
Biometric Systems

Lesson 2 - Performance



Maria De Marsico
demarsico@di.uniroma1.it



SAPIENZA
UNIVERSITÀ DI ROMA



*Dipartimento di
Informatica*



All that glitters... is not gold ...



www.HelloCrazy.com



Problems: possible wide intra-class variations



Problems: possible very small inter-class variations



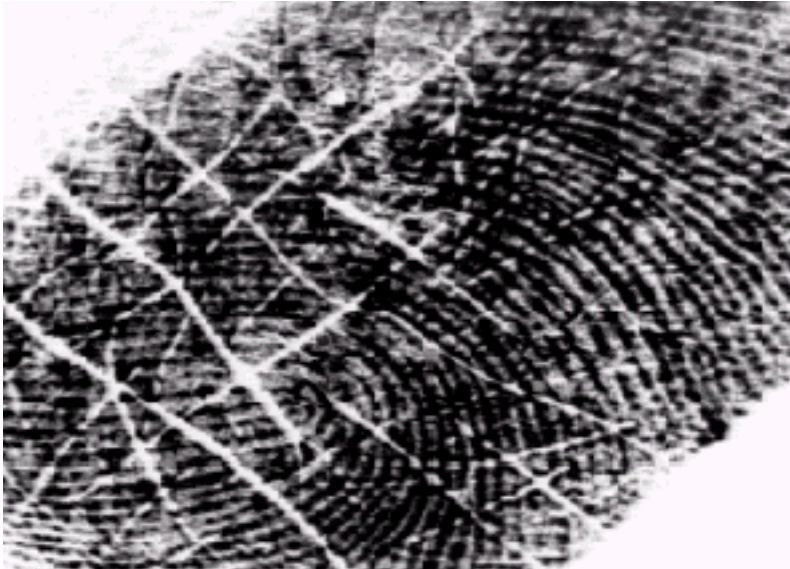
Twins



Father and son



Problems: noisy and/or distorted acquisitions



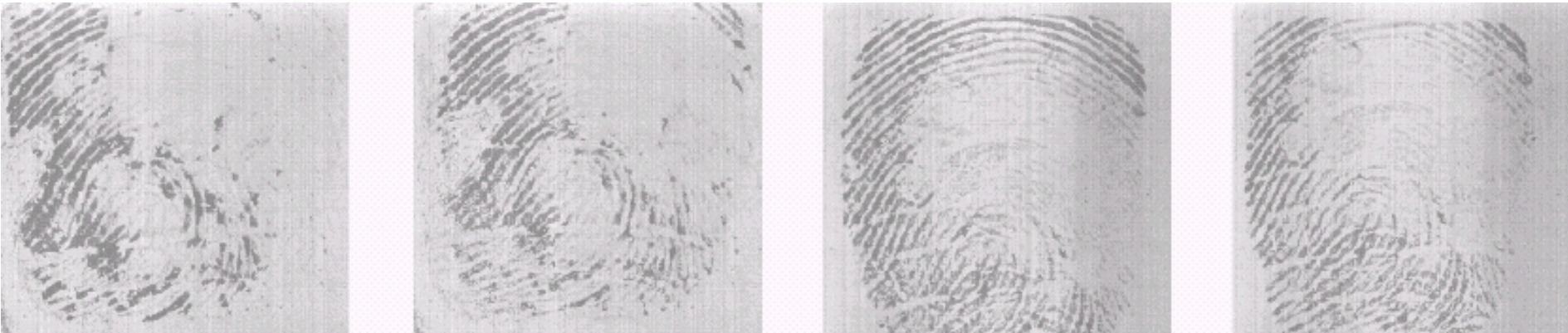
**Poor quality fingerprints
(eg. heavy worker)**

Non uniform lighting





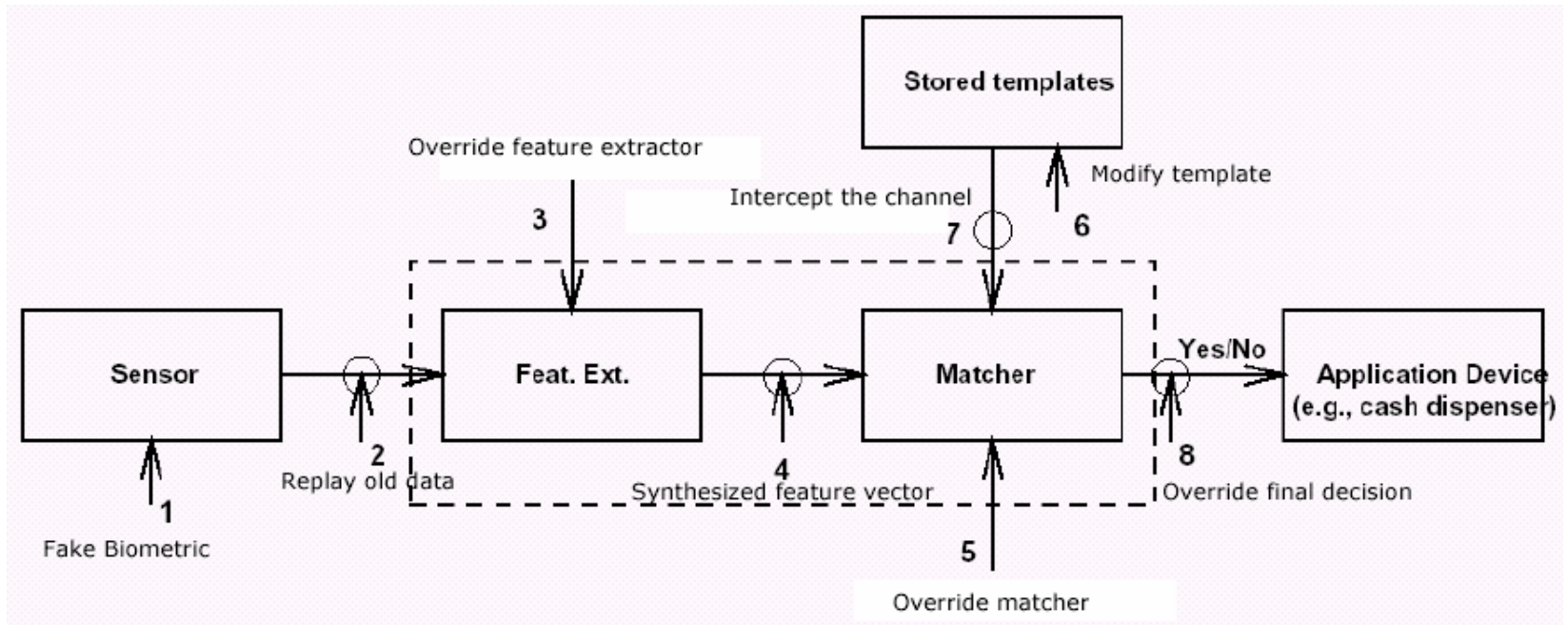
Problems: non universality



4% of population presents poor quality fingerprints
In some groups it is a particularly widespread characteristic (eg. elderly people)



Problems: possible attacks (spoofing) in different moments





What to compare?

- **Sample** = the raw captured data: an image, a voice recording, a fingerprint, etc.
- **Hand-crafted features** = features that are manually engineered by the data scientist and extracted from samples
- **Template** = collection of features extracted from the raw data:
 - a histogram representing the frequencies of relevant values in the image, e.g., greylevel values
 - a vector of values each representing a relevant measure, e.g., Bertillon measures
 - a time series of acceleration values (actually 3, one for each accelerometer axis)
 - a set of triplets as for relevant fingerprint points $\{(x_1, y_1, \theta_1), \dots (x_n, y_n, \theta_n)\}$ representing the coordinates of the points and the direction of the tangent to the ridge in that point



How to compare templates

- In many cases templates are vectors, therefore Euclidean distance

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

or cosine similarity may provide either a distance measure or a similarity measure

Given two n -dimensional **vectors** of attributes, **A** and **B**, the cosine similarity, $\cos(\theta)$, is represented using a **dot product** and **magnitude** as

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}},$$

where A_i and B_i are the i th **components** of vectors **A** and **B**, respectively.



How to compare templates

- (Pearson) Correlation (a similarity measure) can be used for histograms or sets of points

Given paired data $\{(x_1, y_1), \dots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (the sample mean); and analogously for \bar{y} .

Rearranging gives us this formula for r_{xy} :

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

where n, x_i, y_i are defined as above.



