Biometric Systems
Lesson 3 – Recognition Reliability

Maria De Marsico

demarsico@di.uniroma1.it

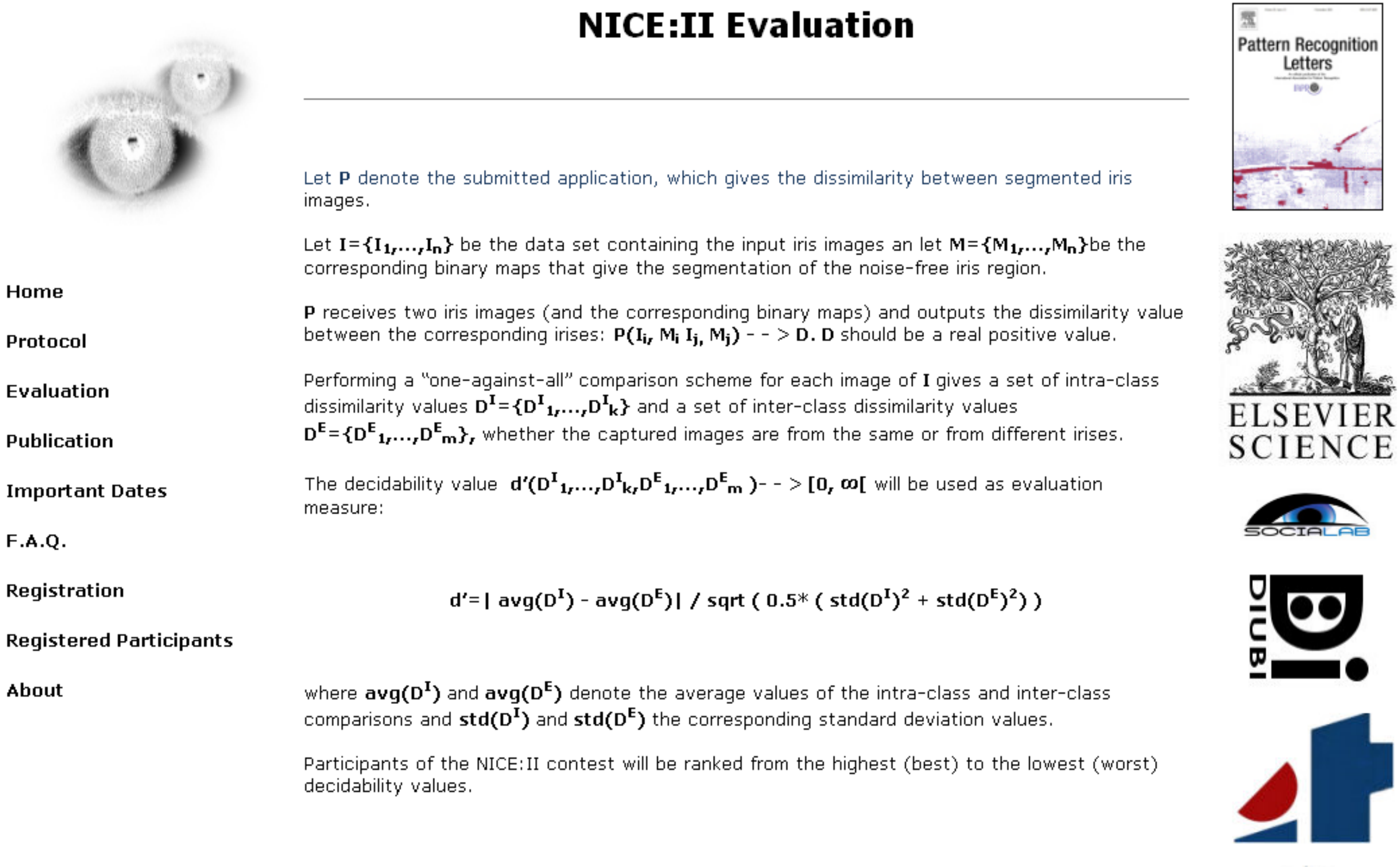
SAPIENZA
UNIVERSITÀ DI ROMA

Dipartimento di
Informatica

Possible in recognition errors

- In practice, the identification task is much more difficult for biometric systems (but also for human operators) than the verification task.
- It is critical to think in terms of the proper task, and their associated statistics, in order to avoid confusion and errors.

Extending the performance measures …



NICE:II Evaluation

Let **P** denote the submitted application, which gives the dissimilarity between segmented iris images.

Let $\mathbf{I=\{I_1,...,I_n\}}$ be the data set containing the input iris images an let $\mathbf{M=\{M_1,...,M_n\}}$be the corresponding binary maps that give the segmentation of the noise-free iris region.

P receives two iris images (and the corresponding binary maps) and outputs the dissimilarity value between the corresponding irises: $\mathbf{P(I_i, M_i\ I_j, M_j)}$ - - > **D**. **D** should be a real positive value.

Performing a "one-against-all" comparison scheme for each image of **I** gives a set of intra-class dissimilarity values $\mathbf{D^I=\{D^I{}_1,...,D^I{}_k\}}$ and a set of inter-class dissimilarity values $\mathbf{D^E=\{D^E{}_1,...,D^E{}_m\}}$, whether the captured images are from the same or from different irises.

The decidability value $\mathbf{d'(D^I{}_1,...,D^I{}_k,D^E{}_1,...,D^E{}_m\ )}$- - > **[0, ∞[** will be used as evaluation measure:

$$d' = |\ avg(D^I) - avg(D^E)|\ /\ sqrt\ (\ 0.5 * (\ std(D^I)^2 + std(D^E)^2)\ )$$

where $\mathbf{avg(D^I)}$ and $\mathbf{avg(D^E)}$ denote the average values of the intra-class and inter-class comparisons and $\mathbf{std(D^I)}$ and $\mathbf{std(D^E)}$ the corresponding standard deviation values.

Participants of the NICE:II contest will be ranked from the highest (best) to the lowest (worst) decidability values.

Closer considerations …

- Measures like FAR, FRR, CMS, … are not enough to give a thorough evaluation of algorithms:
  - Ex-post measures based on ex-ante training on limited datasets (interesting observation from Torralba et al. (2011)
  - Operation context can be different or even change, causing score distributions to change in turn.

- For reliable comparison we have to consider:
  - Number and characteristics of the databases used;
  - Size of images;
  - Size of Probe and Gallery;
  - Amount and quality of addressed as well as tolerated variations;
  - Possible interoperability (e.g., cross-dataset generalization)

Closer considerations …

- From Torralba et al. (2011)

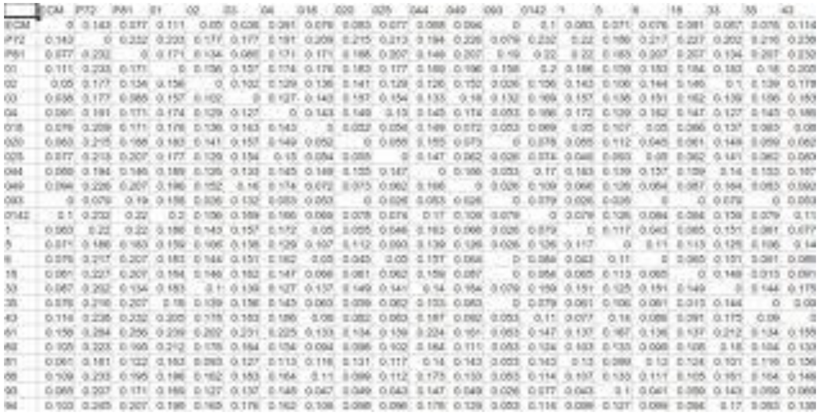
« Datasets are an integral part of contemporary object recognition research. They have been the chief reason for the considerable progress in the field, not just as source of large amounts of training data, but also as means of measuring and comparing performance of competing algorithms. At the same time, datasets have often been blamed for narrowing the focus of object recognition research, reducing it to a single benchmark performance number. Indeed, some datasets, that started out as data capture efforts aimed at representing the visual world, have become closed worlds unto themselves (e.g. the Corel world, the Caltech-101 world, the PASCAL VOC world). With the focus on beating the latest benchmark numbers on the latest dataset, have we perhaps lost sight of the original purpose? »

How it works in practice?

- For each pair(probe$_i$,gallery$_j$) we compute the distance and put it in the corresponding position of a distance matrix DM.

