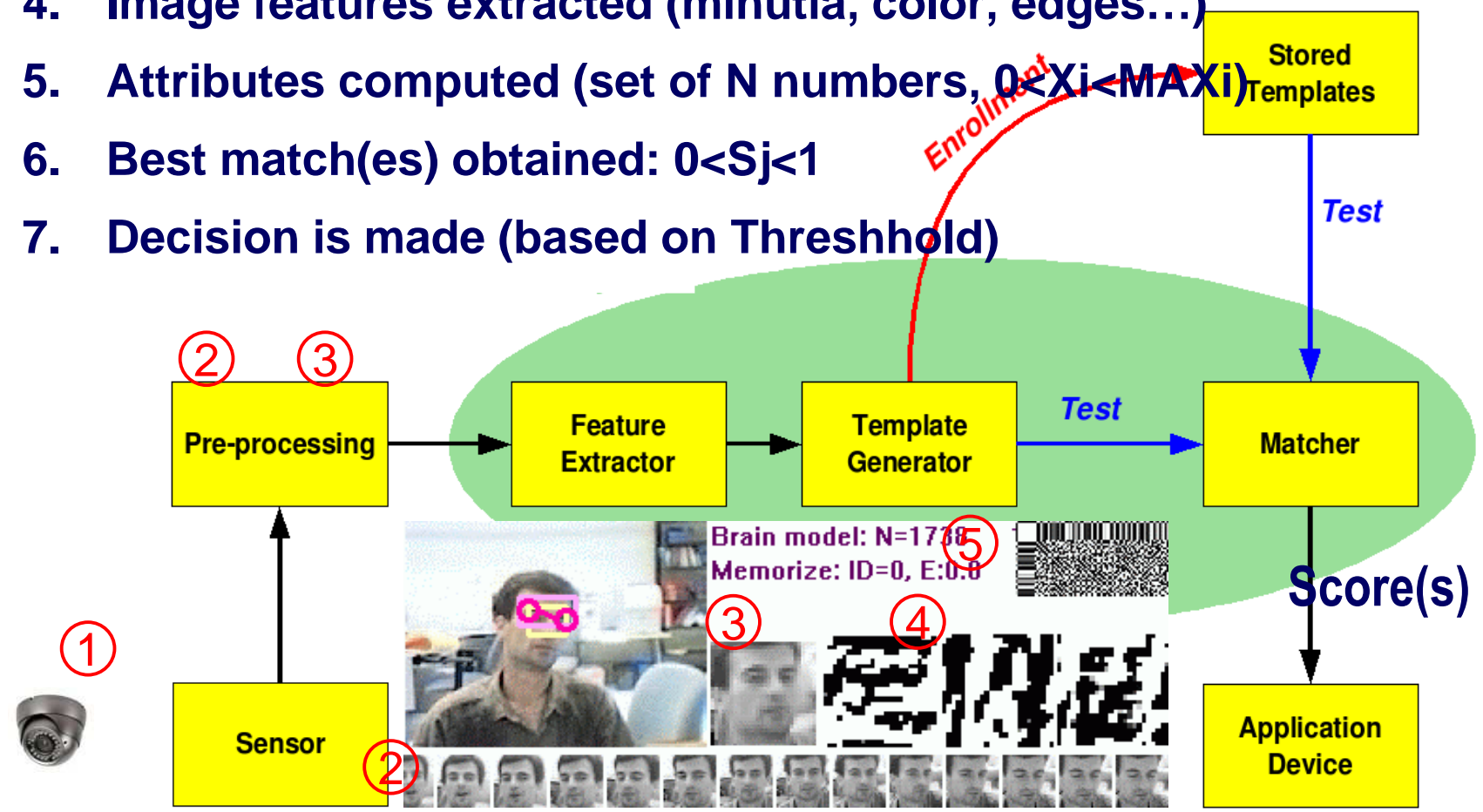


Why Biometrics may fail?



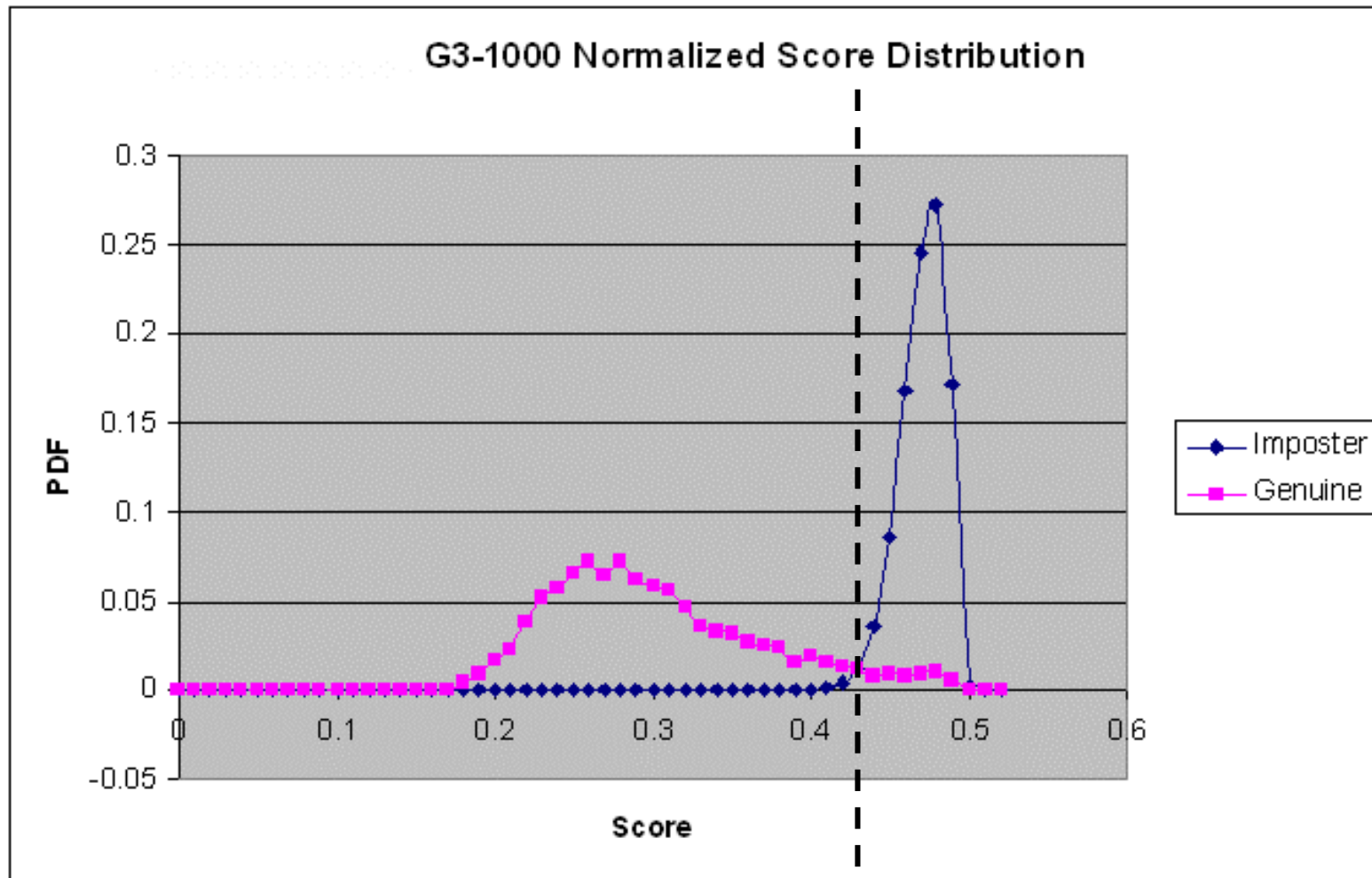
- 1. Image(s) captured
- 2. Best image(s) selected and enhanced - preprocessing
- 3. Biometric region extracted - segmentation
- 4. Image features extracted (minutia, color, edges...)
- 5. Attributes computed (set of N numbers, $0 < X_i < MAX_i$)
- 6. Best match(es) obtained: $0 < S_j < 1$
- 7. Decision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

IP {
PR {



Brain model: N=1778
Memorize: ID=0, E:0.0

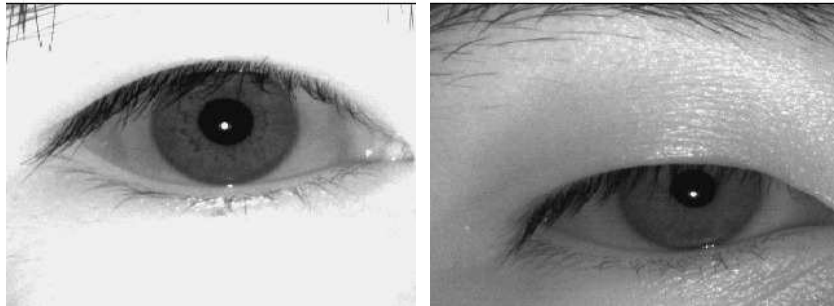
Scores for Genuine and Imposter users:



- NB: Traditionally, Match is when score is lower a threshold (ie. And it could be not the smallest score!)

Iris Recognition example

② Is image quality good?

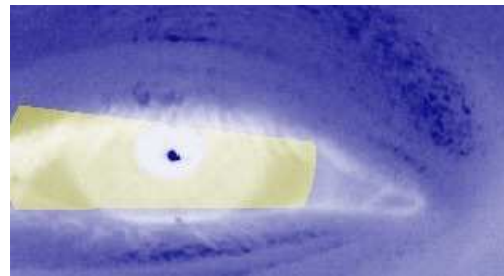
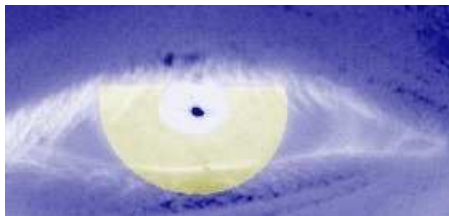


④ Are used features informative?

Is feature detail/numbers sufficient? ⑤



③ Is iris extraction good?



⑥ 5 Best scores:

0.51

0.38 *

0.39 *

0.41 *

0.67

⑦ Where to put Threshold?
What about confidence level?