How to compare templates

- For histograms, other kinds of comparison, e.g., Bhattacharyya distance, can be used

For **probability distributions** P and Q on the same **domain** \mathcal{X} , the Bhattacharyya distance is defined as

$$D_B(P, Q) = -\ln(BC(P, Q))$$

where

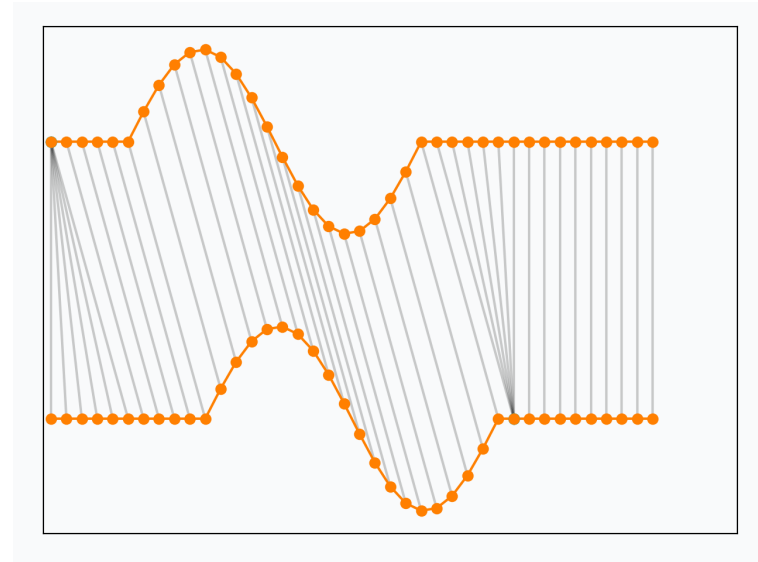
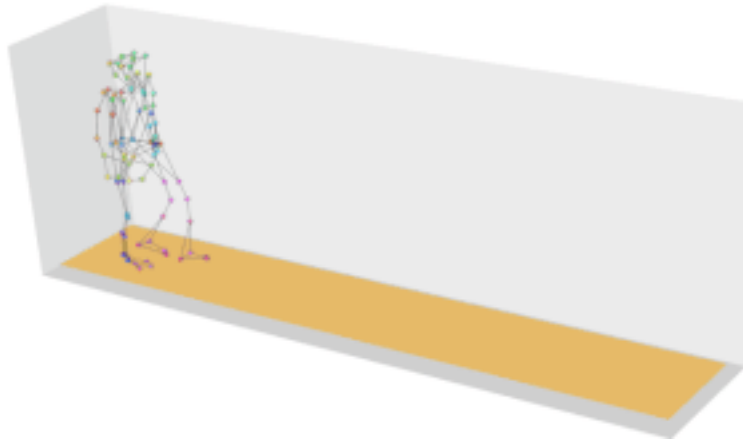
$$BC(P, Q) = \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)}$$

is the Bhattacharyya coefficient for **discrete probability distributions**.



How to compare templates

- Dynamic Time Warping (distance) for time series

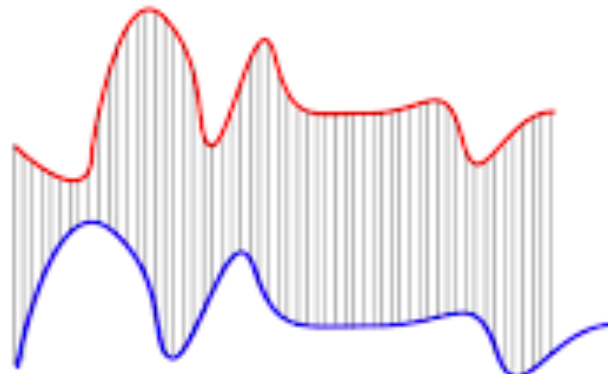


Two repetitions of a walking sequence recorded using a motion-capture system. While there are differences in walking speed between repetitions, the spatial paths of limbs remain highly similar (from Wikipedia).

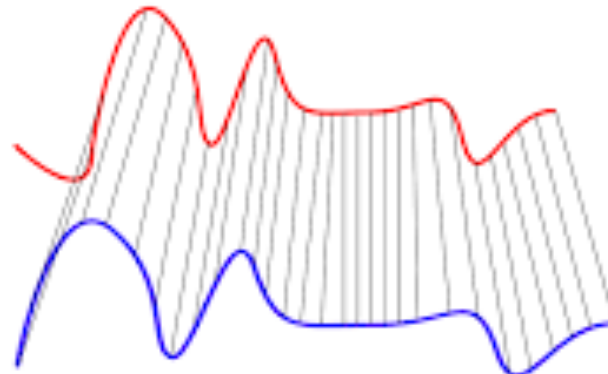


How to compare templates

- Dynamic Time Warping (distance) for time series



Euclidean Matching

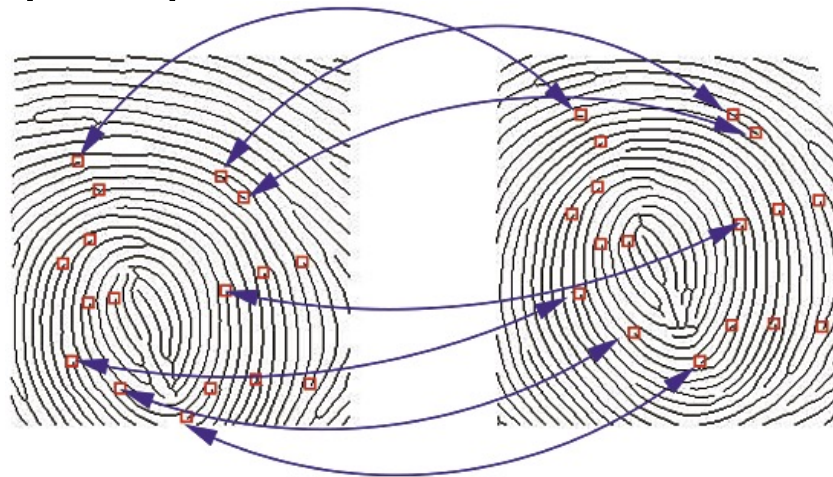


Dynamic Time Warping Matching



How to compare templates

- How to compare the results of submitting a template to a Deep Learning model (**learned features**)?
 - Delete the final classification layer (usually a softmax layer) in order to get the *embeddings* that the architecture would use for the final classification.
 - The embeddings can be compared as they were vectors of hand-crafted features
- Other features may require a more complex comparison process, e.g., for fingerprint points we must also find the best pairing





What after?

- Once a similarity or distance has been obtained, it is compared with an acceptance threshold if in verification or identification open set
- The evaluation of performance analyses the behaviour of the system (system errors) for different thresholds
- Reasoning about similarities or distances is perfectly symmetrical



In summary






- Reminder:
 - Select and extract *good* (discriminative enough) features
 - Devise a reasonable matching strategy
 - Analyse the behaviour based on different acceptance thresholds (similarity must be equal or higher than a similarity threshold, distance must be lower or equal to a distance threshold)



Possible errors: verification

A subject is accepted if the similarity (or score) achieved from matching with the gallery template(s) corresponding to the claimed identity is greater than or equal to the acceptance threshold (or, if the distance with such gallery template(s) is less than or equal to the acceptance threshold). Otherwise it is rejected.

We can identify 4 possible cases:

- The claimed identity is true and the subject is accepted (Genuine Acceptance – GA, also indicated as Genuine Match - GM) 
- **The claimed identity is true but the subject is rejected (False Rejection – FR, also indicated as False Non Match – FNM, or type I error)** 
- An impostor subject is rejected (Genuine Reject – GR, also indicated as Genuine Non Match - GNM) 
- **An impostor subject is accepted (False Acceptance – FA, also indicated as False Match – FM, or type II error)** 
- It would be nice to have something in the middle to leave the choice to a human in uncertain cases ... we will see ... 



Possible errors: verification ... how to compare systems

- In the last years the research community has proposed a great number of systems addressing biometric recognition.
- It is necessary to **measure** and **compare** performance.
- Simple **count** of errors is **not** suited.

False Acceptance Rate - FAR (False Match Rate - FMR)

The FAR is defined as the percentage of recognition operations with an impostor claim in which false acceptance occurs. This can be expressed as a probability. For example, if the **FAR is 0.1 percent**, it means that on the average, **one out of every 1000 impostors** attempting to breach the system will be successful. Stated another way, it means that the probability of an unauthorized person being identified as an authorized person is 0.1 percent.