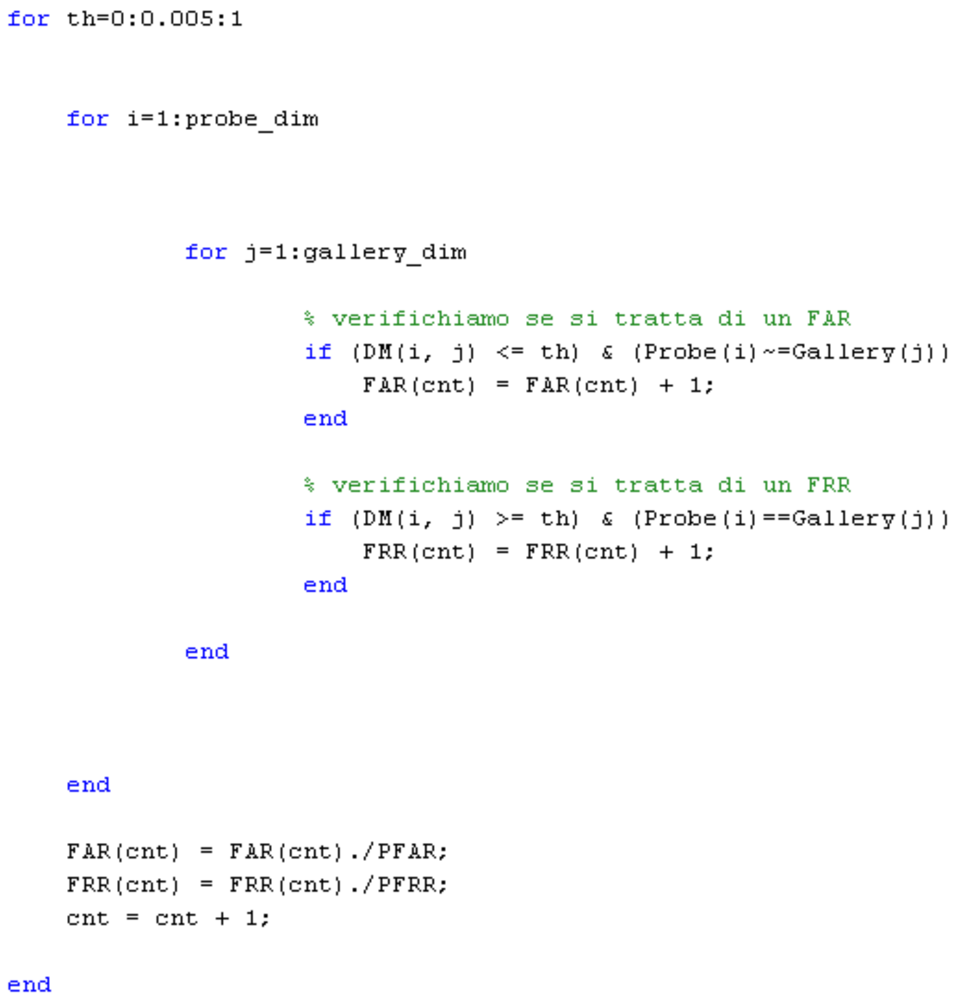
$$\begin{pmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{12} & 0 & d_{23} & d_{24} \\ d_{13} & d_{23} & 0 & d_{34} \\ d_{14} & d_{24} & d_{34} & 0 \end{pmatrix}$$

- Example: if the test is all-against-all (each gallery template matched with all) and the distance measure is symmetric, DM is symmetric with null diagonal.



Example: from DM to FAR/FRR

- For each threshold value *th*
    - for each pair (probe$_i$, gallery$_j$) we check if we have:
    - FAR: distance is lower than the threshold DM(i,j) < th, but samples are not from the same subject, i.e. they have different labels Probe(i)$\neq$Gallery(j)
    - FRR: distance is lower than the threshold DM(i,j) > th, but samples are from the same subjecti.e., they have the same label Probe(i)=Gallery(j)



```
for th=0:0.005:1

    for i=1:probe_dim


            for j=1:gallery_dim

                    % verifichiamo se si tratta di un FAR
                    if (DM(i, j) <= th) & (Probe(i)~=Gallery(j))
                        FAR(cnt) = FAR(cnt) + 1;
                    end

                    % verifichiamo se si tratta di un FRR
                    if (DM(i, j) >= th) & (Probe(i)==Gallery(j))
                        FRR(cnt) = FRR(cnt) + 1;
                    end

            end



    end

    FAR(cnt) = FAR(cnt)./PFAR;
    FRR(cnt) = FRR(cnt)./PFRR;
    cnt = cnt + 1;

end
```

Note 1

- A **metric** on a set *X* is a function (called the *distance function* or simply **distance**)
- $d : X \times X \to \mathbf{R}$
- (where **R** is the set of real numbers). For all *x*, *y*, *z* in *X*, this function is required to satisfy the following conditions:
- $d(x, y) \geq 0$ (*non-negativity*, or separation axiom)
- $d(x, y) = 0$ if and only if $x = y$ (*identity of indiscernibles*, or coincidence axiom)
- $d(x, y) = d(y, x)$ (*symmetry*)
- $d(x, z) \leq d(x, y) + d(y, z)$ (*subadditivity* / *triangle inequality*).

Note 2

- A **semimetric** on X is a function $d : X \times X \rightarrow \mathbf{R}$ that satisfies the first three axioms, but not necessarily the triangle inequality:
- $d(x, y) \geq 0$
- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) = d(y, x)$

- If the masure is not symmetric, i.e., $d(x, y) \neq d(y, x)$, we can use instead $d^*(x, y) = d^*(y, x)=(d(x, y) + d(y, x))/2$

Relations among evaluation measures

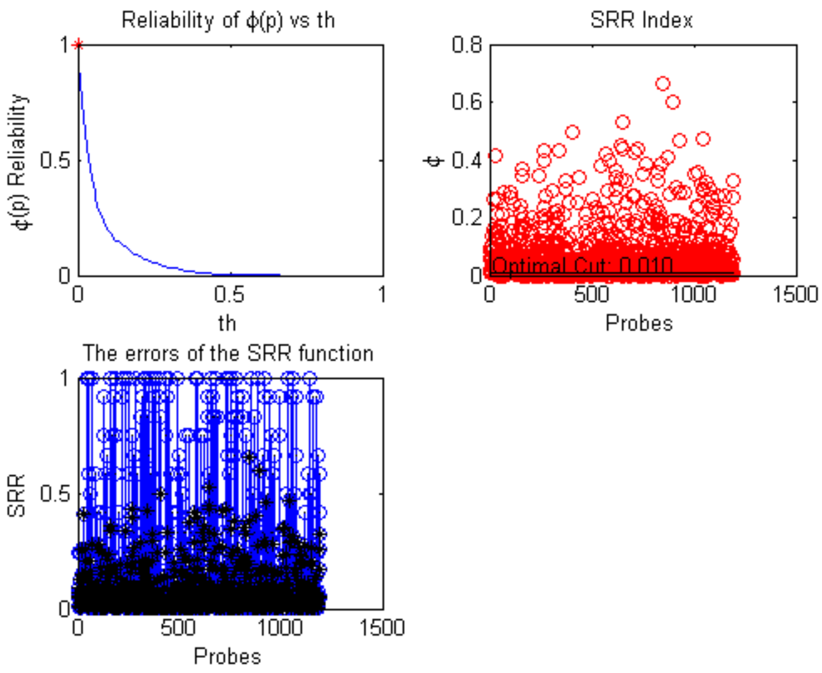
- In 2005 Bolle, Connell, Pankanti, Ratha e Senior demonstrated that CMC is directly related to ROC, interpreted as a measure of the trade off between FAR and FRR as a function of the operation threshold.