Science behind Biometrics



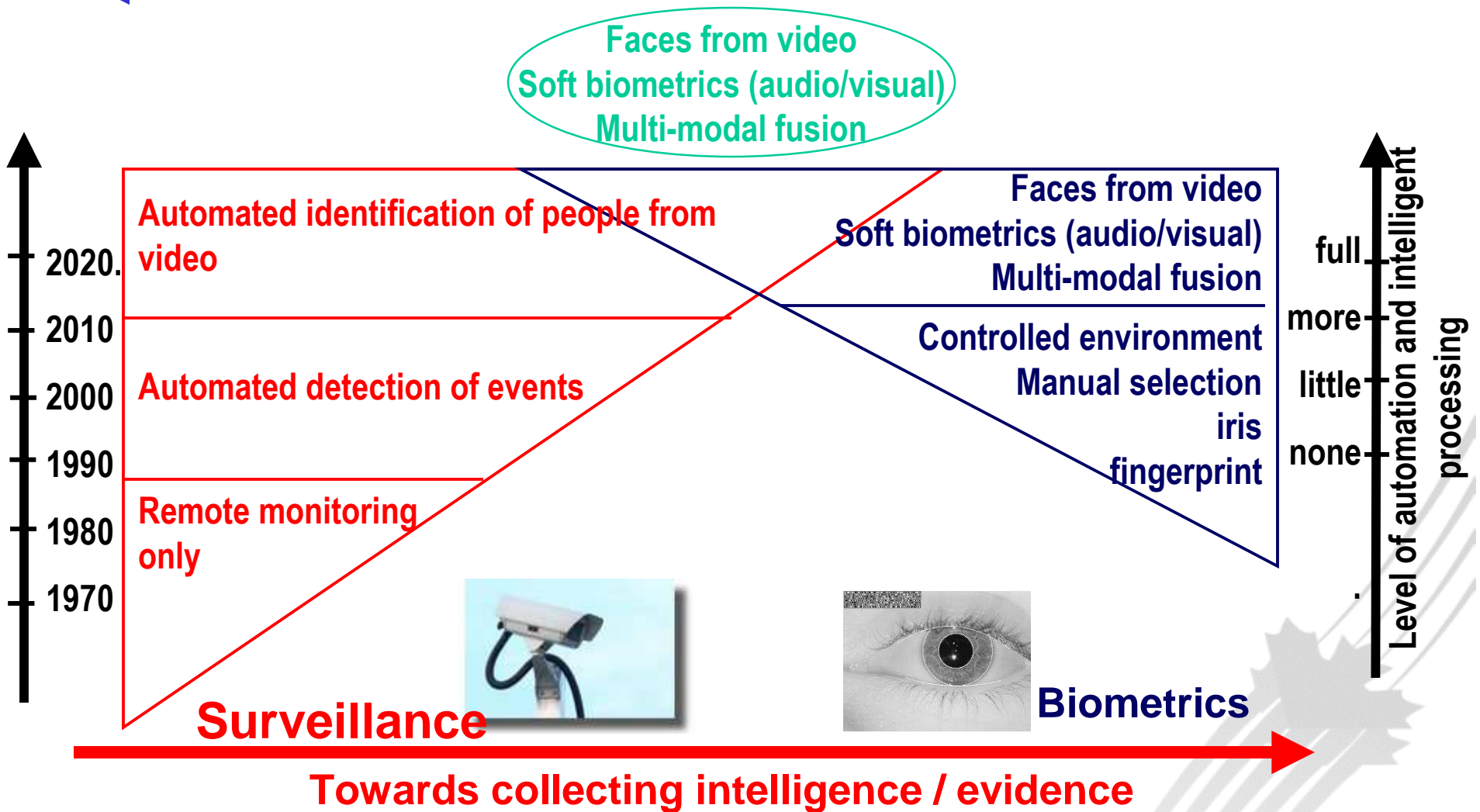
- Biometric recognition success is attributed to success in :
 - **IP: Image Processing** theory to deal with variability of the quality of the captured iris.
 - **PR: Pattern Recognition** (statistical machine learning algorithms) to determine the similarity between templates.

- Academic Conferences: ~10 big ones, ~ 10000 papers / year
 - Canadian IAPR Conference on Robot and Computer Vision (CRV)
 - IEEE CVPR, ECCV, ICCV, ICIP, BCMV ...

Evolution of Biometrics



Towards more collectable, unconstrained environments

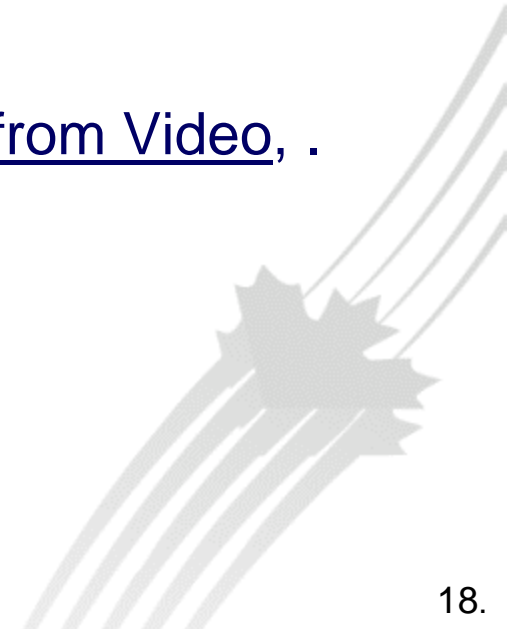


New biometric technologies



As a result of evolution, the arrival of such biometric technologies as

- Biometric Surveillance,
- Soft Biometrics,
- *Stand-off Biometrics*, also identified as Biometrics at a Distance, Remote Biometrics, Biometrics on the Move or Biometrics on the Go
 - And increased demand for Face Recognition from Video, .



Stand-off vs. Stand-in biometrics



Stand-in biometrics: a person intentionally comes in contact with a biometric sensor

Stand-off biometrics: without person's direct engagement with the sensor (In many cases, s/he may 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 is normally much less identifying. This means two things:

1. → there could be more than one match below the threshold, or two or more very close matching scores.
2. → final recognition decision is not based on a single measurement, but on a number of biometric measurements taken from the same or different sensor, combined together using some data fusion technique

To conclude...



Why to conduct evaluation ?

Because ...

- **Biometric system is not a “magic box”, but a statistically-derived 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!**



3. How to do it and what to watch for



What is performance evaluation ?



Definition: *Performance evaluation* is a procedure that involves the testing of a system on a database and/or in a specific setup for the purpose of obtaining measurable statistics (metrics) that can be used to compare systems or system setups to one another.

Basic Measurements:

- Scores
- Decisions (correct vs. incorrect) :
 - False Match (FM), True Match
 - False Non-match (FNM), True Non-Match

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)

Factors that affect the performance



➤ THREE sources of problem:

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






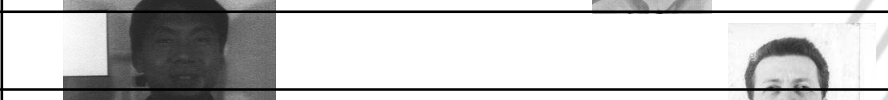






Example: seven Face Recognition factors

1. face image resolution*
2. facial image quality*
 1. blurred due to motion,
 2. lack of focus,
 3. low contrast due to insufficient camera exposure or aperture,
3. head orientation*
4. facial expression/variation
5. lighting conditions*
 1. location of the source of light with
 - respect to the camera, - users and - other objects
6. occlusion*
7. aging and facial surgery*

* - concerns ALL image-based biometrics

Example: Face Recognition databases*

Database (year created) / Users	#individuals/ # images	IOD / image width	Orient ation	Expres sion	Lightin g / quality	Occ lusi on	situ atio ns	Representative Facial image
AT & T Olivetti (1992-1994)	40 / 400	~60 / 92	yes	yes	yes	yes		
FERET (1993-1996)	1999 / 14,126	~ 80 / 256	9-20	2	2		2	
PIE 2000	68 / 41,368	~75/ 640	13	3	43			
Cas-peal 2003	1040	~45 / 360	21	15	6		1-5	
Korean	1000/ 52000	~80/ 640	7	5	16			
Equinox	91	~100 / 240	1	3	3 -IR images			
Cmu- hyperspectral	54	80/ 640	1		4 -IR images		5	
U of Texas 2002	284	~80/ 720	video	video				
FRVT * HCInt 1999-2002	37,437 / 121,589	>100	1				3	
FRVT * MCInt 1999-2002	>100 63	>80, <80	sever al	Still and video	severa l		yes	

* From [Gor2008]



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

Basic performance metrics

- False Match Rate (False Accept, False Positive, Type 1 Error)
- False Non-Match Rate (False Reject, False Negative, Type 2 Error, Miss rate)
 - which are tradeoff off one another !
 - functions of many parameters (main - Threshold)

Consider Cost (Impact) of these errors!

- In “White list”: FM cost > FNM cost
- In “Black list” – the opposite

Focus on those error that costs the most!