False Rejection Rate - FRR (False Non Match Rate - FNMR)

The FRR is defined as the percentage of recognition operations with a genuine claim in which false rejection occurs. This can be expressed as a probability. For example, if the **FRR is 0.05 percent**, it means that on the average, **one out of every 2000 authorized persons** attempting to access the system will **not be recognized** by that system.



Possible errors: verification ... how to compare systems

Most common measures for **verification**:



- **FAR**: False Acceptance Rate
 - **FRR**: False Rejection Rate
 - **ERR**: Equal Error Rate
 - **DET**: Detection Error Trade-off
 - **ROC**: Receiving Operating Curve
- All such measures depend on the adopted **acceptance threshold**
 - **All the data (samples) used for the evaluation experiments is labeled with the correct identity (ground truth - this is not true during real world operations!)**
 - **Let us assume to have a ground truth function $\text{id}(\text{template})$ that, given a template, returns its true identity, for instance:**
 - **$\text{id}(p_j)$ is the true identity associated with the j -th probe**
 - **$\text{id}(g_x)$ is the true identity associated with template x in the gallery**
 - **i is the identity claimed by a probe p_j**
 - **$\text{topMatch}(p_j, \text{identity})$ returns the best match between p_j and the (possibly more than one) templates associated to the claimed identity in the gallery**
 - **$s(t1, t2)$ returns the similarity between template $t1$ and template $t2$**



i

Possible errors: verification ... how to compare systems

Error measures:

- P_G = set of probes belonging to subjects in the gallery (**genuine claims can only come from here**)
- P_N = set of probes belonging to subjects not in the gallery (but in the dataset, so id function works)
- The sets of subjects in the gallery/not in the gallery is decided during experiment set up but all samples/templates are labeled in any case**

$$FRR(t) = \frac{|\{p_j : s_{xj} \leq t, id(g_x) = id(p_j)\}|}{|\{p_j : id(g_x) = id(p_j)\}|}$$

$$g_x = \text{topMatch}(p_j, id(p_j))$$

$$s_{xj} = s(g_x, p_j)$$

$$\forall p_j \in P_G$$

- When the user submitting the probe sample (template) p_j declares the true identity, then such identity will be the same returned by the ground truth function for p_j that is $id(p_j)$
- We use $id(p_j)$ instead of a generic i to underline that the claim is genuine**

$$FAR(t) = \frac{|\{p_j : s_{xj} \geq t \wedge id(g_x) \neq id(p_j)\}|}{|\{p_j : id(g_x) \neq id(p_j)\}|}$$

$$g_x = \text{topMatch}(p_j, i)$$

$$s_{xj} = s(g_x, p_j)$$

$$\text{Scenario 1: } \forall p_j \in P_G \cup P_N \quad \forall i \in I$$

$$\text{Scenario 2: } \forall p_j \in P_N \quad \forall i \in I'$$

- When the user submitting the probe sample (template) p_j declares a false identity, then such identity will be a generic identity i which is different from that returned by the ground truth function for p_j that is $id(p_j)$
- Two scenarios differ only for the fact that the impostor can either ALSO belong to the gallery (a registered subject, whose probe template is therefore in P_G , but declaring another identity) or not (a subject who is not even registered, whose probe template is in P_N).
- The difference is not important for the computation, since in both cases we have a false claim.



Possible errors: verification ... how to compare systems

Hypothesis:

H_0 : different person

H_1 : same person

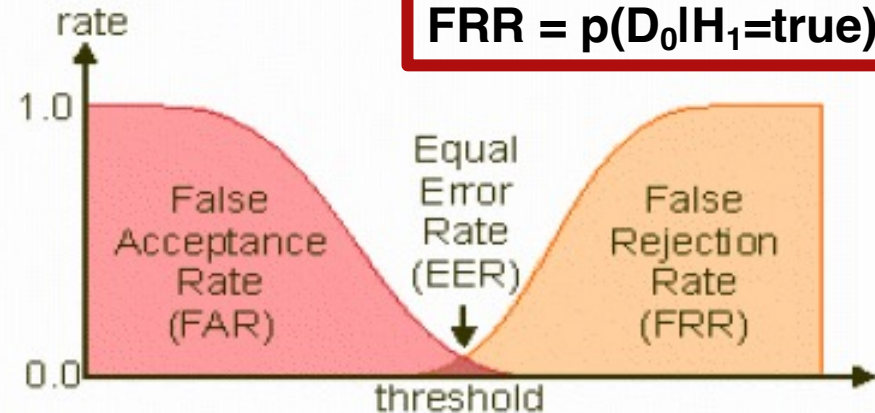
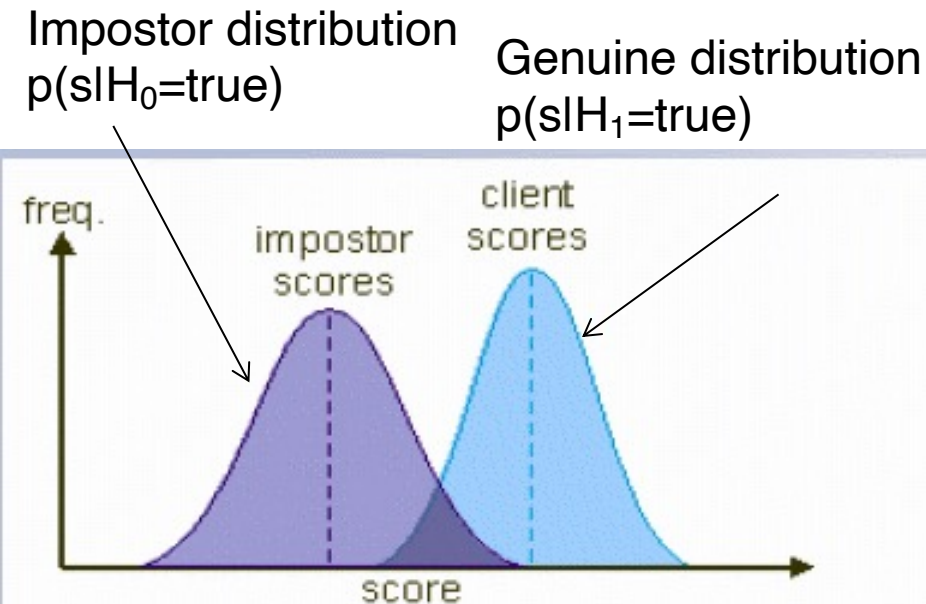
Possible decisions:

D_0 : different person

D_1 : same person

$$\text{FAR} = p(D_1 | H_0 = \text{true})$$

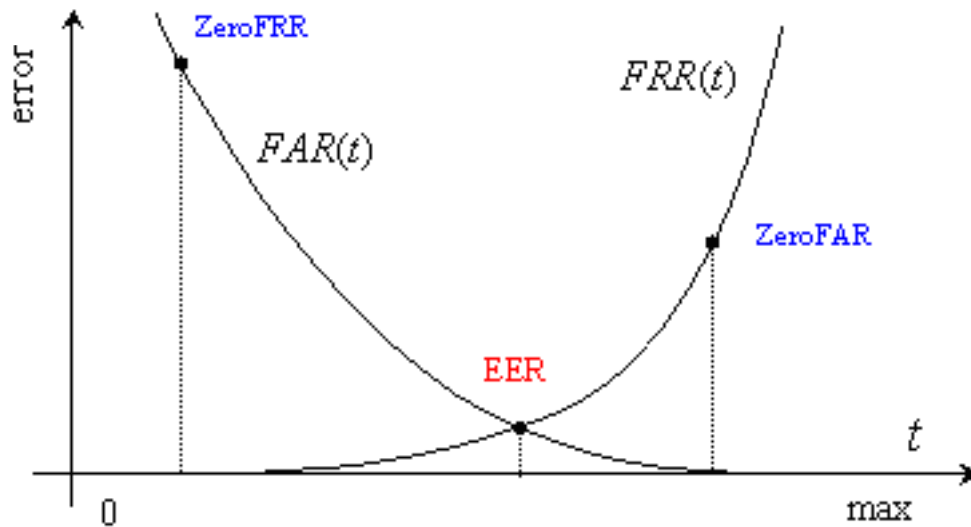
$$\text{FRR} = p(D_0 | H_1 = \text{true})$$



A score is said *genuine* (authentic) if it results from matching two samples of the biometric trait of a same enrolled individual; it is said *impostor* if it results from matching the sample of a non-enrolled individual.