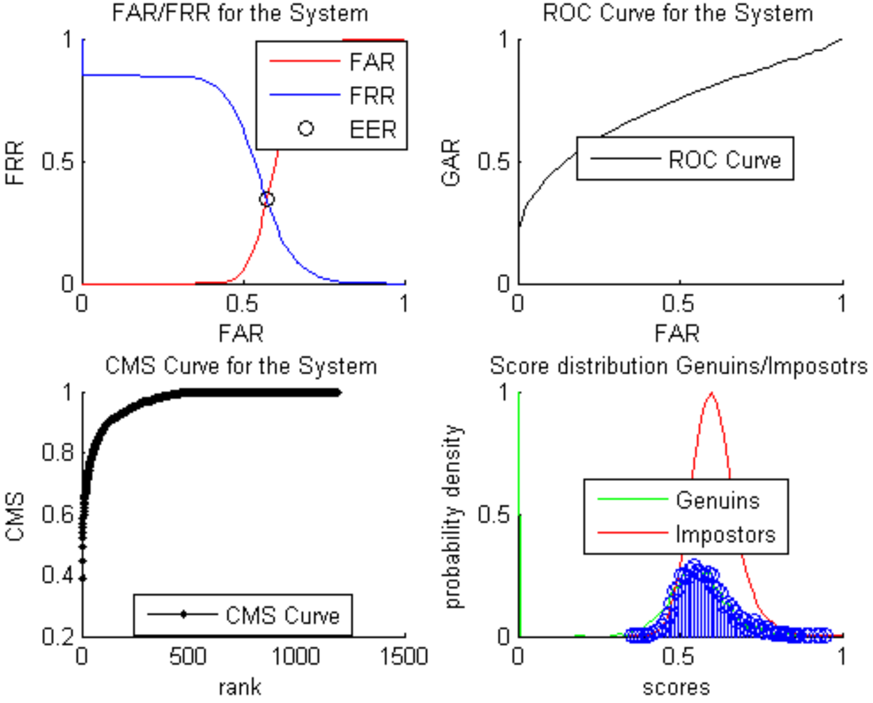
- They show that the CMC is also related to the FAR and FRR of a 1:1 matcher, i.e., the matcher that is used to rank the candidates by sorting the scores. This has as a consequence that when a 1:1 matcher is used for identification, that is, for sorting match scores from high to low, the CMC does not offer any additional information beyond the FAR and FRR curves. The CMC is just another way of displaying the data and can be computed from the FAR and FRR.

R.M. Bolle, J.H. Connell. S. Pananti, N.K. Ratha and A.W. Senior, “The Relation Between the ROC Curve and the CMC”. Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AUTOID ’05), pp. 15-20, 2005.

From DM to performance measures

- Starting from the distance matrix DM it is possible to compute many performance measures.
- Often called Figures of Merit (FoMs)





Doddington zoo

Doddington et al. defined Sheep, Goats, Lambs and Wolves in the context of speaker recognition systems:

- **Sheep**: A person who is a Sheep produces a biometric that matches well to other biometrics of themselves and poorly to those of other people. As such, Sheep generate fewer false accepts and rejects than average. (Normal average behaviour)

- **Goats**: A person who is a Goat produces a biometric that matches poorly to other biometrics of themselves. These low match scores imply a higher than average false reject rate for Goats.

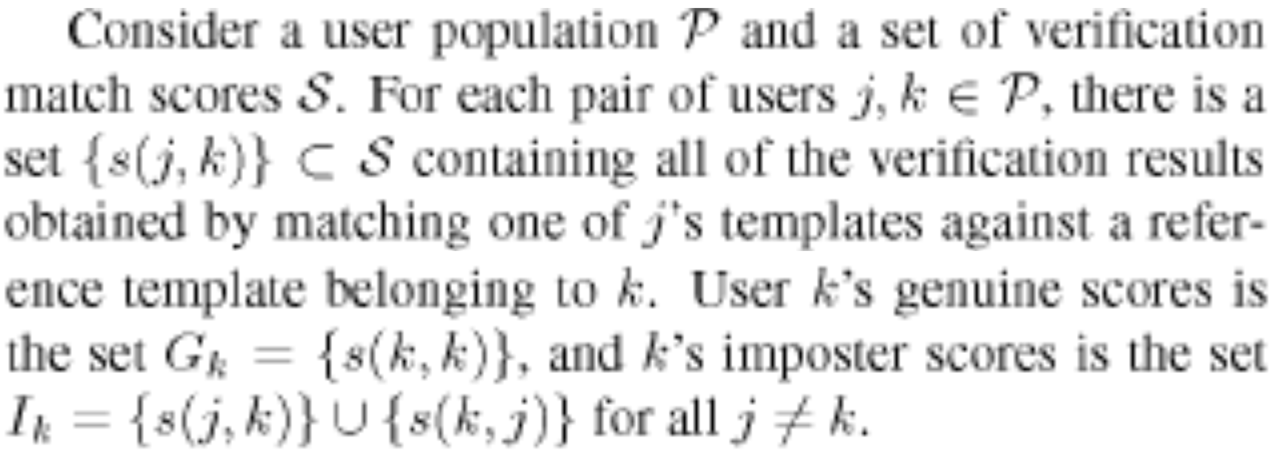
Doddington zoo

- **Lambs**: A person who is a Lamb can be easily impersonated. When the biometric of such a person is paired to a biometric from a different person the resulting match score will be higher than average. Consequently, false matches are more likely.

- **Wolves**: A person who is a Wolf is good at impersonation. When such a person presents a biometric for comparison they have an above average chance of generating a higher than average match score when compared to a stored biometric of a different person.

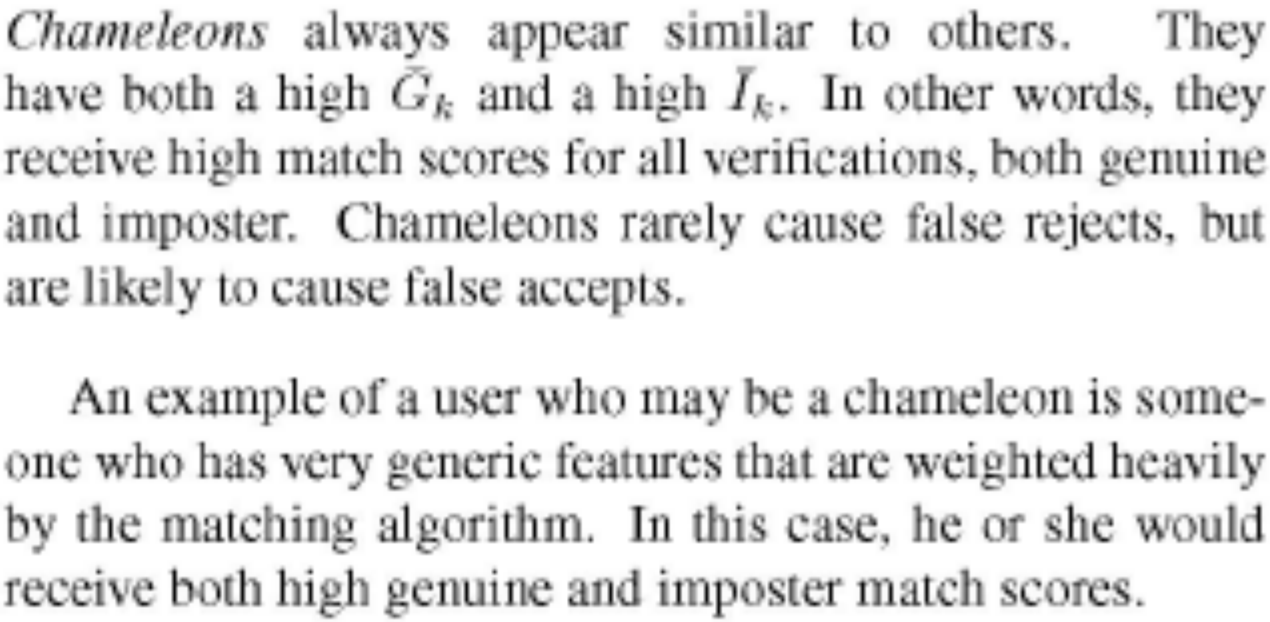
Extending the managerie …

Goats, lambs, and wolves are defined in terms of a user's average genuine or imposter scores. New additions to the biometric menagerie by Yager and Dunstone are defined in terms of both a user's genuine and imposter scores.



Consider a user population $\mathcal{P}$ and a set of verification match scores $\mathcal{S}$. For each pair of users $j, k \in \mathcal{P}$, there is a set $\{s(j,k)\} \subset \mathcal{S}$ containing all of the verification results obtained by matching one of j's templates against a reference template belonging to k. User k's genuine scores is the set $G_k = \{s(k,k)\}$, and k's imposter scores is the set $I_k = \{s(j,k)\} \cup \{s(k,j)\}$ for all $j \neq k$.