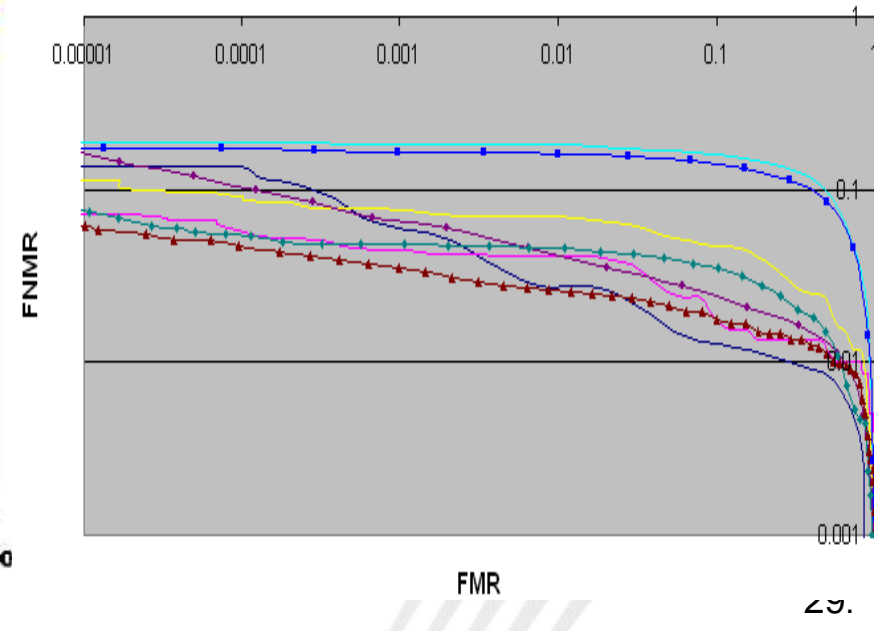
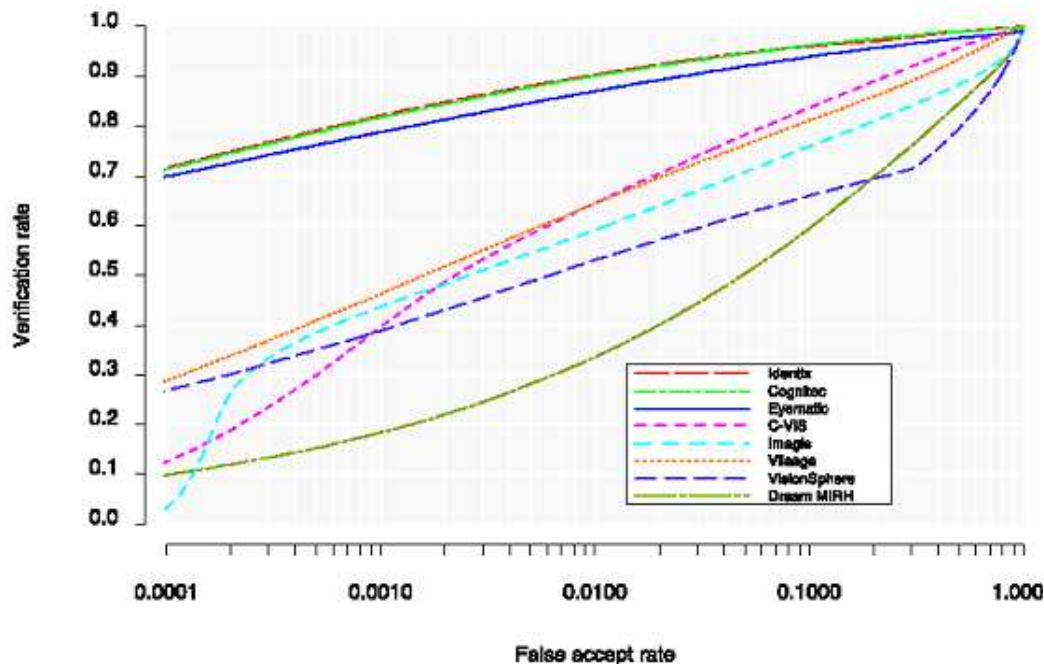
0.19	0	36
0.2	0	69
0.21	0	93
0.22	0	155
0.23	0	211
0.24	0	230
0.25	0	269
0.26	0	295
0.27	0	263
0.28	0	292
0.29	0	253
0.3	0	236
0.31	1	226
0.32	1	191
0.33	6	146
0.34	7	134
0.35	19	130
0.36	47	108
0.37	85	102
0.38	224	98
0.39	559	64
0.4	1563	79
0.41	4464	64

Error trade-off curves



Detection Error Trade-off (DET) curve: graph of FMR vs FNMR, which is obtained by varying the system parameters such as **match threshold**.

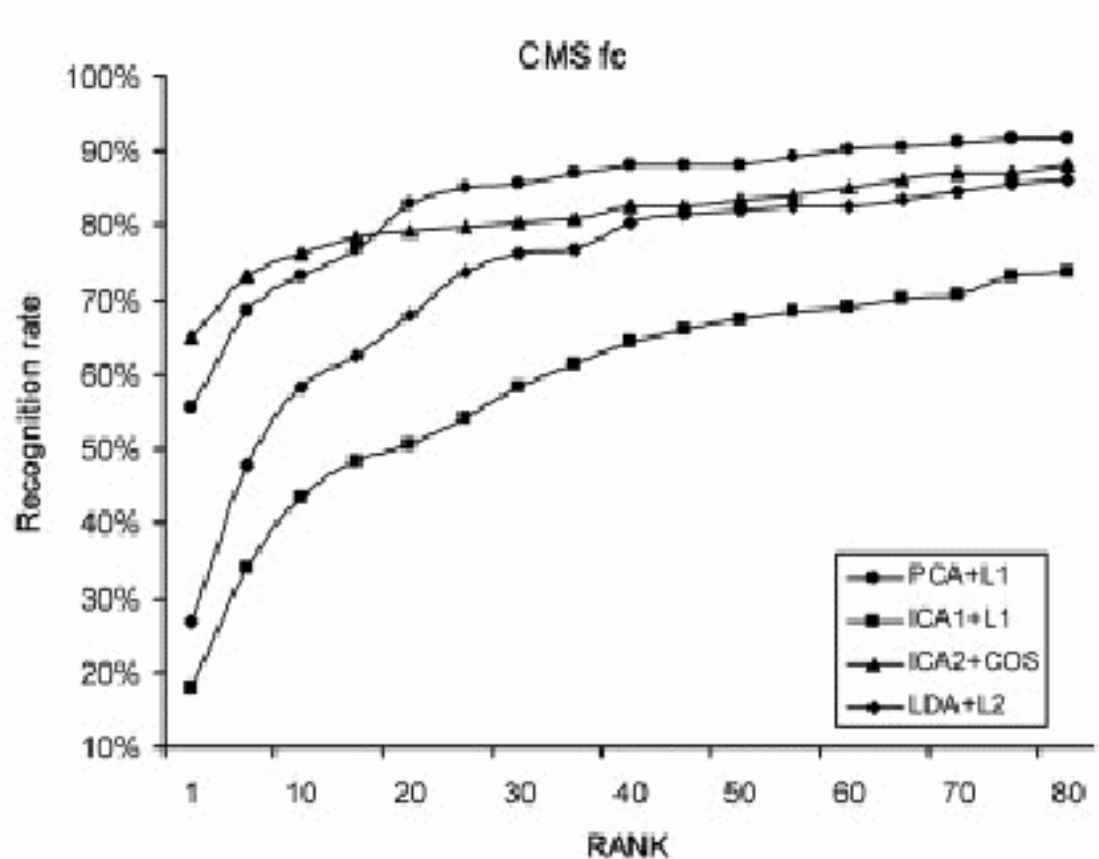
Receiver Operator Characteristic (ROC) curve similar to DET curve, but plots TMR (True Match Rate) against FMR.



Identification curves

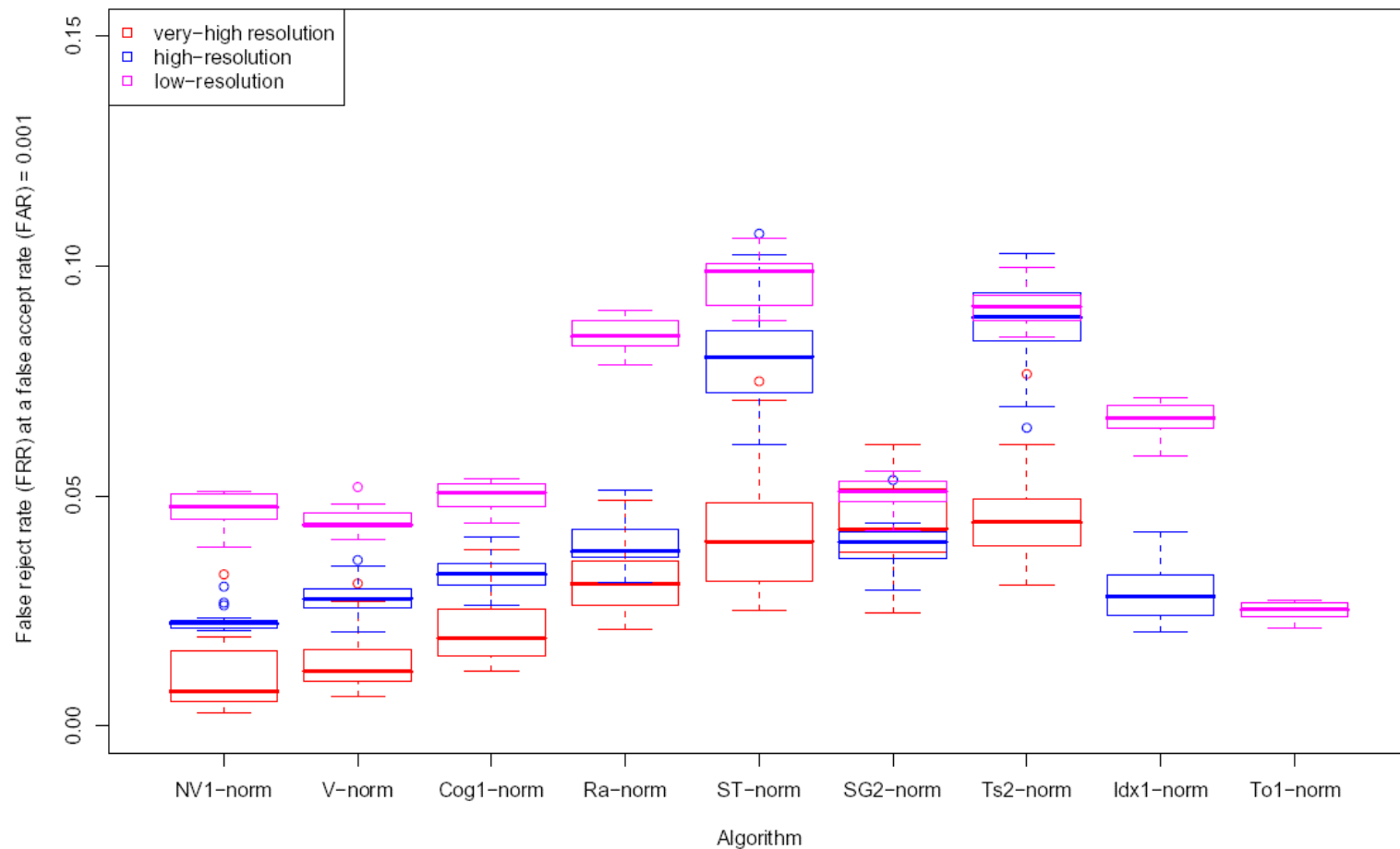


- *Rank-k identification rate (R_k)* - the number of times the correct identity is in the top k most likely candidates.
- *Cumulative Match Characteristic (CMC) curve* - plots the rank-k identification rate against k.