Possible errors: verification ... how to compare systems



Genuine Acceptance Rate (GAR) or Genuine Match Rate (GMR)

1-FRR (1-FNMR)

- **Equal Error Rate (EER)**

Error Rate when $FAR=FRR$ ($FMR = FNMR$).

- **ZeroFRR (ZeroFNMR)**

FAR when $FRR = 0$ (FMR when $FNMR = 0$)

- **Zero FAR (Zero FMR)**

FRR when $FAR = 0$ ($FNMR$ quando $FMR = 0$)



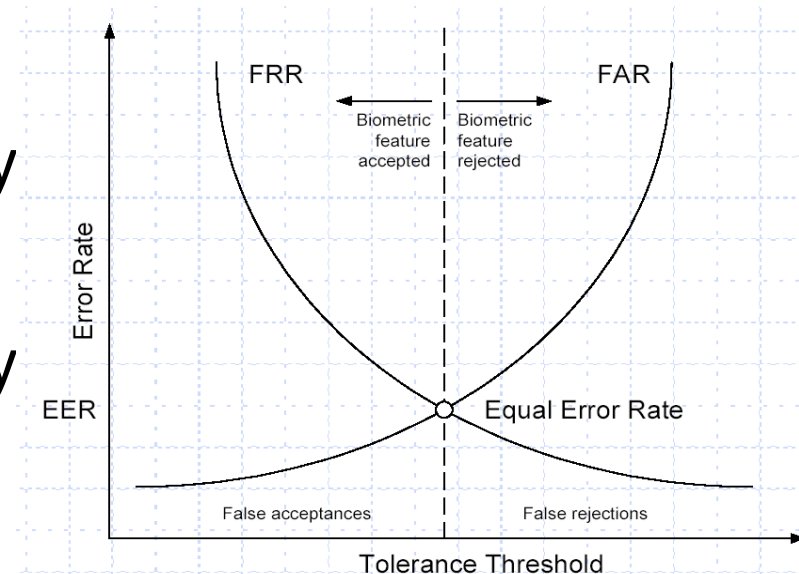
Possible errors: verification ... how to compare systems

- **Acceptance threshold** is crucial and depends from the application needs

Beware!
Here we have distances!

Considering distances:

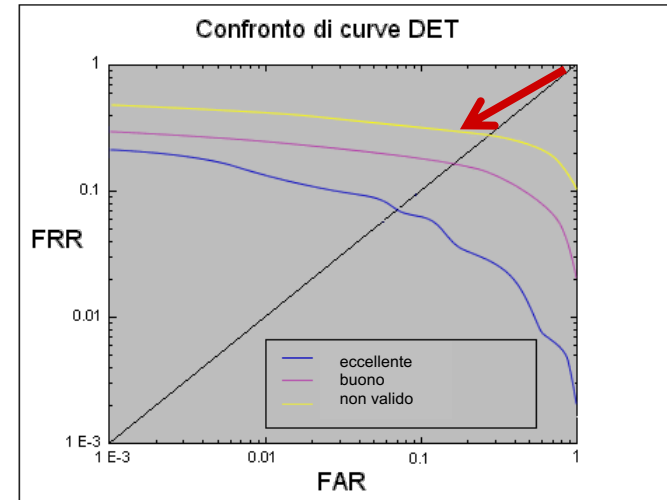
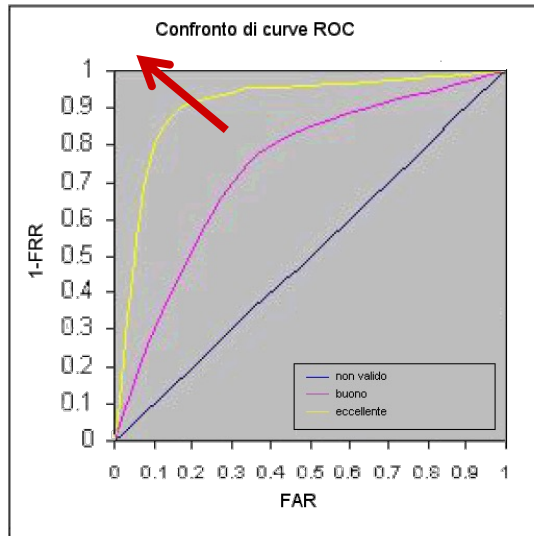
- A too low threshold causes many type I errors – rejected genuine **FRR**
- A too low threshold causes many type II errors – accepted impostors - **FAR**
- Popular choice is the threshold corresponding to **ERR**: **FAR = FRR**





Possible errors: verification ... how to compare systems

- **ROC** (Receiver Operating Characteristic) – ROC depicts the probability of Genuine Accept (GAR) of the system, expressed as $1 - \text{FRR}$, vs False Accept Rate (FAR) variation.



- **DET** (Detection Error TradeOff) - DET depicts the probability of False Reject (FRR) of the system, vs False Accept Rate (FAR) variation. It is plotted in logarithmic form.



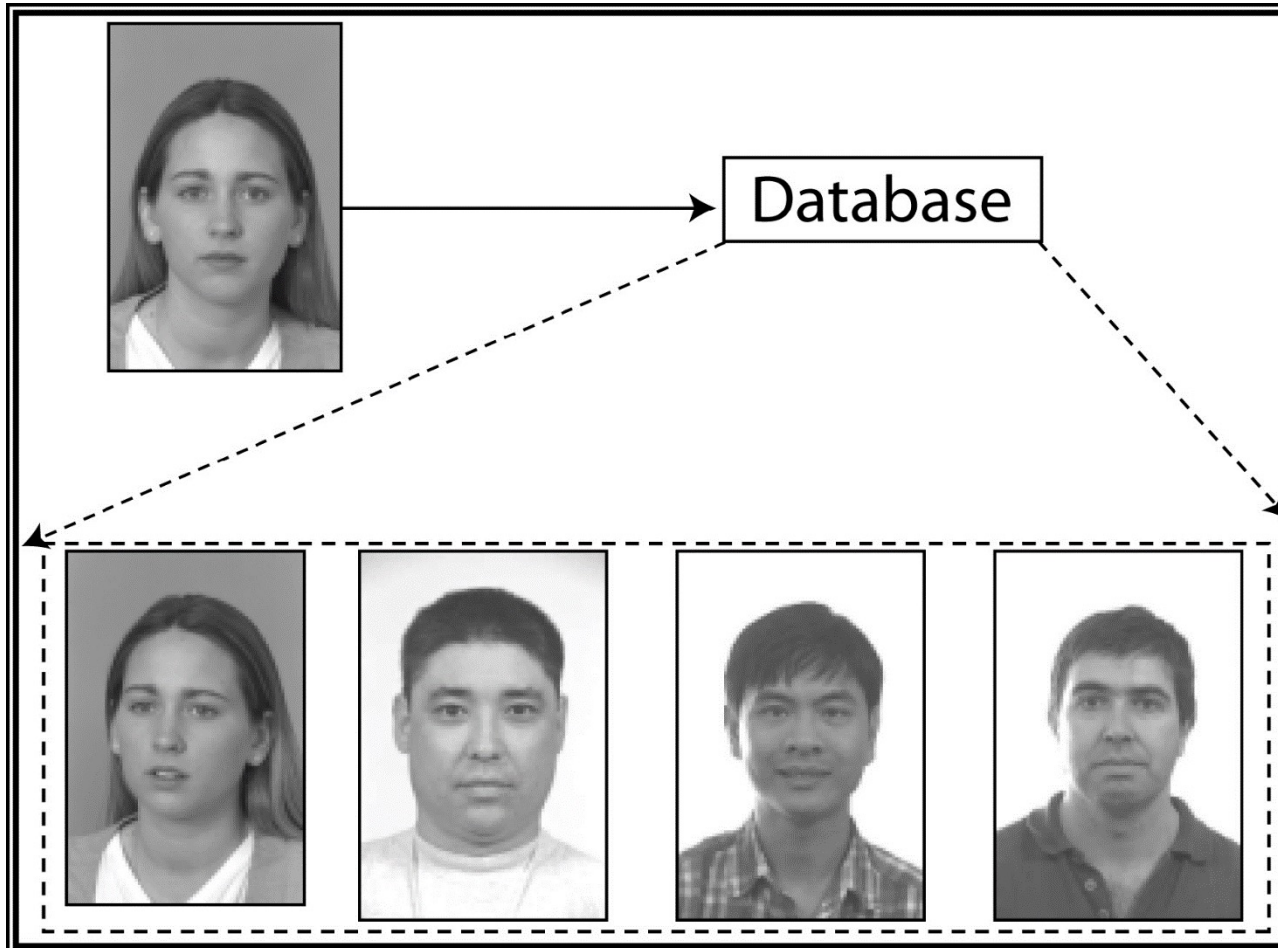
Possible errors: identification – open set