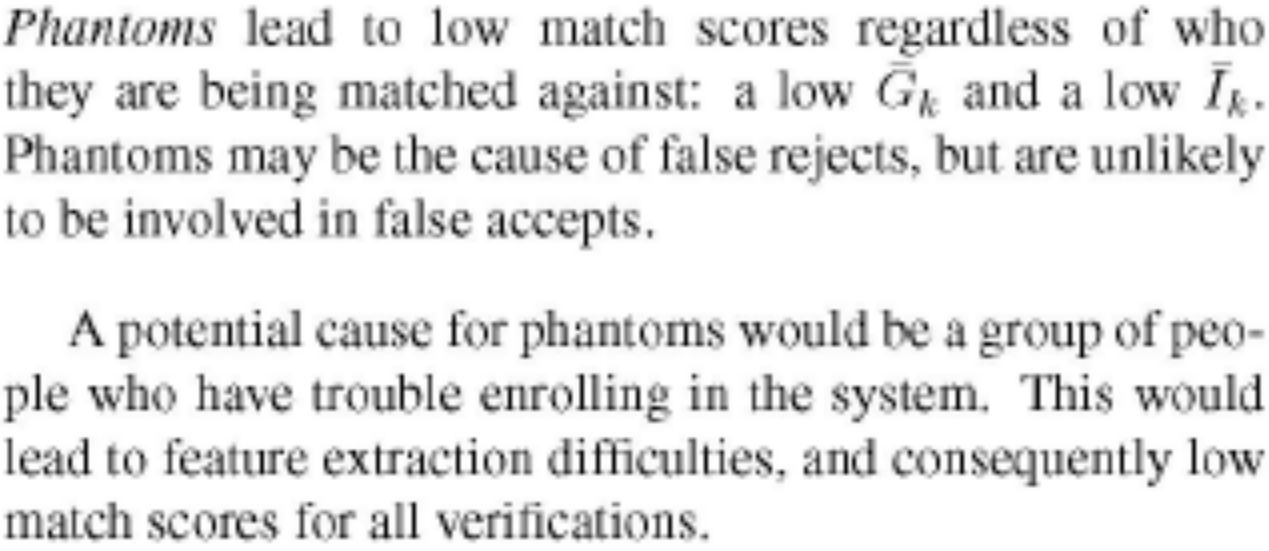
Extending the managerie …



Chameleons always appear similar to others. They have both a high $\bar{G}_k$ and a high $\bar{I}_k$. In other words, they receive high match scores for all verifications, both genuine and imposter. Chameleons rarely cause false rejects, but are likely to cause false accepts.

An example of a user who may be a chameleon is someone who has very generic features that are weighted heavily by the matching algorithm. In this case, he or she would receive both high genuine and imposter match scores.

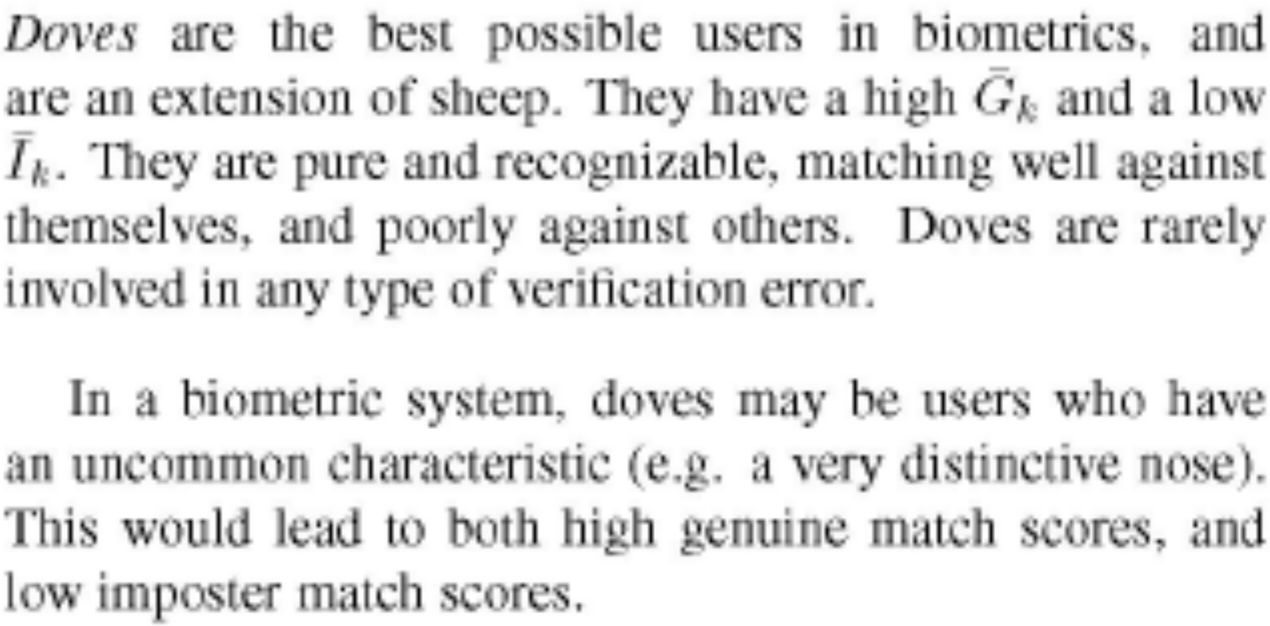
Extending the managerie …



Phantoms lead to low match scores regardless of who they are being matched against: a low $\bar{G}_k$ and a low $\bar{I}_k$. Phantoms may be the cause of false rejects, but are unlikely to be involved in false accepts.

A potential cause for phantoms would be a group of people who have trouble enrolling in the system. This would lead to feature extraction difficulties, and consequently low match scores for all verifications.

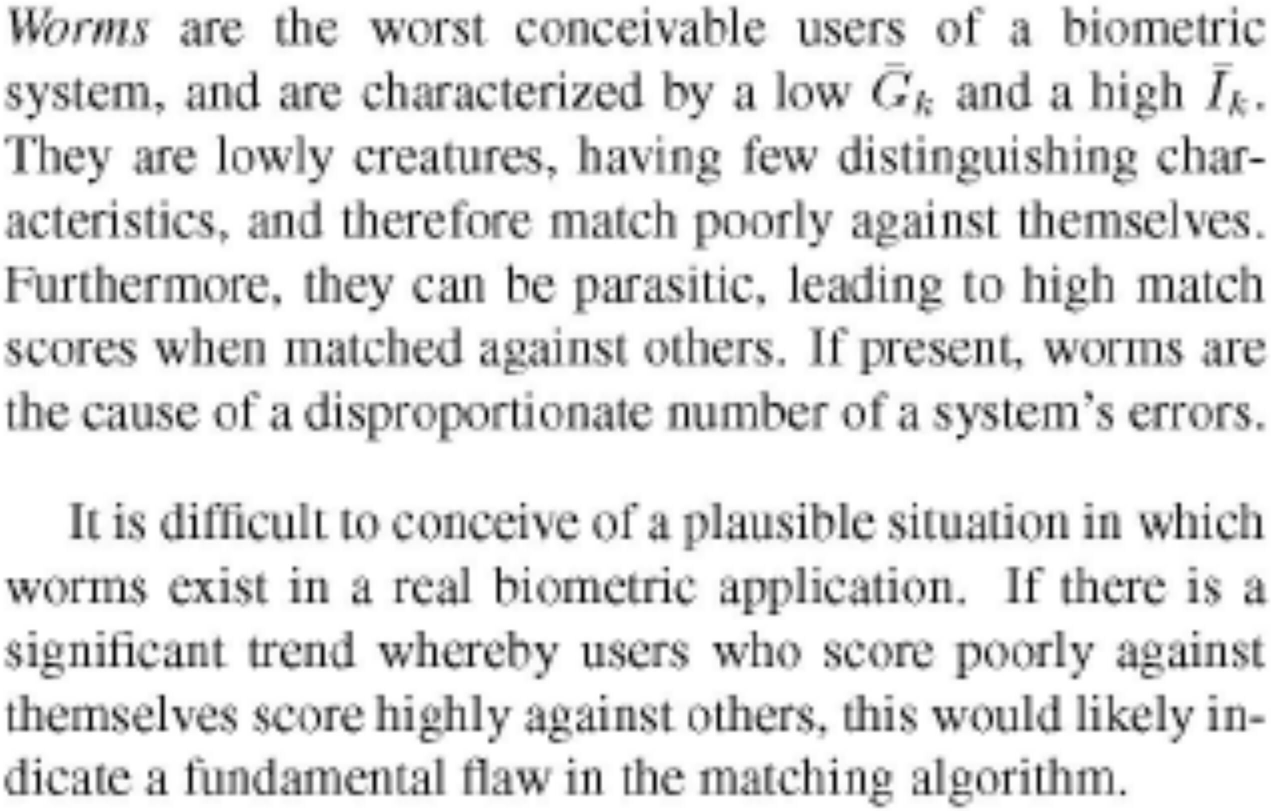
Extending the managerie …



Doves are the best possible users in biometrics, and are an extension of sheep. They have a high $\bar{G}_k$ and a low $\bar{I}_k$. They are pure and recognizable, matching well against themselves, and poorly against others. Doves are rarely involved in any type of verification error.