Other measures

- Equal Error Rate
- FNMR at fixed FMR (eg. 0.001, if cost of FM is high)



Reported error rates



Biometrics	EER	FMR	FNMR	Subjects	Comment	Reference
–	–	–	–	–		
<u>Face</u> 	n.a.	1%	10%	37437	Varied lighting, indoor/outdoor	FRVT (2002) ^[24]
<u>Fingerprint</u>	n.a.	1%	0.1%	25000	US Government operational data	FpVTE (2003) ^[25]
<u>Fingerprint</u>	2%	2%	2%	100	Rotation and exaggerated skin distortion	FVC (2004) ^[26]
<u>Hand geometry</u>	1%	2%	0.1%	129	With rings and improper placement	(2005) ^[27]
<u>Iris</u>	< 1%	0.94%	0.99%	1224	Indoor environment	ITIRT (2005) ^[28]
<u>Iris</u>	0.01%	0.000 1%	0.2%	132	Best conditions	NIST (2005) ^[29]
<u>Keystrokes</u>	1.8%	7%	0.1%	15	During 6 months period	(2005) ^[30]
<u>Voice</u>	6%	2%	10%	310	Text independent, multilingual	NIST (2004) ^[31]

- Obtained for a specific data/setup/conditions/constraints
- These rates can not be relied upon, but should be used as a guide only!

Limitations of basic metrics



What if:

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:

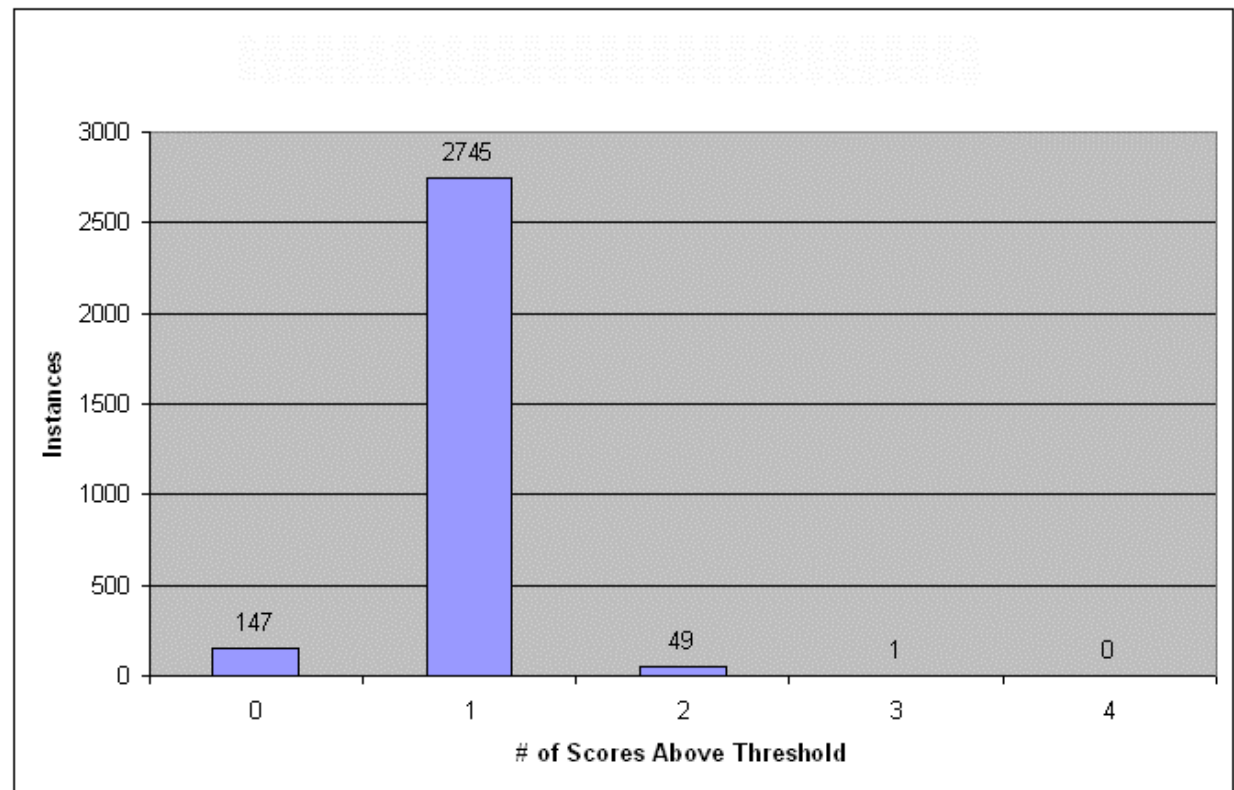
0.51

0.38 *

0.39 *

0.41 *

0.67



Request for new ISO Biometrics standard

ISO/IEC SC 37 N 3060 (May 2009) – “Canadian contribution on Future Testing needs”:

“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.

Such standard would be most applicable for evaluation of stand-off identification systems, such as Face or Iris Recognition from Video. The standard however would also be applicable for verification systems, in particular to those that make the decision based on examining the entire database of enrolled identities.”

ISO/IEC SC 37 N 3371 (WG 5 Roadmap, July 2009) -

7. Perceived Requirements for New Standards and Technical Reports:

“A contribution from Canadian National Body (37N3145) might form the basis of future work”

Multi-order biometric performance analysis

Order 0:

- See ALLs scores distributions

Order 1 (Traditional):

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

Order 2:

- Examine all scores and see best (smallest) scores:
“Do they coincide with genuine score?”

Order 3:

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

General workflow



- 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

Estimate time required



- Step 2b (getting the scores) is most time consuming

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

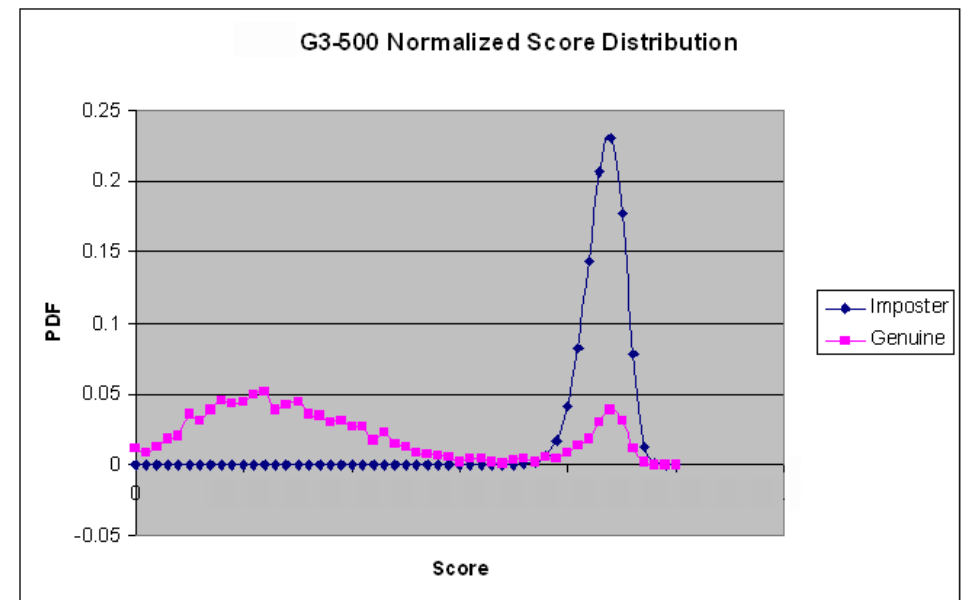
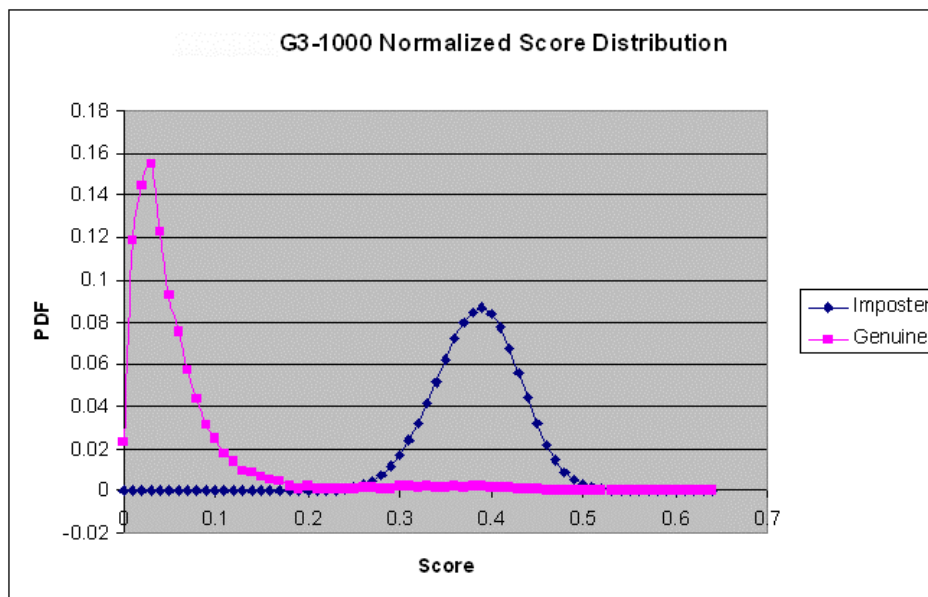