- In the **open set identification** task (e.g. **watchlist**) the biometric system determines if the individual's biometric signature matches a biometric signature of someone in the gallery.
- The individual **does not make** an identity claim.
- Examples:
 - comparing customers in an airport against a terror
 - comparing “John Doe” to a missing persons datab
- Two questions:
 - Is the probe subject in the database?
 - Who is the probe subject ?
- More possible error situations ... depending on the **matcher** and on the recognition **threshold** (score/similarity/distance)





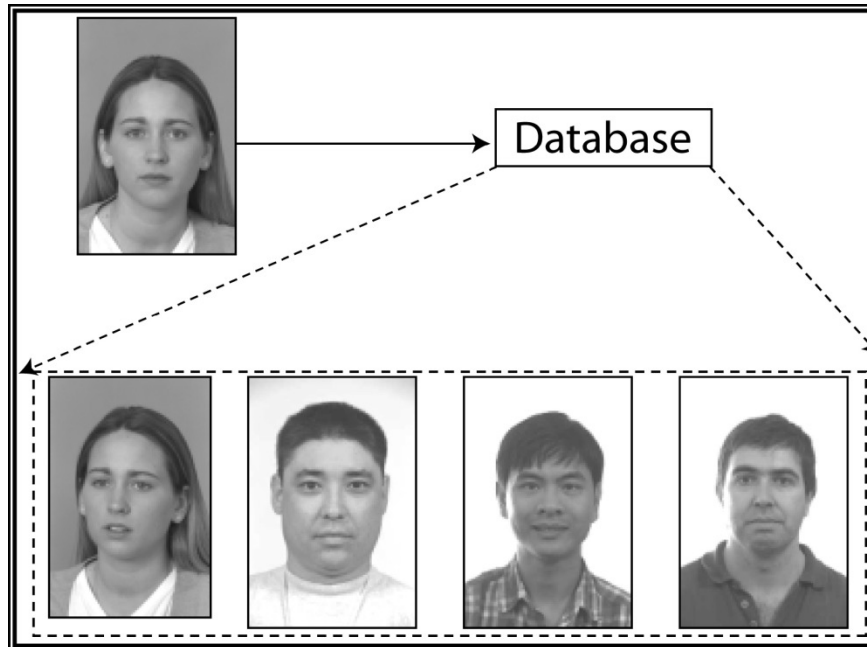
Possible errors: identification – open set



*Face images are from FERET database



Possible errors: identification – open set



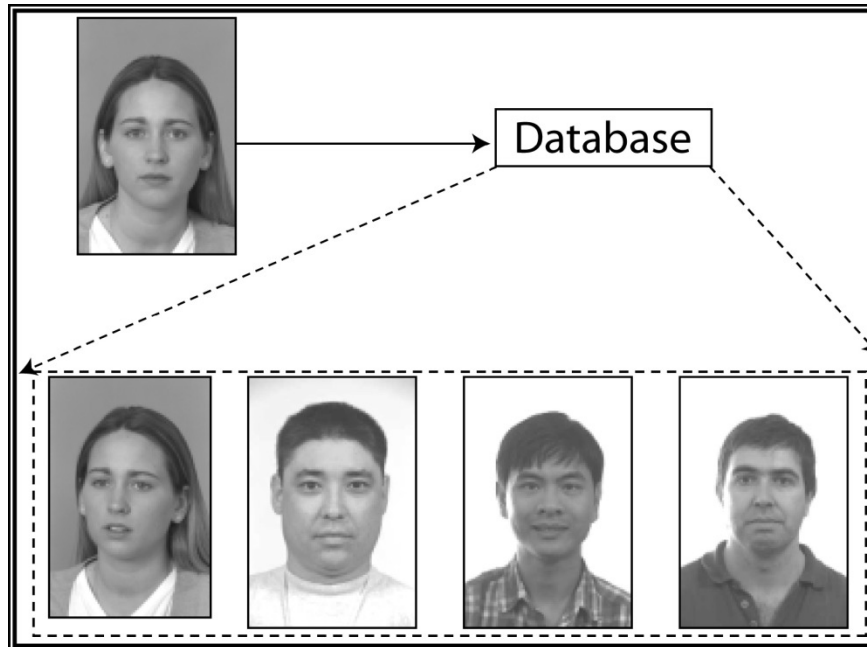
Scores = 0.9 0.86 0.6 0.4 threshold = 0.85

- Two individuals above the threshold = correct detect (alarm)
- The first individual is the right one = correct identification

→ correct detect and identify



Possible errors: identification – open set



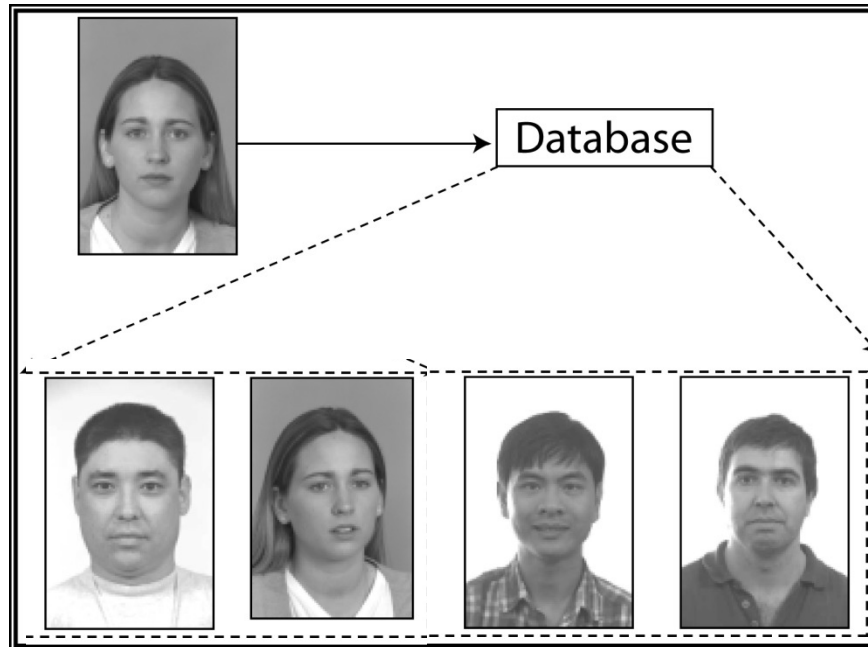
Scores = 0.9 0.86 0.6 0.4 threshold = 0.95

- No individual above the threshold = no detect (no alarm)
→ We do not care about looking at the top individual = no correct identification

→ no correct detect and identify



Possible errors: identification – open set



Scores = 0.86

0.8

0.6

0.4

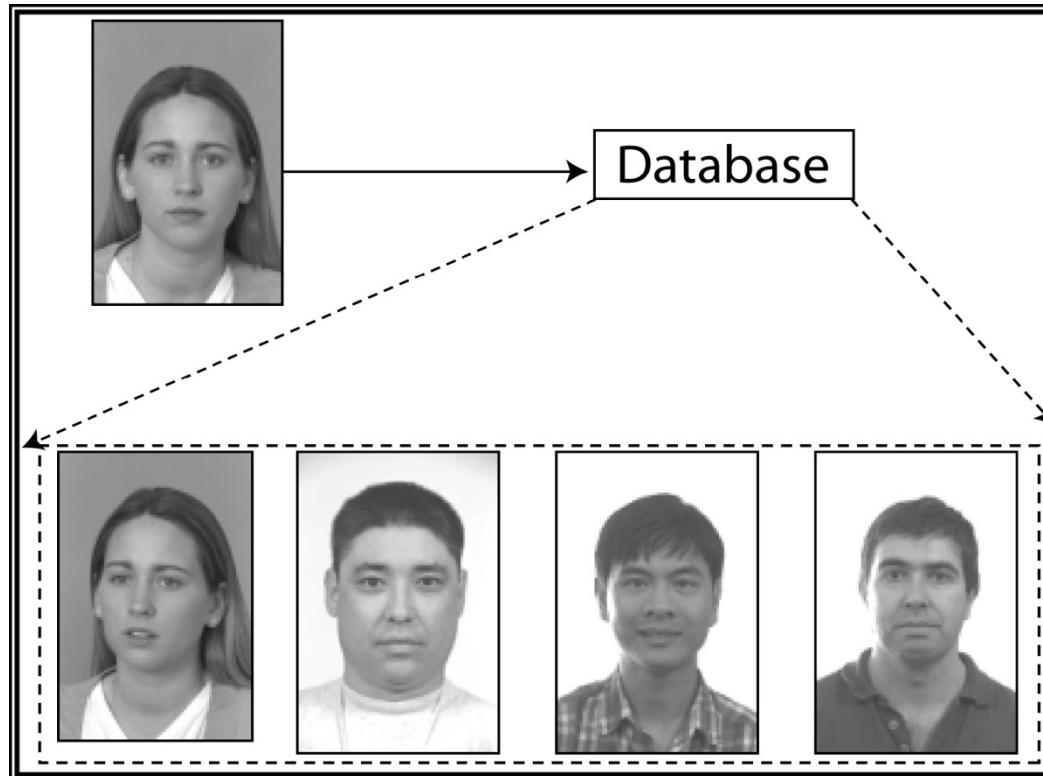
threshold = 0.75

- Two individuals above the threshold = correct detect (alarm)
- The first individual is not the right one = no correct identification