In a biometric system, doves may be users who have an uncommon characteristic (e.g. a very distinctive nose). This would lead to both high genuine match scores, and low imposter match scores.

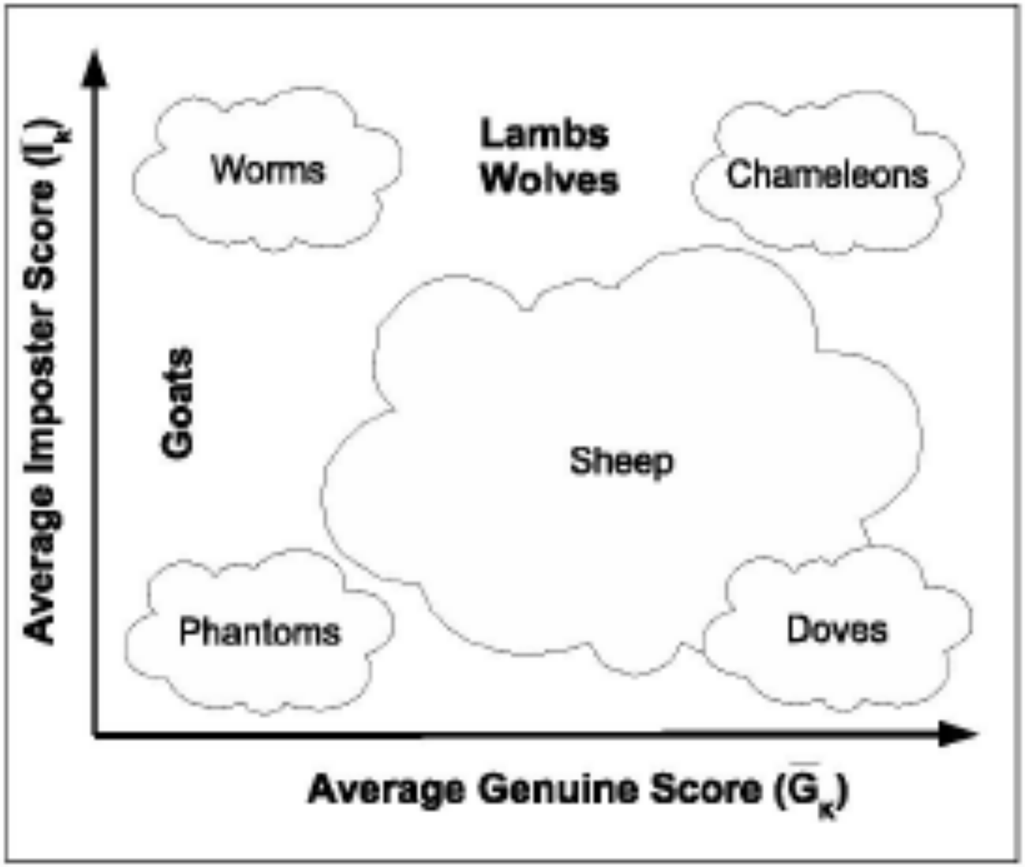
Extending the managerie …



Worms are the worst conceivable users of a biometric system, and are characterized by a low $\bar{G}_k$ and a high $\bar{I}_k$. They are lowly creatures, having few distinguishing characteristics, and therefore match poorly against themselves. Furthermore, they can be parasitic, leading to high match scores when matched against others. If present, worms are the cause of a disproportionate number of a system's errors.

It is difficult to conceive of a plausible situation in which worms exist in a real biometric application. If there is a significant trend whereby users who score poorly against themselves score highly against others, this would likely indicate a fundamental flaw in the matching algorithm.

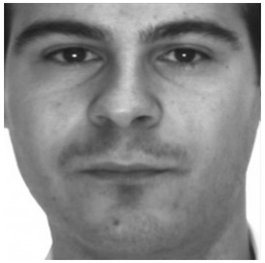
Extending the managerie …


Average Imposter Score ($\bar{I}_k$)
Worms
Lambs
Wolves
Chameleons
Goats
Sheep
Phantoms
Doves
Average Genuine Score ($\bar{G}_k$)

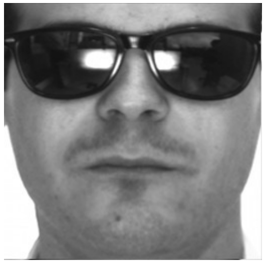
Reliability of an identification system

- Due to the possible different quality of input to different systems, and to possible accuracy in recognition procedures, it may happen that, notwithstanding global FOMs, not all responses are equally reliable.
- The definition of a reliability measure for each single response from a system provides further information to be used in setting up an operation policy (e.g., is f the identification is not reliable enough and if possible, repeat capture), but also to merge results from different systems (multibiometric architectures).

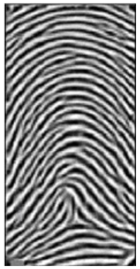
Reliable



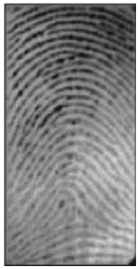
Not Reliable



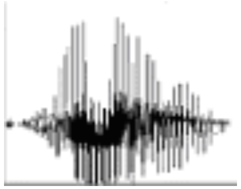
Reliable



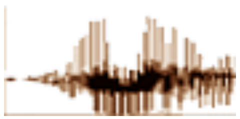
Not Reliable



Reliable

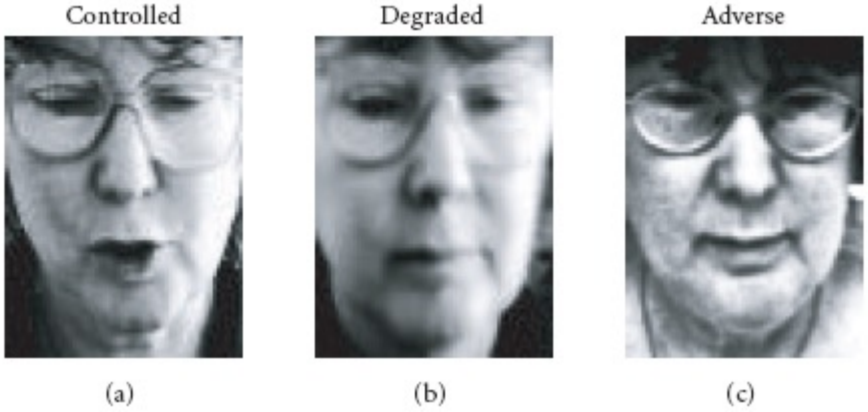


Not Reliable



Some approaches: 1) image quality

- Margins based on quality
- (Kryszczuk, Richiardi, Prodanov and Drygajlo, 2006):



Examples from BANCA database

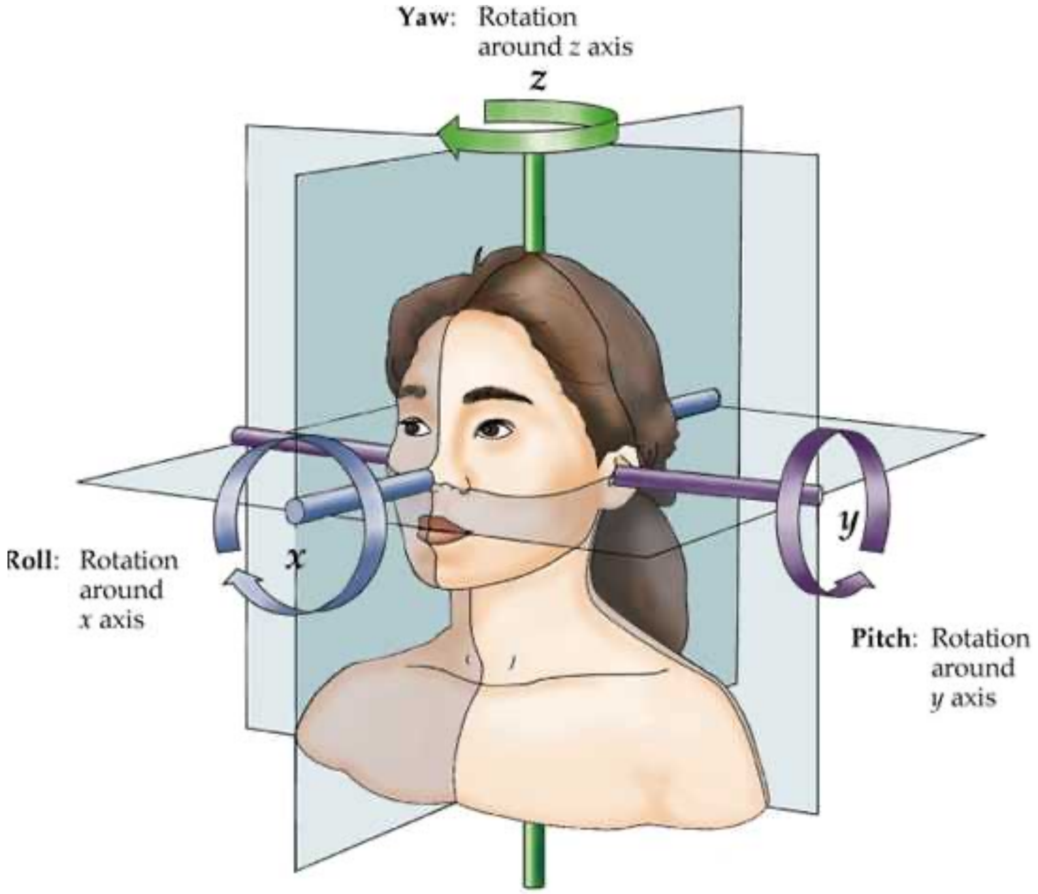
Correlation with an "average" image

The quality of training images can be modeled by creating an "average" template from all faces, the quality of which is taken as a reference*.*