Order-0 analysis



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

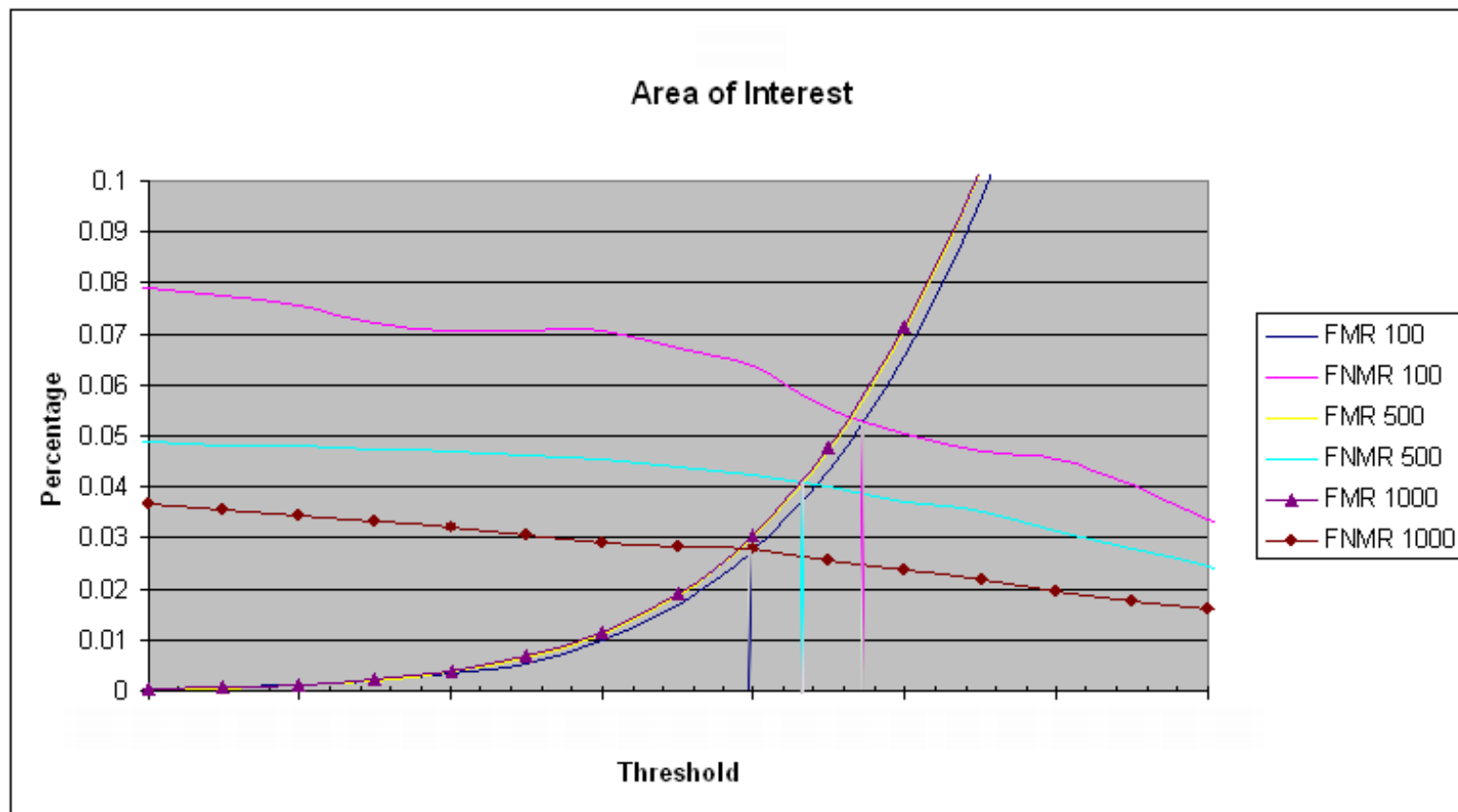


One system displays intolerance to one (or more) factors present in the enrolled images.

→ Modify your setup or buy another system!

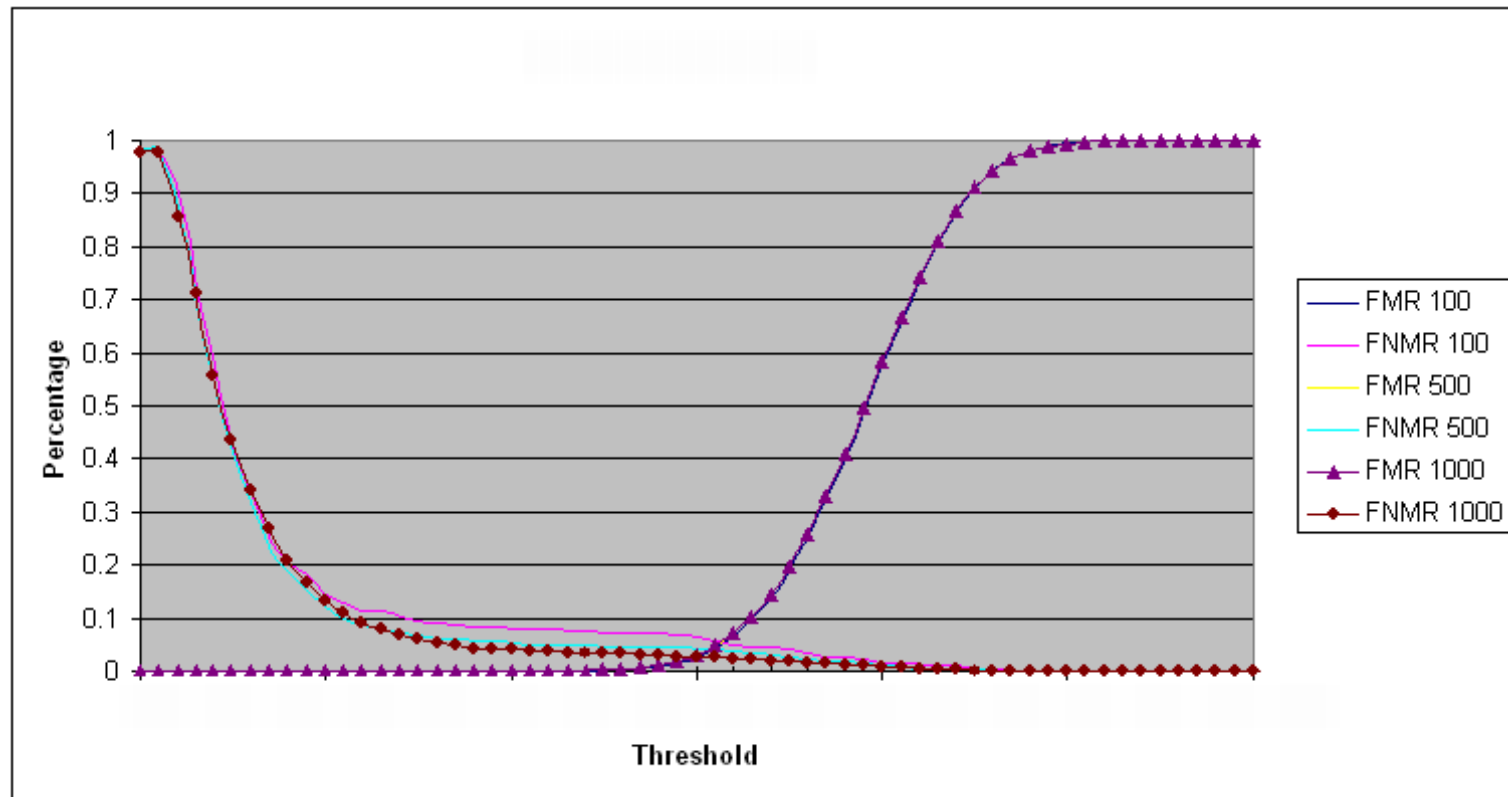
Order-1 analysis: FMR / FMNR curves

- 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: 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.