→ no correct detect and identify



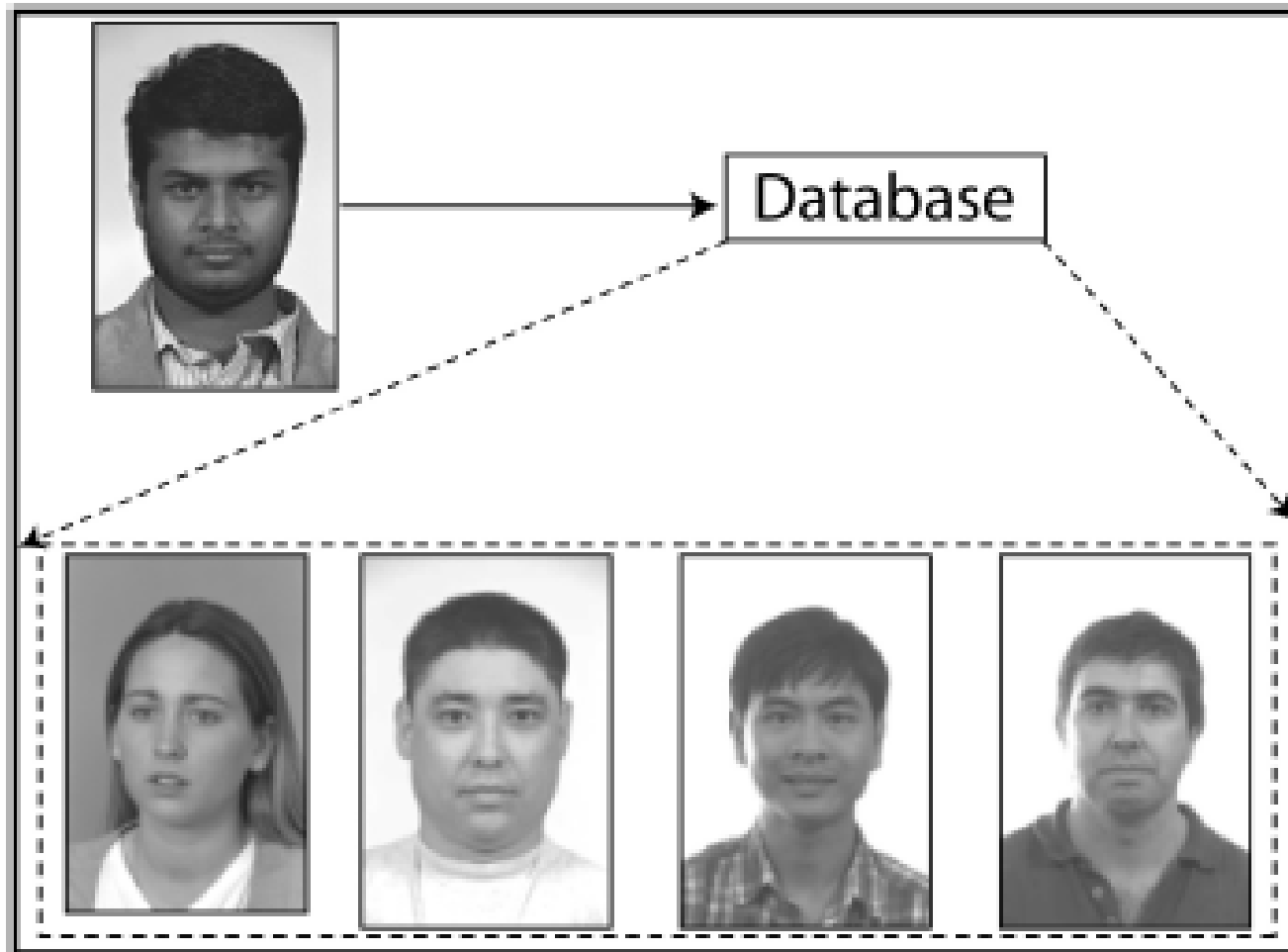
Possible errors: identification – open set



- If we run many trials with probes belonging to the subjects in the database (set P_G), we will know how often the system will return a correct result.
- A correct result occurs when the individual in the probe image is also in the database **AND** the correct individual has the highest similarity score.
- This is called the **correct detect and identify rate**.

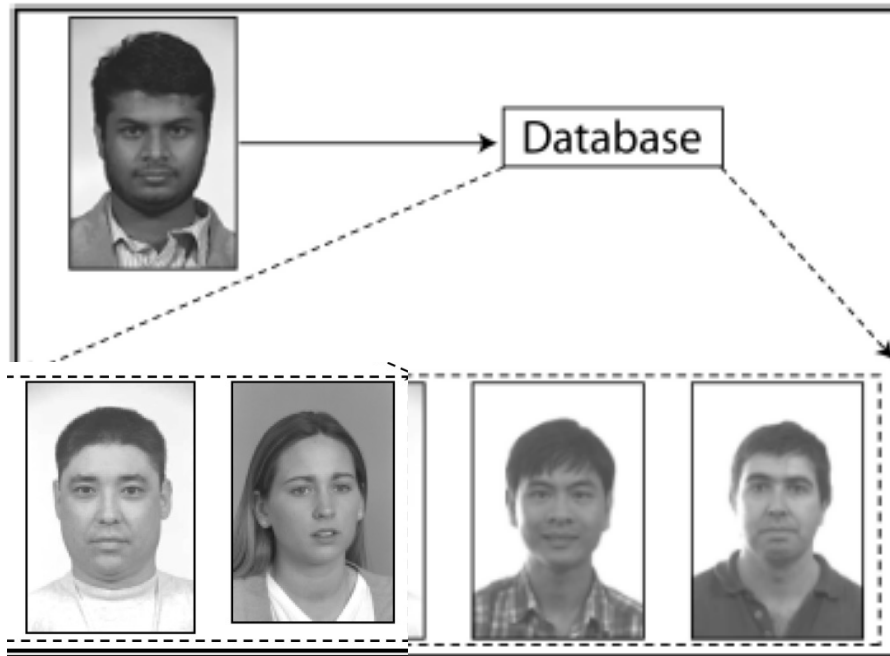


Possible errors: identification – open set





Possible errors: identification – open set



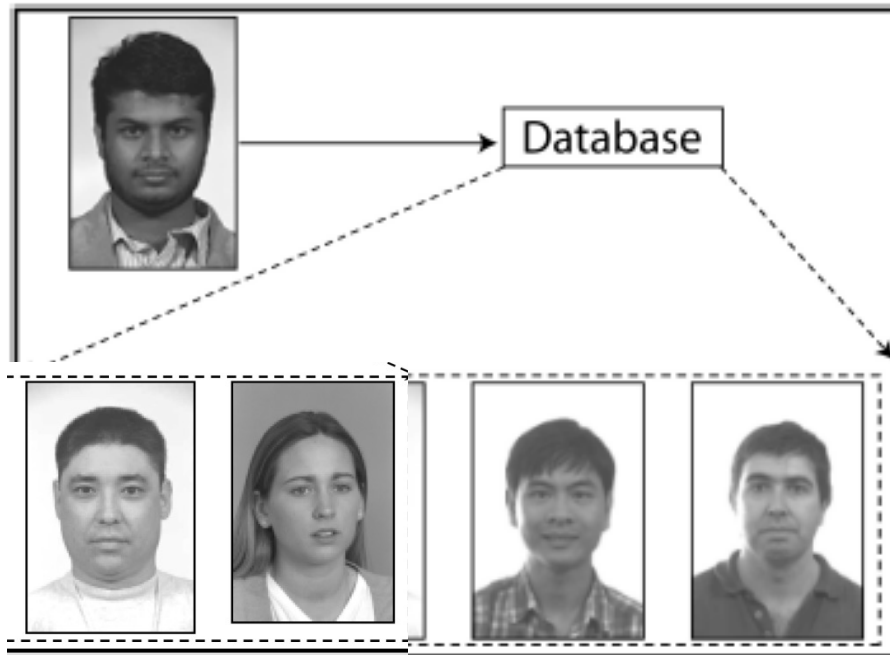
Scores = 0.8 0.7 0.6 0.4 threshold = 0.85