Estimation of image sharpness

The lack of high frequency details in the image can be dscribed as a loss of sharpness (blurring)

Some approaches: 2) face “quality”
How to measure the “quality” of a face image


Yaw: Rotation around z axis
z
Roll: Rotation around x axis
x
y
Pitch: Rotation around y axis

How to measure the “quality” of a face image

De Marsico, Nappi, Riccio (2011)

- SP: measure of distortion with respect to frontal pose, expressed in terms of misalignment of roll (α), yaw (β) and pitch (γ):

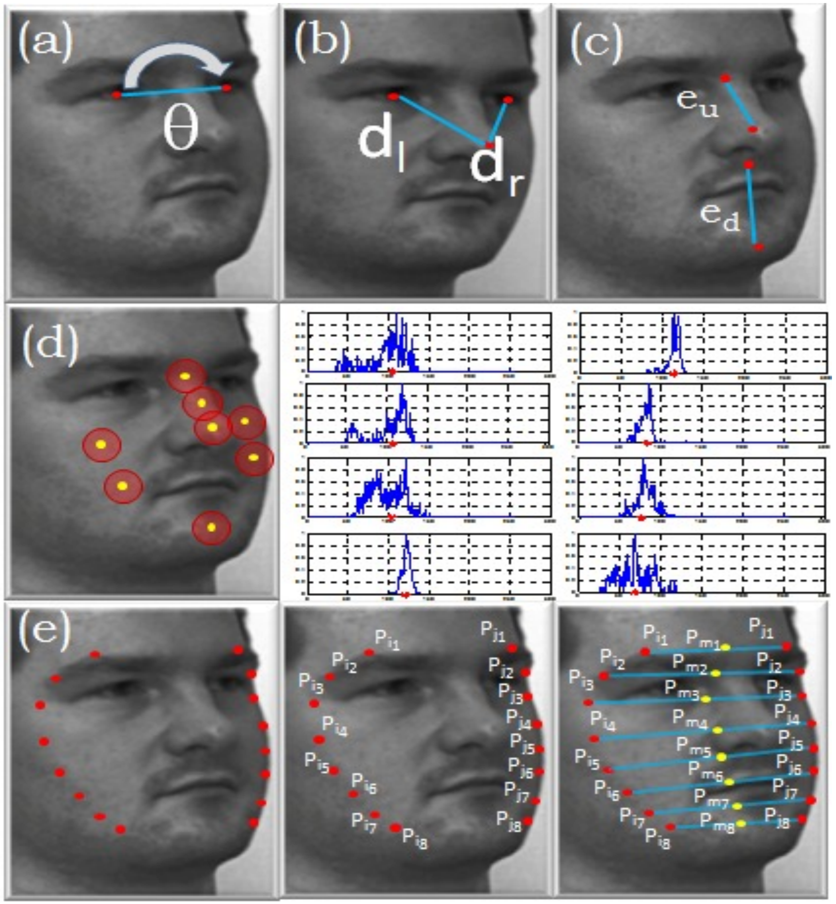
$$SP = \alpha \cdot (1 - roll) + \beta \cdot (1 - yaw) + \gamma \cdot (1 - pitch)$$

- SI: is defined as a measure of homogeneity of grey levels in some pre-determined face regions:

$$SI = 1 - F(std(mc))$$

- SY: is defined as a measure of face symmetry.

$$SY = \sum_{(i,j)\in X} sym(P_i, P_j).$$



Other more general “quality” measures

Universal Image Quality Index (UIQI): any image distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion

Let x={x_i|i=1 … N} and y={y_i|i=1 .. N} be the original and test image respectively. The index is defined as

$$Q = \frac{4\,\sigma_{xy}\,\bar{x}\,\bar{y}}{(\sigma_x^2 + \sigma_y^2)\,[(\bar{x})^2 + (\bar{y})^2]}\,, \qquad (1)$$

where

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i\,,$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\,, \quad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\,,$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})\,.$$

Other “quality” measures

Sharpness Estimation Quality Index: In order to estimate the sharpness of an image I of x × y pixels, we compute the mean of intensity differences between adjacent pixels, taken in both the vertical and horizontal directions:

$$\text{SE} = \frac{1}{2}\left(\frac{1}{(x-1)y}\sum_{m=1}^{y}\sum_{n=1}^{x-1}\left|p_{n,m} - p_{n+1,m}\right| + \frac{1}{(y-1)x}\sum_{m=1}^{y-1}\sum_{n=1}^{x}\left|p_{n,m} - p_{n,m+1}\right|\right)$$

Example tests on face datasets

FERET: first 250 images of the fa (frontal) group, corresponding to 116 subjects

LFW (Labeled Face in the Wild) : 480 images, the first 6 of the first 80 subjects

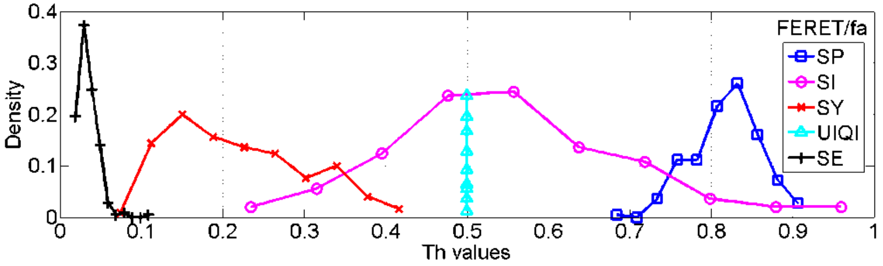
Scface: 650 images of 130 subjects, corresponding to groups cam1-5 (visible light) of the subgroup dist3 (greater distance)

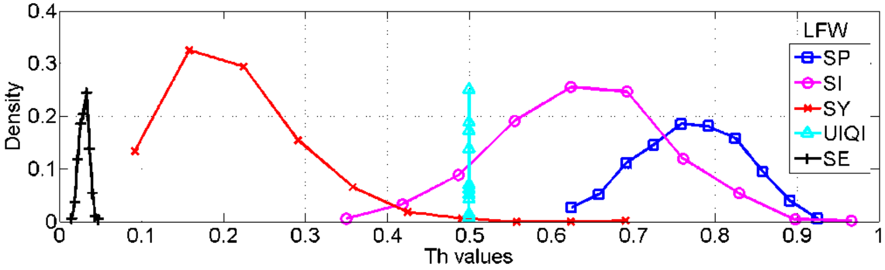
How to measure the “quality” of a quality measure

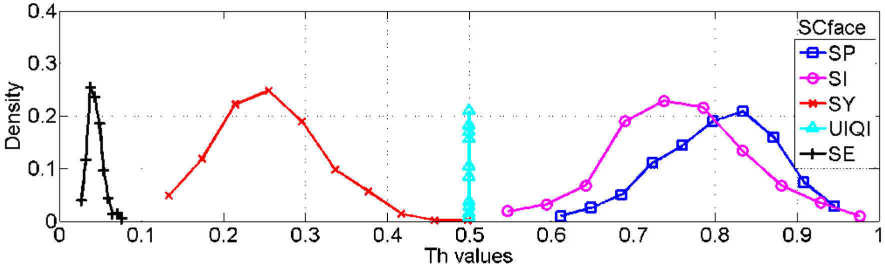
- The first test for a quality measure is to check how the values of a dataset are distributed w.r.t. the returned values.

- This allows to understand which is the average level of quality of a face dataset w.r.t. to a specific measure.