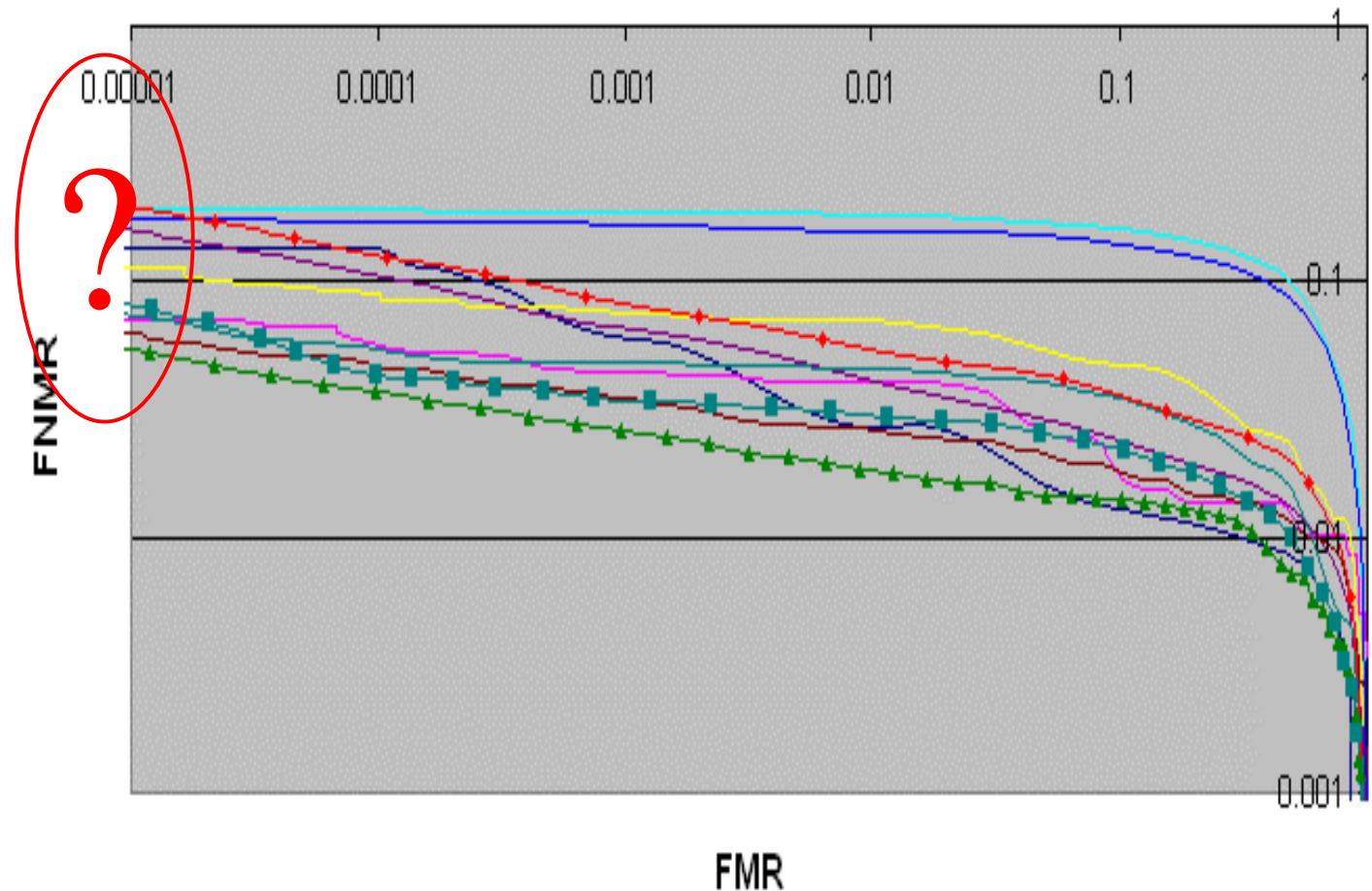


Order-1 analysis: DET curves



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

INCORECT →

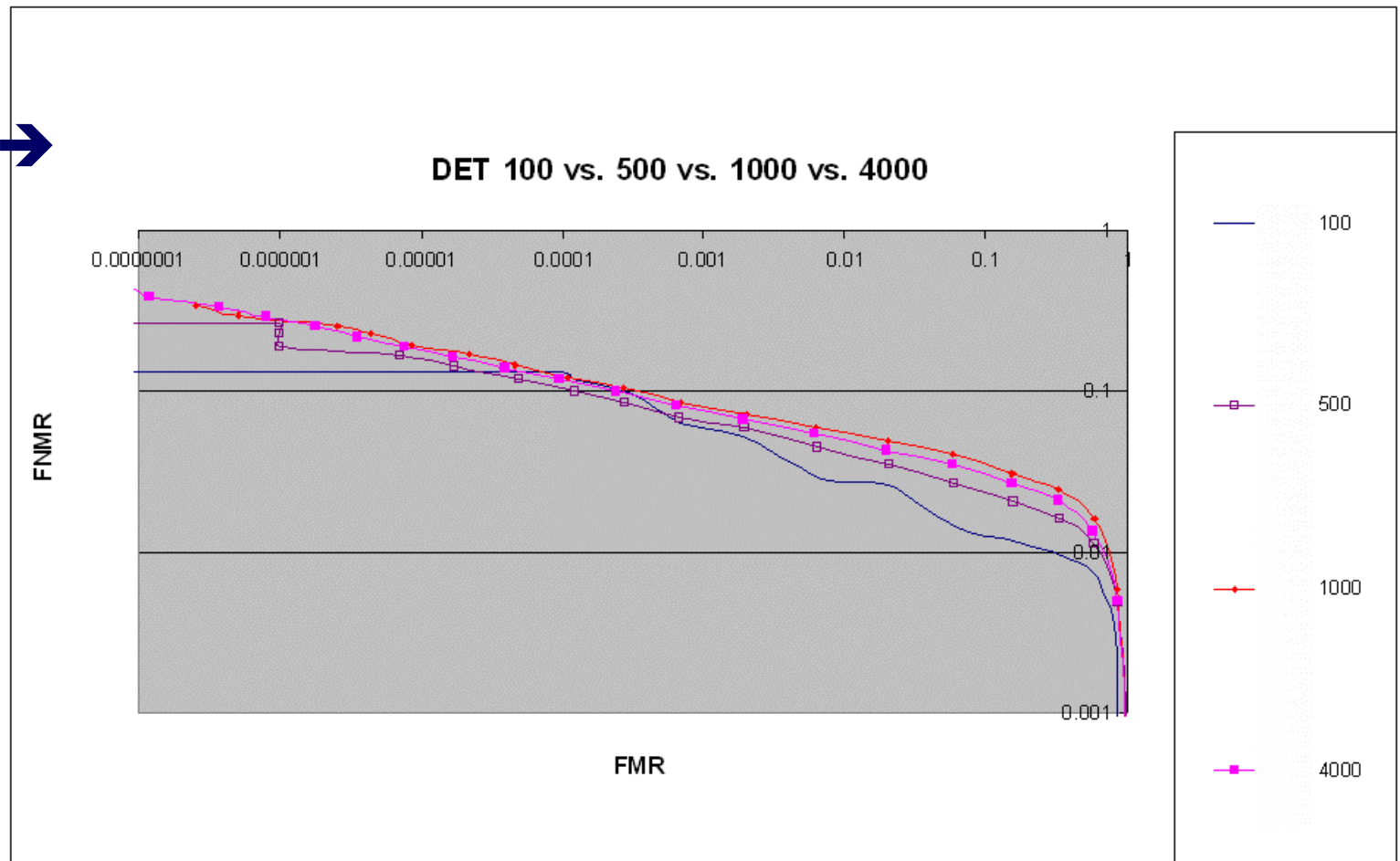


Order-1 analysis: DET curves

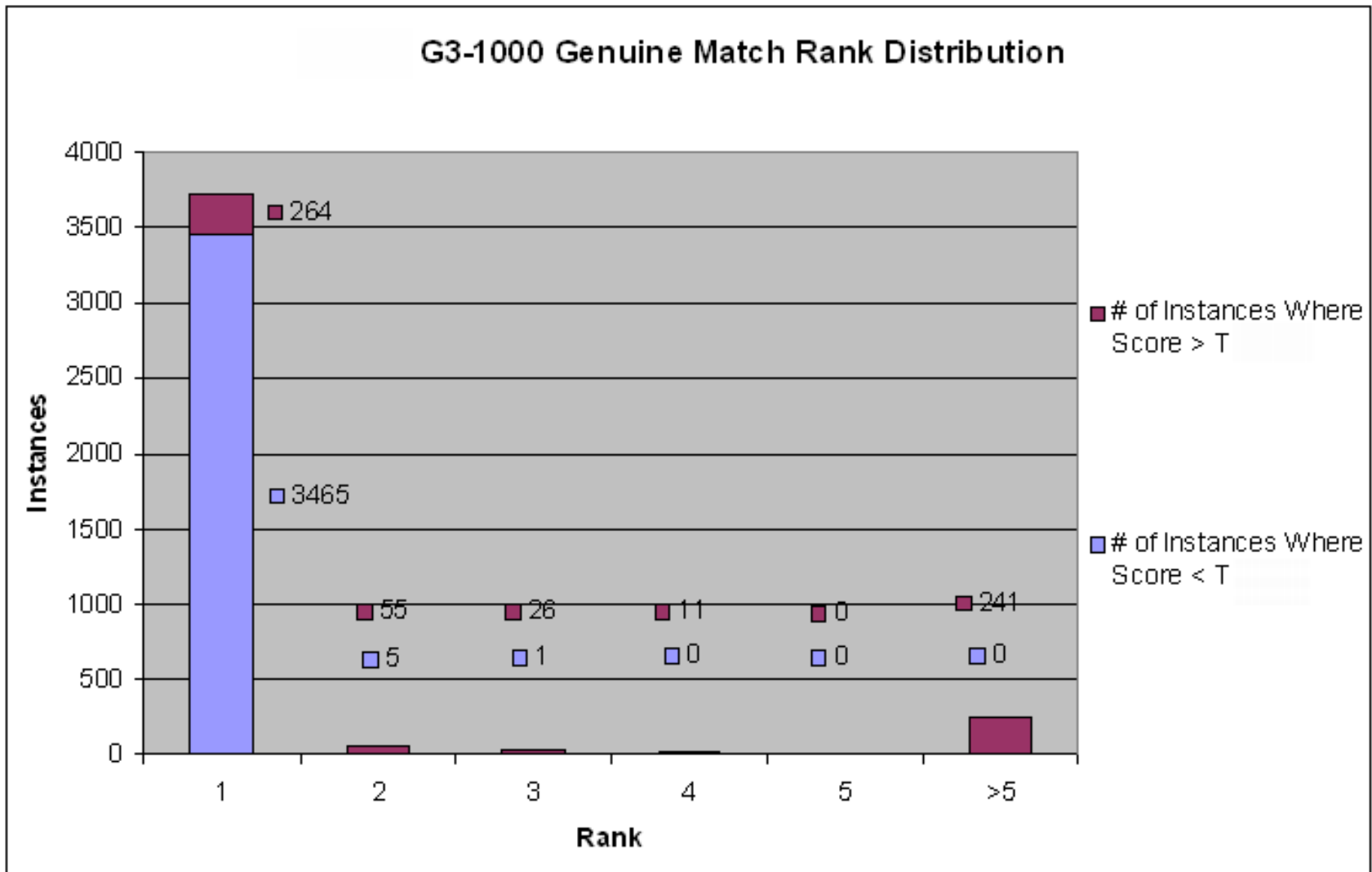


- Measured points must be shown, not only extrapolated lines!
 - Especially in the area of prime interest

CORRECT →



Order 2. Do Genuine data have best scores ?



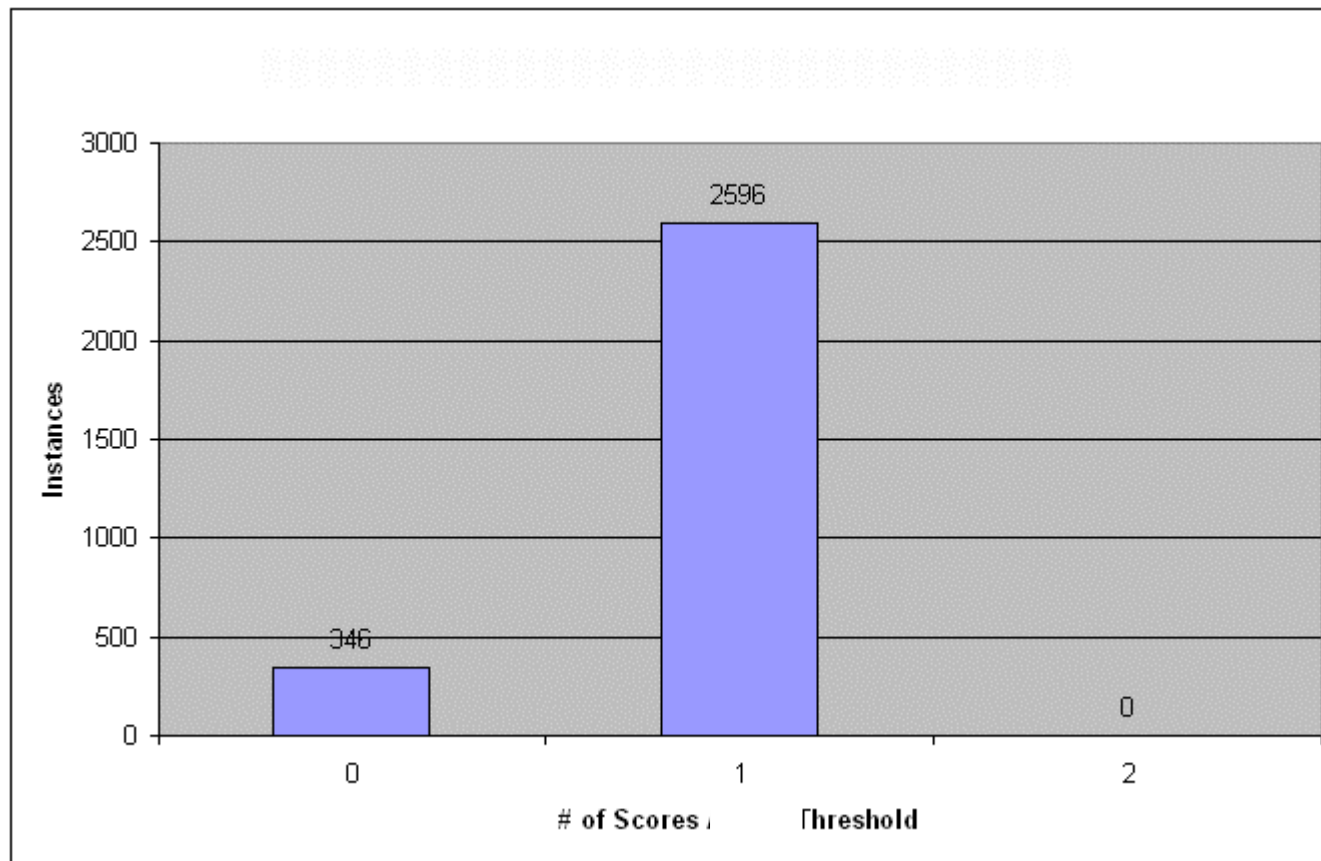
Order 3: Recognition confidence I



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

Is this a good systems ?