- No individual above the threshold = no detect (no alarm)
→ We do not care about looking at the top individual = no identification

→ **correct result**



Possible errors: identification – open set



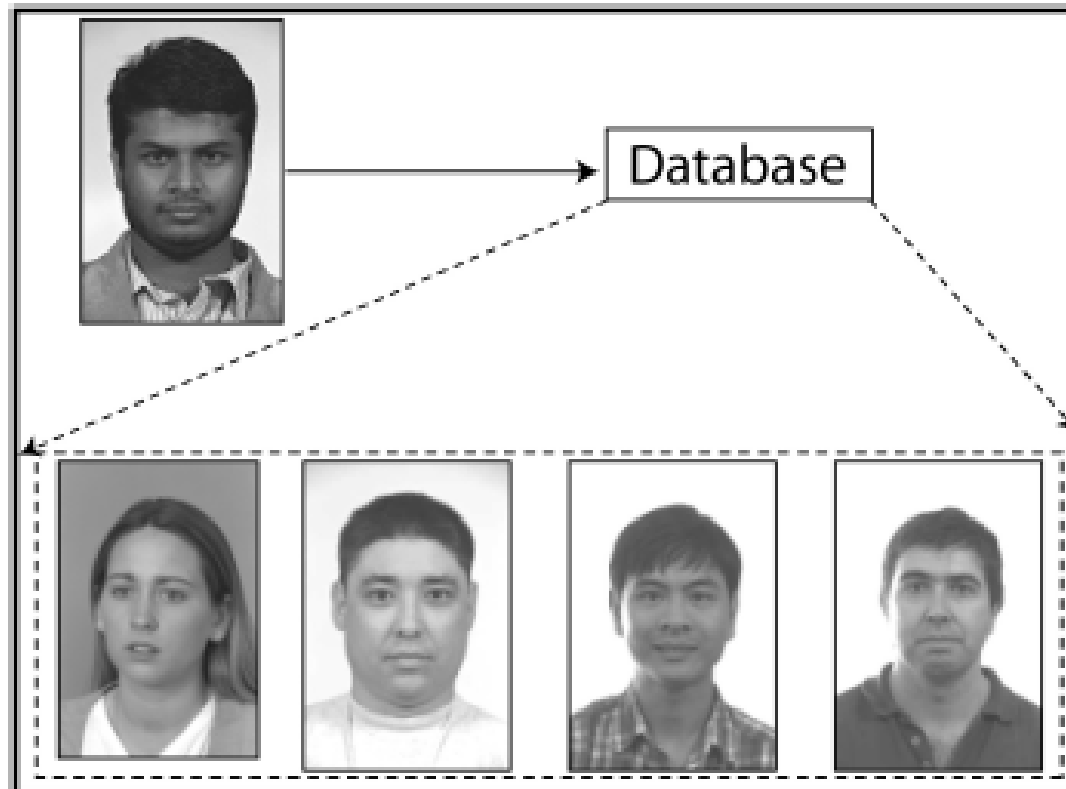
Scores = 0.8 0.7 0.6 0.4 threshold = 0.75

- Two individuals above the threshold = detect (alarm)
- The first individual is not the right one = no correct identification

→ false alarm



Possible errors: identification – open set



- If we run many trials with probes belonging to subjects not in the database (set P_N), we will know how often the system will return an incorrect alarm.
- This is called **false alarm rate**.



Possible errors: identification – open set

- **rango(p_j)** = the position in the list where the first template for the correct identity is returned
- **DIR (at rank k) (Detection and Identification Rate (at rank k))**: the probability of correct identification at rank k (the correct subject is returned at position k)
- The rate between the number of individuals correctly recognized at rank k and the number of probes belonging to individuals in P_G

$$DIR(t, k) = \frac{|\{p_j : rango(p_j) \leq k, s_{ij} \geq t, id(g_i) = id(p_j)\}|}{|P_G|} \quad \forall p_j \in P_G$$

- **FRR or more specifically FNIR (False Reject Rate or False Negative Identification Rate)**: the probability of false reject expressed as 1 - DIR (at rank 1)

$$FRR(t) = 1 - DIR(t, 1)$$



Possible errors: identification – open set

- **FAR or more specifically FPIR (False Acceptance Rate or False Positive Identification Rate) or False Alarm Rate (Watch List):** the probability of false acceptance/alarm
- The rate between the number of impostor recognized by error and the total number of impostors in P_N

$$FAR(t) = \frac{|\{p_j : \max_i s_{ij} \geq t\}|}{|P_N|} \quad \forall p_j \in P_N \quad \forall g_i \in G$$

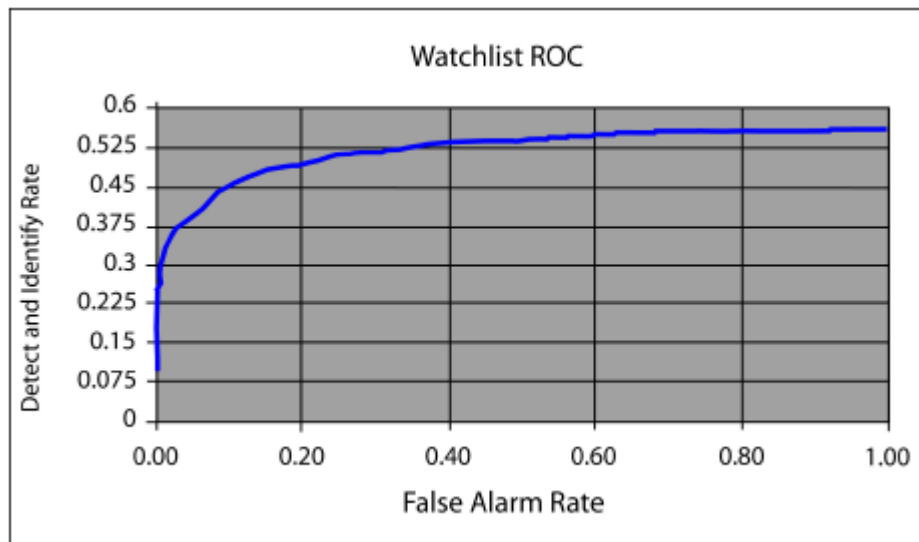
- **EER (Equal Error Rate):** the point where the two probability errors are equal, i.e., $FRR = FAR$

$$EER = \{x : FRR(t) = x \wedge FAR(t) = x\}$$



Possible errors: identification – open set

- As in verification, we would like to be able to set our threshold so that the detect and identify rate is 100%, and the false alarm rate is 0%.
- This is not possible for the same reasons : FAR and FRR both depend from the score/similarity/distance threshold (they are connected) yet in opposite directions.
 - If we raise the threshold, the detect and identify rate decreases, but our false alarm rate also decreases.
 - If we lower the threshold, the detect and identify rate increases, but our false alarm rate also increases.
- We can plot detect and identify rates and their associated false alarm rates.
- We can call this plot Open-set (Watchlist) Receiver Operating Characteristic, or Open-set (Watchlist) ROC.



Each threshold corresponds to a point on the curve



Possible errors: identification – open set

Selection of a watchlist threshold will depend on the kind of application.

In practice, we can identify five operational areas:

- **Applications requiring extremely low false alarm.** When any alarm requires immediate action, this could lead to public disturbance and confusion. Moreover, an alarm and subsequent action may make evident that surveillance is being performed and how, and may minimize the possibility of catching a future suspect.
- **Applications requiring extremely high probability of detect and identify.** The main concern is detecting someone on the watchlist; false alarms are a secondary concern and will be dealt with according to pre-defined procedures.