- Two or more measures can be compared by computing the amount of correlation of the returned values w.r.t. to the images of given dataset.





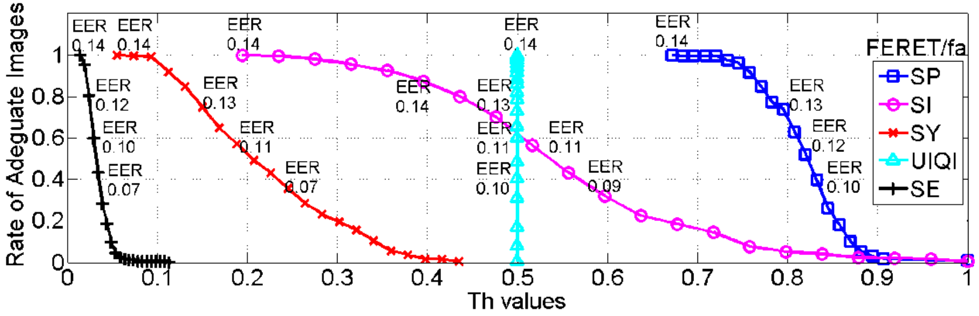


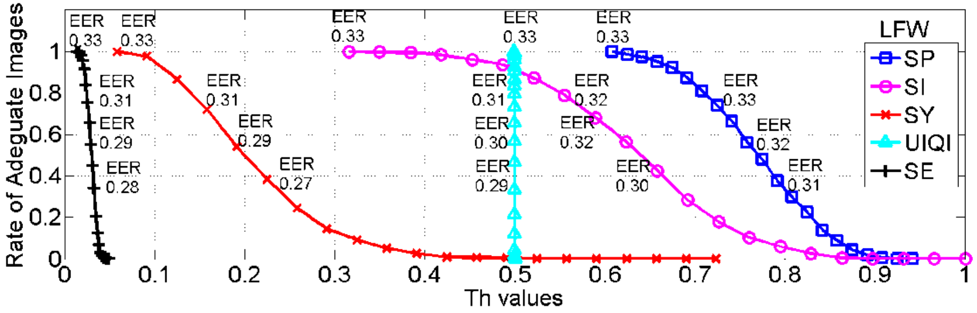
How to measure the “quality” of a quality measure

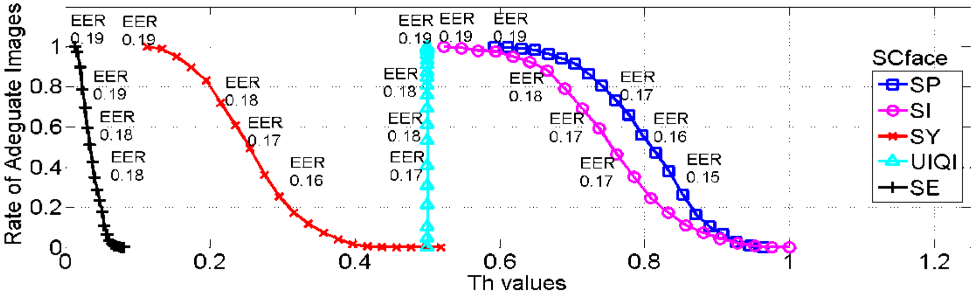
- A measure of quality of input samples allows to discard a-priori (before recognition) those affected by too high distortion which would lead to a wrong response or not reliable from the system.

- A further test for a quality measure is to evaluate how it affects the system performance (EER) by varying a tolerance threshold.

- A good quality measure must provide error good decrease by discarding as few samples as possible.







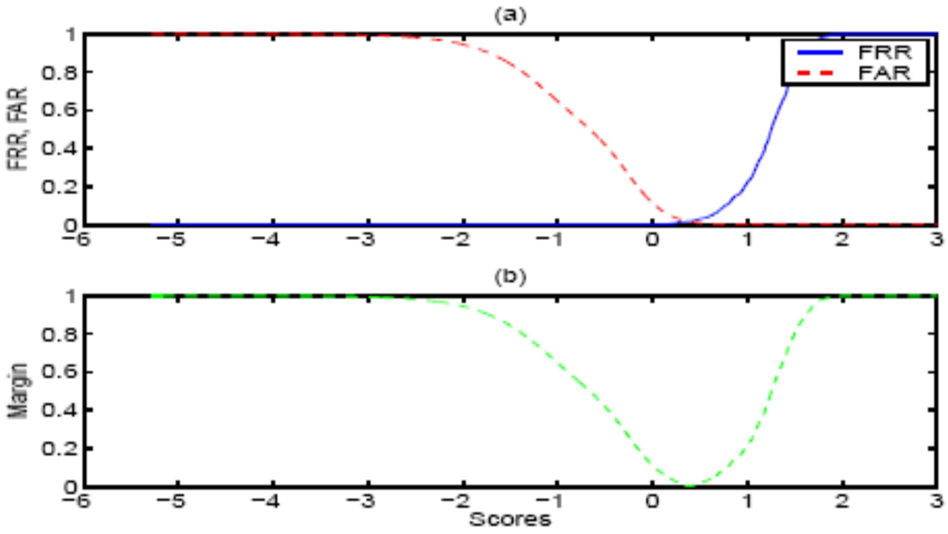
Some approaches: 3) margins based on error estimation

Poh and Bengio, 2004

System performance is measured in terms of:

$$\mathrm{FAR}(\Delta) = \frac{\text{number of FAs}(\Delta)}{\text{number of impostor accesses}},$$

$$\mathrm{FRR}(\Delta) = \frac{\text{number of FRs}(\Delta)}{\text{number of client accesses}}.$$



Margin *M*(*Δ*) is defined as:

$$\mathcal{M}(\Delta) = |\mathrm{FAR}(\Delta) - \mathrm{FRR}(\Delta)|.$$

Some approaches: 4) System Response Reliability (SRR)

There is a major difference between a quality measure for an input sample and a reliability measure for the response of a biometric system.

System Response Reliability (*srr*∈[*0,1*]) index measures the ability of an identification system to separate genuine subjects from impostors on a single probe basis.

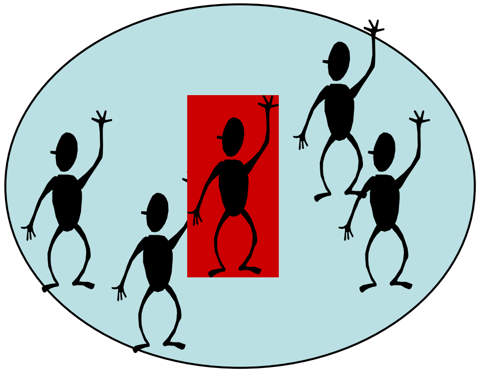
The SRR relies on different versions of function φ. We defined and tested two different φ functions:

- Relative distance;
- Density ratio;

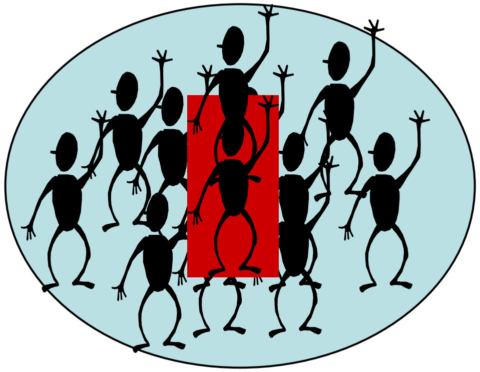
Both functions measure the amount of “confusion” among possible candidates.

We assume that the result of an identification operation is the whole gallery ordered by distance from the probe, or a short list at least.

SRR



Cloud around the returned subject
"less crowded" =
More reliable response

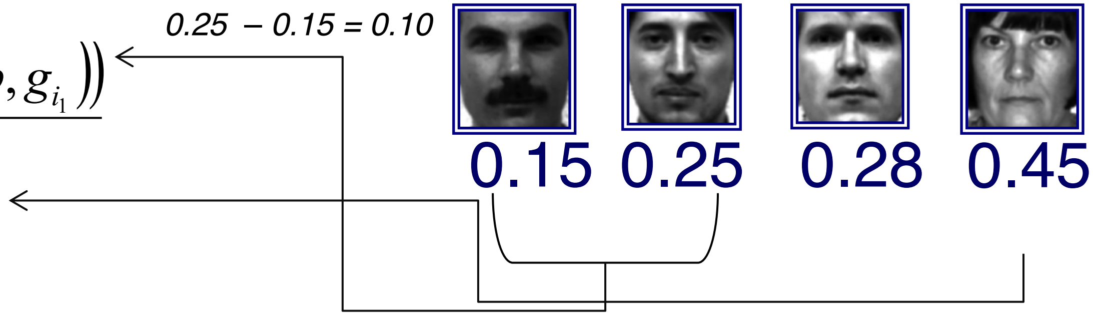


Cloud around the returned subject
"more crowded" =
Less reliable response

SRR

Given a probe *p* and a system *A* with gallery *G*, the ***relative distance*** is defined as:

$$\varphi(p)=\frac{F\left(d\left(p,g_{i_2}\right)\right)-F\left(d\left(p,g_{i_1}\right)\right)}{F\left(d\left(p,g_{i_{|G|}}\right)\right)}$$



The lower the difference in the numerator with respect to the denominator (the maximum computed difference with the probe), the higher the possible confusion related to the first two candidates, the lower the reliability.