Hits=2596, Misses=346

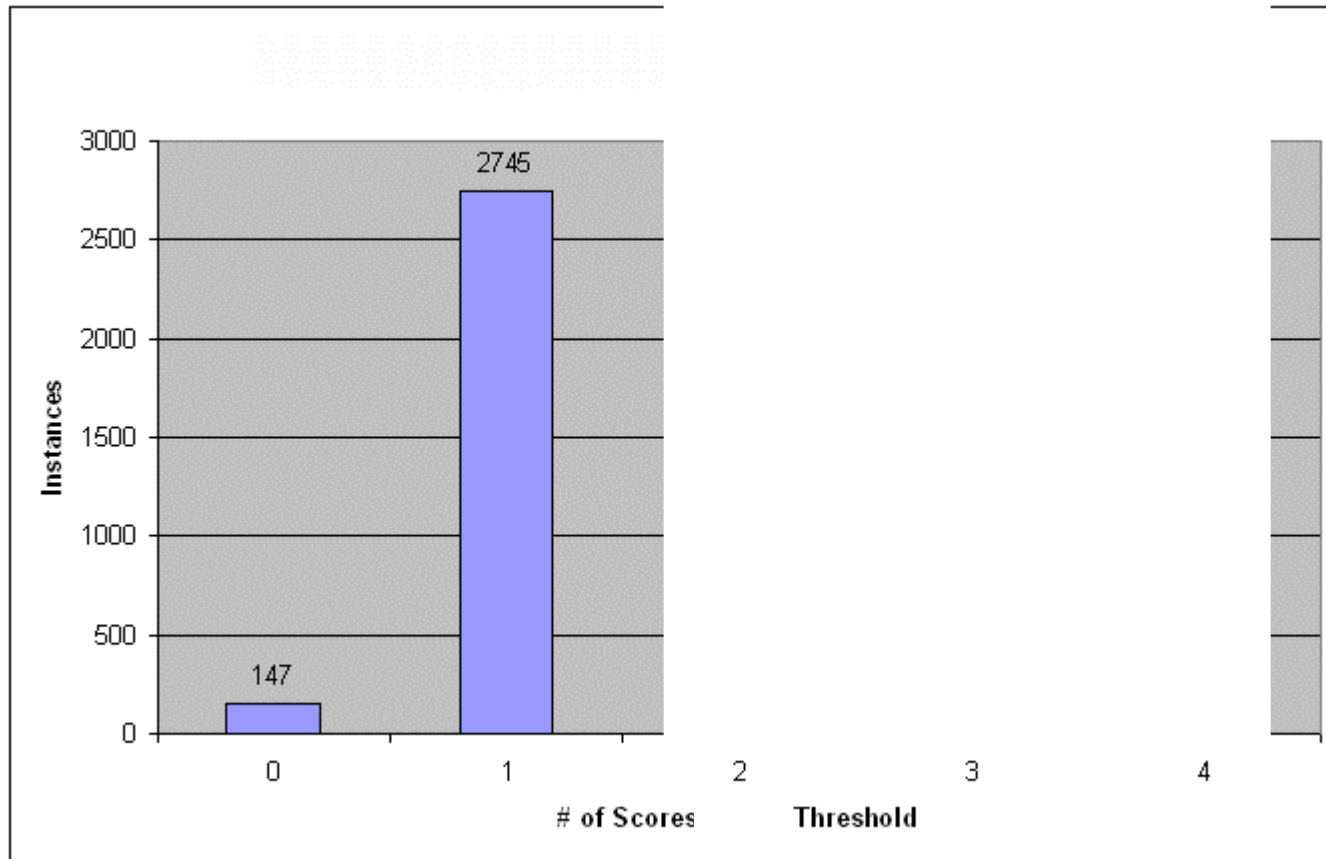


Order 3: Recognition confidence I (cnt)

- Number of scores below a threshold

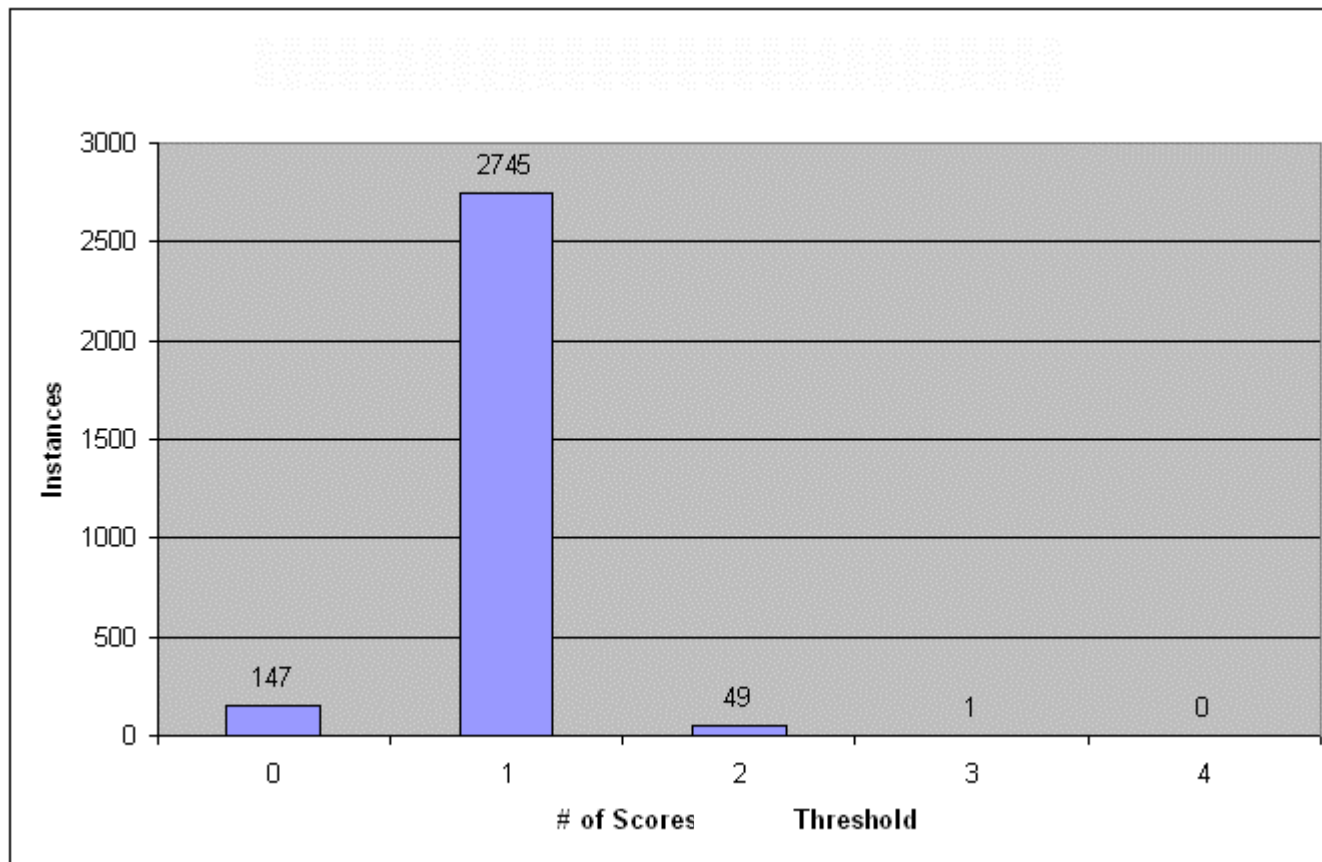
Is this one better ?

Hits=2745, Misses=147



Order 3: Recognition confidence I (cnt)

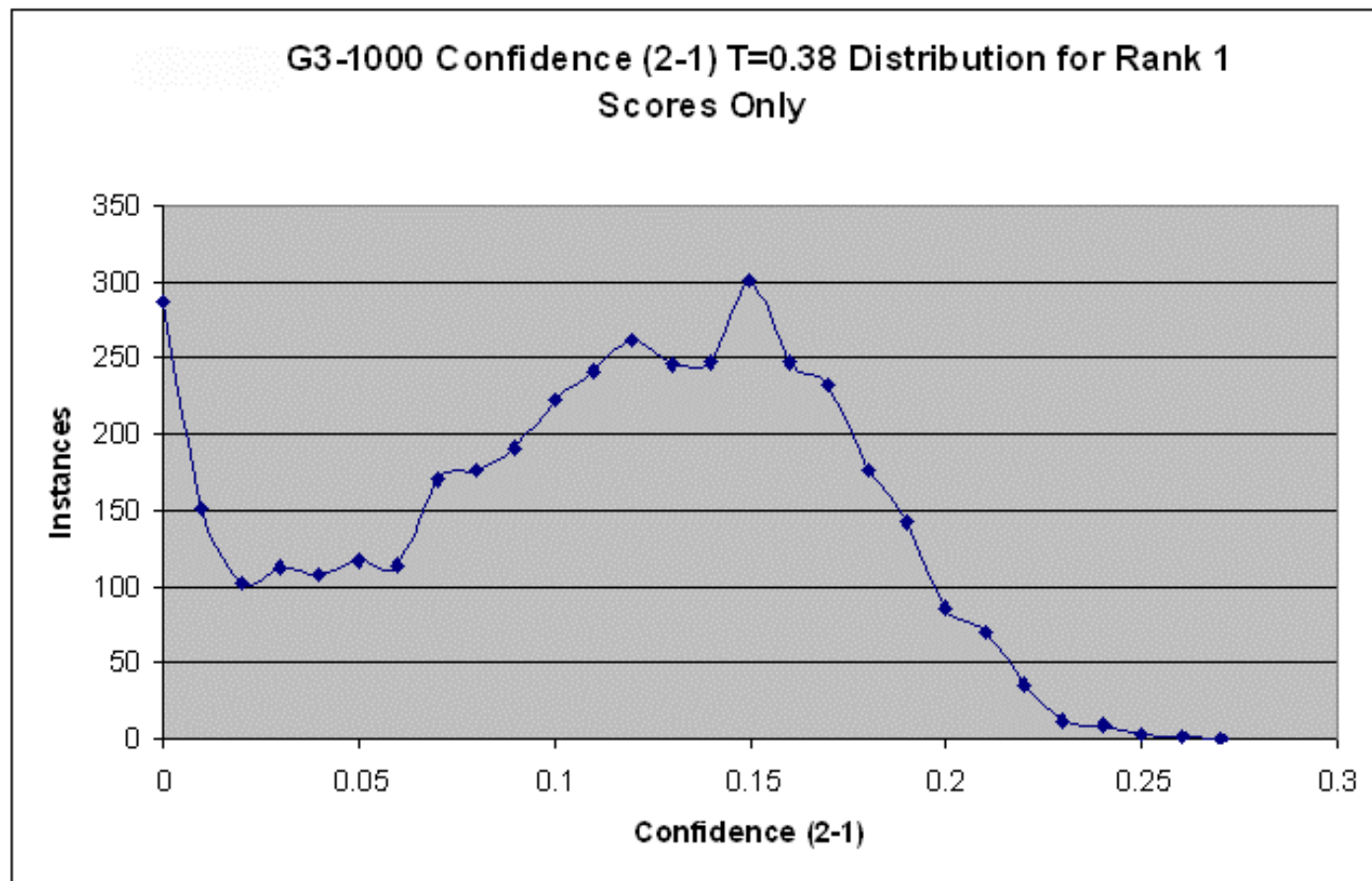
Many can improve the match/non-match tradeoff at the cost of allowing more than one scores below a threshold. (by raising the threshold) - Will you deploy it for Access Control ?!