Possible errors: identification – open set

- **Applications requiring low false alarm and detect/identify.** The main concern is lower false alarms and it is acceptable to deal with low detect/identify.
- **Applications requiring high false alarm and detect/identify.** The main concern is higher detect/identify performance and it is acceptable deal with a high false alarm rate as well.
- **Applications requiring no threshold.** The user wants all results with corresponding confidence measures for investigation.



Possible errors: identification – closed set

- Closed set identification is a special case of the open set identification (watchlist) task
- We can assume for sure that every single probe image has a corresponding match in the database.
 - The first question of the watchlist task (is this person in the database) is already answered.
- The remaining question is how close it is to the gallery template(s) belonging to the same subject.
- In practice, there are very few applications that operate under the closed set identification task
 - FBI's Integrated Automated Fingerprint Identification System (IAFIS) actually operates as a watchlist, not identification, task.



Possible errors: identification – closed set



Scores=

0.9

0.86

0.6

0.4

In this example, the correct match has the **top similarity score**.

If we run many trials with different subjects, we will know how often the system will return a correct result **with the top match**. This is termed the **probability of identification at rank 1**.



Possible errors: identification – closed set



Scores=

0.86

0.8

0.6

0.4

In this example, the correct match has the **second highest similarity score**.

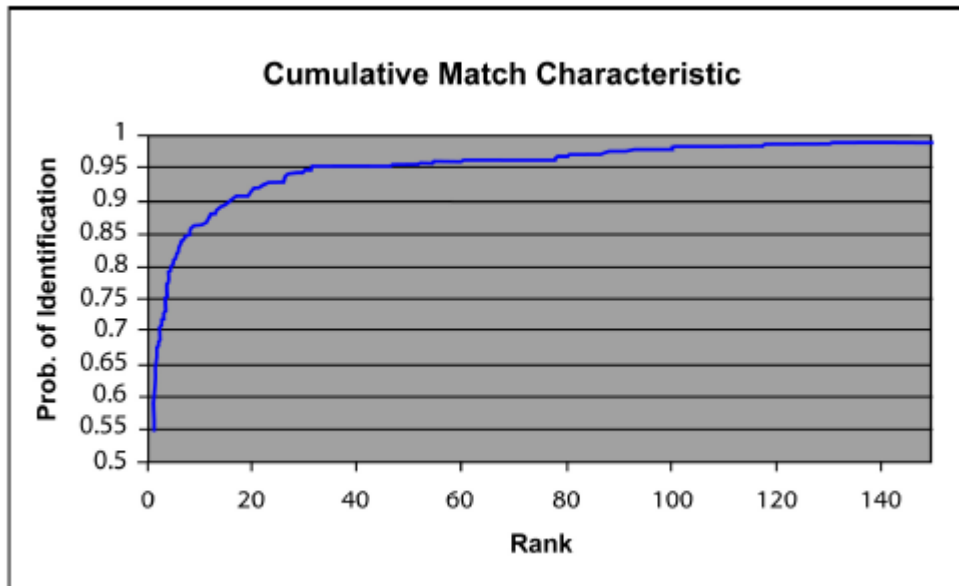
If we run many trials with different subjects, we will know how often the system will return a correct result **with either the top or second similarity score** (we do not necessarily care if they are in the top or second, just that they are in one of those positions). This is termed the **probability of identification at rank 2**.

The probability of correct identification at rank 20 means: what is the probability that the correct match is somewhere in the top 20 similarity scores?



Possible errors: identification – closed set

- **CMS (at rank k)** (Cumulative Match Score (at rank k) – The *probability of identification at rank k*, or even the ratio between the number of individuals which are correctly recognized among the first k and the total number of individuals in the test set (probe).
- **CMC** (Cumulative Match Characteristic) – A *Cumulative Match Characteristic (CMC)* curve shows the *CMS* value for a certain number of ranks (clearly, each implying the following ones). It therefore reports the probability that the correct identity is returned at the first place in the ordered list (*CMS at rank 1*), or at the first or second place (*CMS at rank 2*), or in general among the first k places (*CMS at rank k*). If the number *n* of ranks in the curve equals the size of the gallery, we will surely have a probability value of 1 at point *n*.



RR (Recognition Rate)
- CMS at rank 1 is also
defined as
Recognition Rate.



Beware of CONFUSION!

It often happens is students' reports to find a GREAT confusion between some of the performance measures used in biometrics and those used in Machine Learning.

PRECISION, RECALL and F-SCORE (a combination of them) DO NOT express exactly the same information, or better the same error statistics (that is what we are most interested in!)

PRECISION = Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances

RECALL (also known as sensitivity) = the fraction of relevant instances that were retrieved.

Both precision and recall are therefore based on relevance. (P = Positive, N = Negative, T = True, F = False)

$\text{PRECISION} = \text{TP} / (\text{TP} + \text{FP})$

$\text{RECALL} = \text{TP} / (\text{TP} + \text{FN})$

F-SCORE = a combination of Precision and Recall



Beware of CONFUSION!

In biometric terms (A = Acceptance, R = Rejection)

PRECISION can be compared with $GA / \text{all accepted} = GA / (FA + GA)$ (positive predictive value)

RECALL can be compared with $GA / \text{genuine} = GA / (GA + FR)$ (true positive rate)

They both start from the correct positive responses!

On the other hand we are interested in

$$FRR = FR / (GA + FR)$$

$$FAR = FA / (FA + GR)$$

Or, using the ML terms Positive (P) for A and Negative (N) for R

$$FRR = FN / (TP + FN)$$

$$FAR = FP / (FP + TN)$$

They are clearly different, since they start from the two types of error!



Beware of CONFUSION!

Making a comparison with Machine Learning

FRR can be compared to Miss Rate or false negative rate (FNR)

FAR can be compared to Fall-Out or False Positive Rate



A further possible source of CONFUSION!

Identification

- The aim is to determine the identity of an unknown subject
- Used in the medium-long term

Re-identification (re-ID)

- The aim is to match a person's *identity* (intended as presence) across different cameras or locations in a video or in an image sequence or in an event set.
- It involves detecting and tracking a person and its actions and then using features such as appearance, body shape, or behaviour to determine the *presence* of the same *identity* in different frames or moments in time.
- The goal is to associate the same person across multiple non-overlapping time slices.
- Used in the short-term



A further possible source of CONFUSION!

- Re-identification is evaluated using separate metrics
 - CMC in case there is a single mated gallery image for each query
 - mAP when there are more mated images for each query
 - AP is the area under Precision-Recall curve computer different thresholds
 - mAP is the average computed over each query
- Re-identification can be linked to identification when we find a subject repeatedly in a video and then check whether he or she appears in a watchlist



Some references

- R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, A. W. Senior, "The Relation between the ROC Curve and the CMC," AUTOID, pp.15 -20, Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), 2005
- G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST1998 speaker recognition evaluation. In Proc. ICSLD, 1998.
<http://www.dtic.mil/dtic/tr/fulltext/u2/a528610.pdf>
- Mohammad Nayeem Teli, J. Ross Beveridge, P. Jonathon Phillips, Geof H. Givens, David S. Bolme, Bruce A. Draper. Biometric Zoos: Theory and Experimental Evidence.
<http://www.csis.pace.edu/~ctappert/dps/2011IJCB/papers/327.pdf>
- Yager, N.; Dunstone, T., "Worms, Chameleons, Phantoms and Doves: New Additions to the Biometric Menagerie," *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on* , vol., no., pp.1,6, 7-8 June 2007.
http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4263204&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxppls%2Fabs_all.jsp%3Farnumber%3D4263204
- Torralba, A., & Efros, A. A. (2011, June). Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1521-1528). IEEE.



Some references

- K. Kryszczuk, J. Richiardi, P. Prodanov and A. Drygajlo, “Reliability-based decision fusion in multimodal biometric verification”, EURASIP Journal on Advances in Signal Processing 2006, Volume 2007 (2007), Article ID 86572, 9 pages.
- De Marsico, M., Nappi, M., & Riccio, D. (2011, November). Measuring measures for face sample quality. In *Proceedings of the 3rd international ACM workshop on Multimedia in forensics and intelligence* (pp. 7-12). ACM.
- N. Poh, S. Bengio, Improving Fusion with Margin-Derived Confidence In Biometric Authentication Tasks, IDIAP-RR 04-63, November 2004.
- M. De Marsico, M. Nappi, D. Riccio, G. Tortora. NABS: Novel Approaches for Biometric Systems. IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews. Volume: 41 Issue:4, July 2011, pp. 481-493