SRR

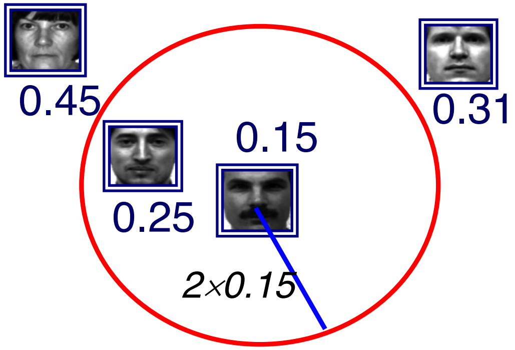
Given a probe *p* and a system *A* with gallery *G*, the ***density ratio*** is defined as:

$$\varphi(p) = 1 - |N_b| / |G|$$

With

$$N_b = \{g_{i_k} \in G \mid F(d(p, g_{i_k})) < 2 \cdot F((d(p, g_1)))\}$$

Both IN_bI and IGI are computed WITHOUT considering the element in the first list position in order to have a maximum value =1 (but this is not strictly necessary)



2×0.15 = 0.30

Density Ratio = 1 – 1/3
= 0.66 …

This function is less sensible to outliers, and in fact usually performs better than φ_1.

As a drawback, its definition takes to consider narrower clouds when the first retrieved identity is closer to the probe. On the contrary, a large distance takes to a larger cloud, which can be expected to be more crowded in any case. Any attempt to substitute 2 with an adaptable parameter did not achieve better results.

SRR

We need to identify a value fostering a correct separation between wrong rejections of enrolled subjects and wrong recognitions of not enrolled ones, both supported by the reliability value*.*

The critical φ_k is given by that value able to minimize the wrong estimates of function $\varphi(p)$, i.e. not enrolled subjects erroneously recognized (FA caused by a distance below the acceptance threshold or a similarity above) with $\varphi(p)$ higher than φ_k, or genuine subjects wrongly rejected (FR caused by a distance above the acceptance threshold or a similarity below) because recognized with $\varphi(p)$ lower than φ_k.

The distance between $\varphi(p)$ and φ_k is significant for reliability.

SRR

We also define as the width of the subinterval from to the proper extreme of the overall [0,1) interval of possible values, depending on the comparison between the current *φ(p)* and :

$$S\left(\varphi(p),\overline{\varphi}\right)=\begin{cases} 1-\overline{\varphi} & if \quad \varphi(p)>\overline{\varphi} \\ \overline{\varphi} & otherwise \end{cases}$$

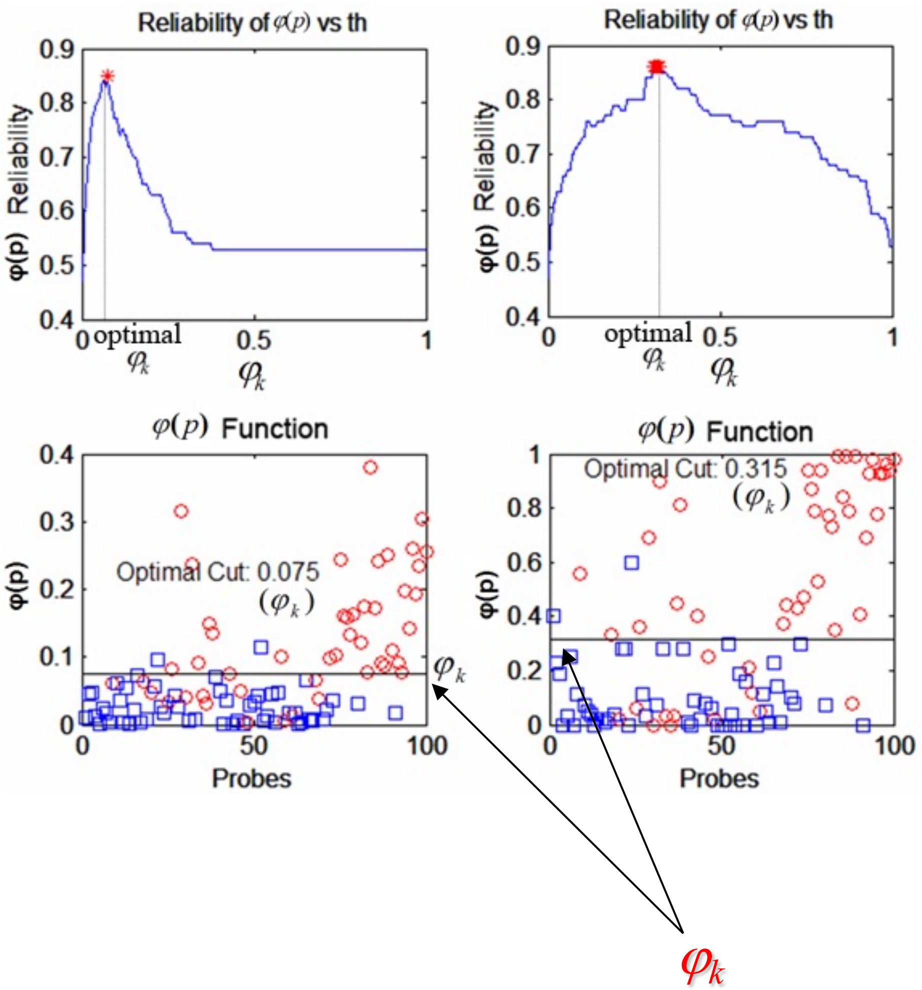
SRR index can finally be defined as:

$$\boxed{SRR=\left(\varphi(p)-\overline{\varphi}\right)\big/S\left(\overline{\varphi}\right)}$$

SRR

Density Ratio

Relative Distance



Red circles = GAs
Blue squares = FAs

Blue squares **above** the critical value = FAs **confirmed** by a high value of φ

Red circles **below** the critical value = GAs **not confirmed** due to a low value of φ

A threshold for SRR

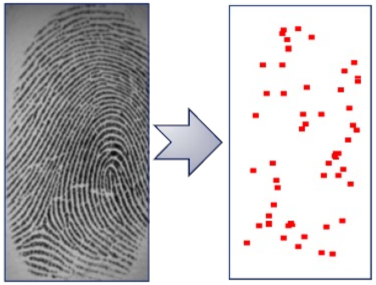
- The reliability threshold *th* can be automatically estimated by exploiting a certain number M of successive observations.

- We would desire to have a high average (the system is generally reliable) and a low variance (the system is stable)-

- We can summarize as:

$$th_i = \left| \frac{\mathrm{E}[\bar{S}_i]^2 - \sigma[\bar{S}_i]}{\mathrm{E}[\bar{S}_i]} \right|$$

Among possible solutions to increase quality/reliability: Template Updating



- Features extracted from a sample of a biometric trait, which are labeled with the individual’s identity, represent its *template.*
  - Matching exploits the template, not the sample
  - A template “should not allow ro reconstruct” a valid sample
  - Size aids codings and storing on more devices
  - Different template are generated any time the individual provides a biometric sample
- During operation of the recognition system much more biometric data become available, which were acquired over time. The system can use such data to *update the templates* in the gallery on a regular basis in order to address
  - Template ageing
  - Template enhancing

Among possible solutions: Template Updating

- Label assignment
  - Supervised systems
    - They require a supervisor to assign identity labels to newly acquired data during recognition system operation.
    - They usually work offline
  - Semi-supervised
    - They use the union of labeled and unlabeled data.
    - They work both online and offline.
- Most representative template selection to perform.
  - Online
    - Selection is performed as soon as new input data is acquired by the recognition system.
  - Offline
    - Selection is performed after a certain amount of data has been acquired during a specific time elapse.

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%2Fxpls%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

- Wang, Z., & Bovik, A. C. (2002). A universal image quality index. IEEE signal processing letters, 9(3), 81-84.

- 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