Order 3: Recognition confidence II



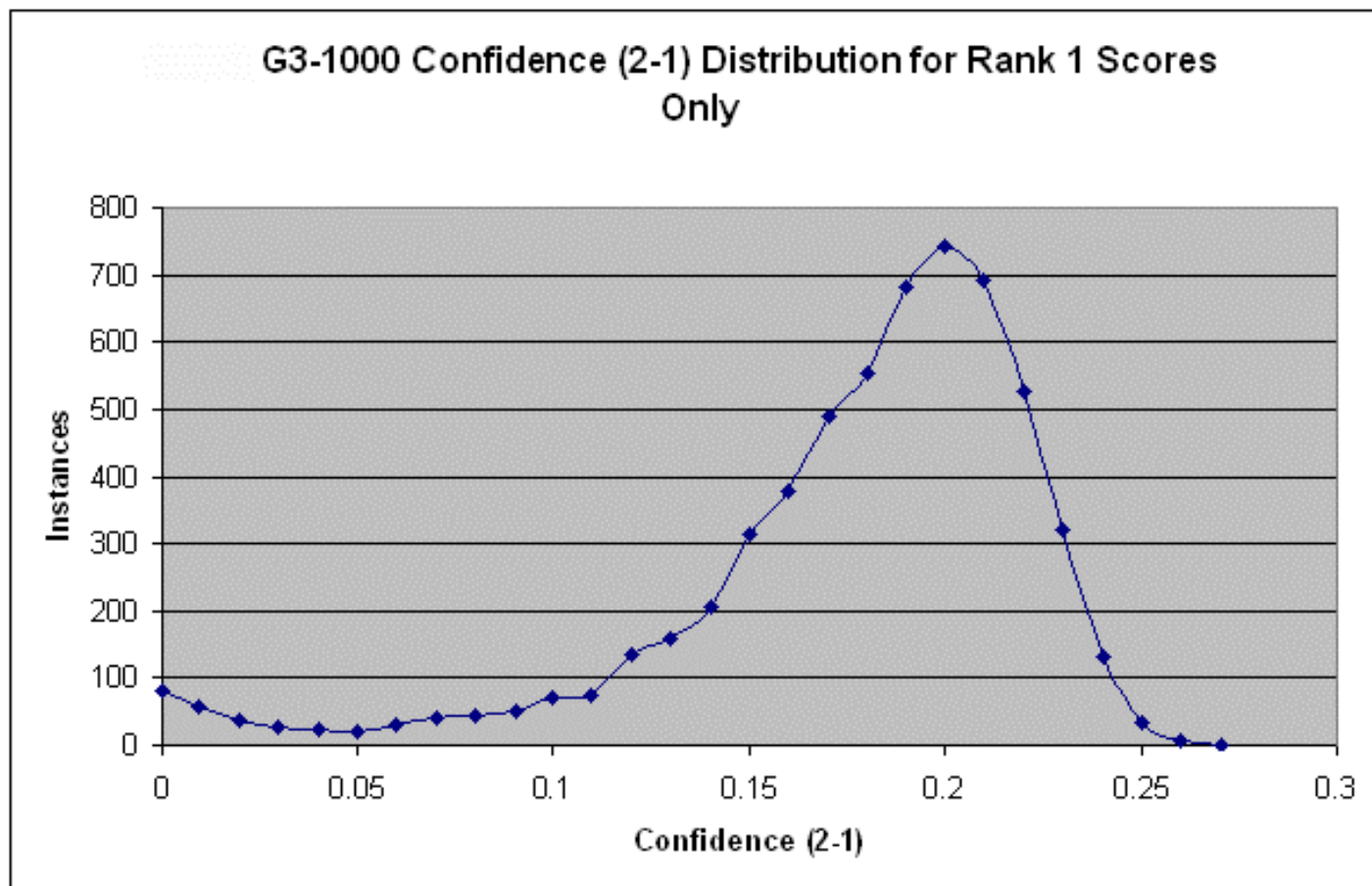
- Distance from “runner-up” and “winning” scores –



Order 3: Recognition confidence II



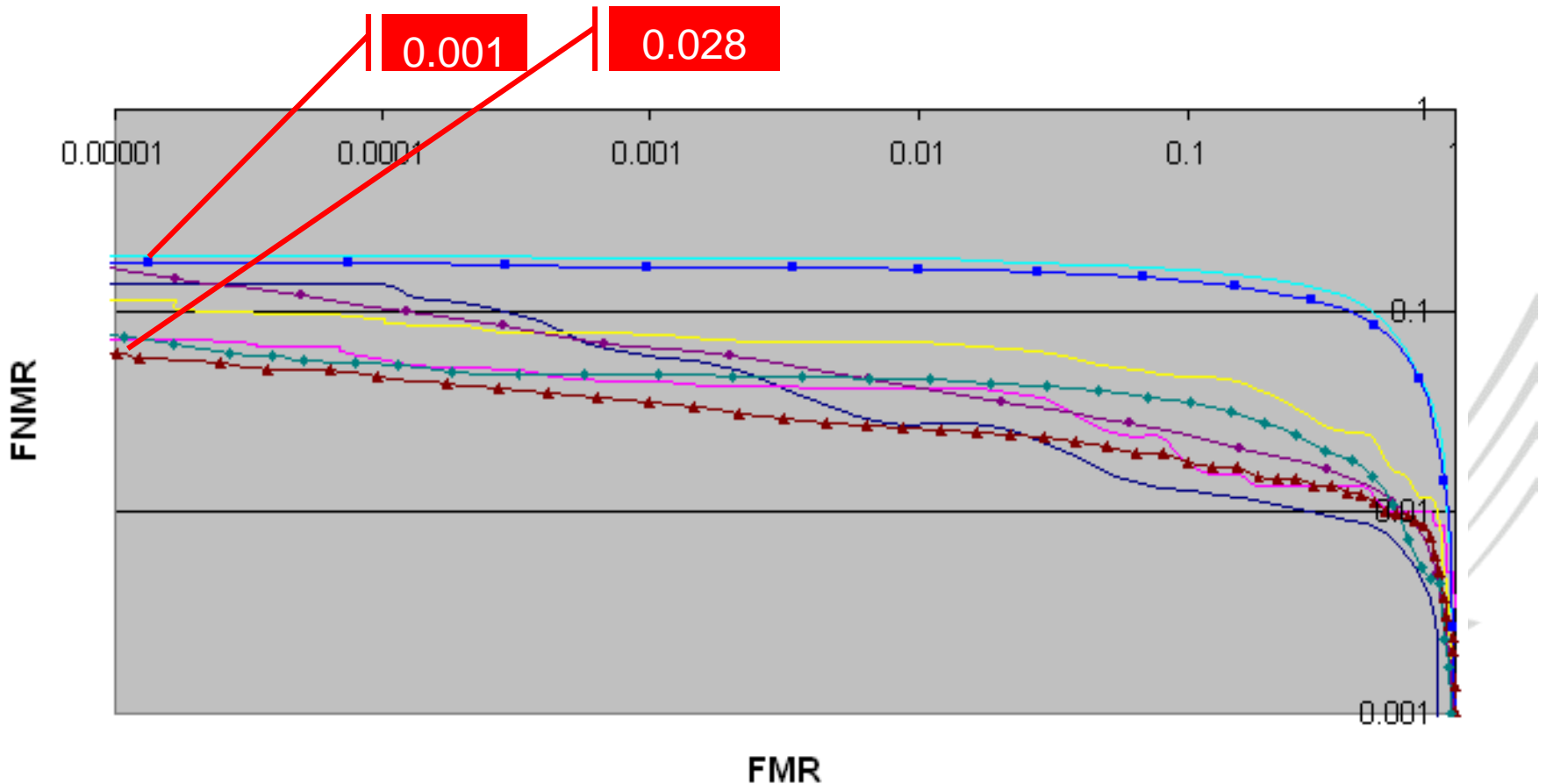
- Distance from “runner-up” and “winning” scores – Which do you prefer?



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



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.

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



- 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.
- 

Main lesson:



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 performance evaluation that

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

Especially because ...



- There are many biometric error types (eg. FMR, FTA, FCR...)
- There are many factors that affect the performance (lighting, location ...)
- Performance may deteriorate over time (as number of stored people increases and spoofing techniques become more sophisticated).

While ...

- There are also many ways to improve the performance (eg. more samples, modalities, constraints, ...)

References



- ISO/IEC 19795: Biometric performance testing and reporting.
- D. Gorodnichy. “Face databases and evaluation”, in **Encyclopedia of Biometrics** (Editor: Stan Li), Elsevier Publisher, 2009.
- 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
- D. Gorodnichy. **Multi-order analysis framework for comprehensive biometric performance evaluation**, SPIE Conference on Defense, Security, and Sensing. Orlando, 5 - 9 April 2010,
- **C-BET (Comprehensive Biometrics Evaluation Toolkit):**
 - developed by Canada Border Services Agency for GoC
 - for selecting new and tuning existing biometric systems.