Multimodal Biometrics with Auxiliary Information

Quality, User-specific, Cohort information and beyond

Norman Poh

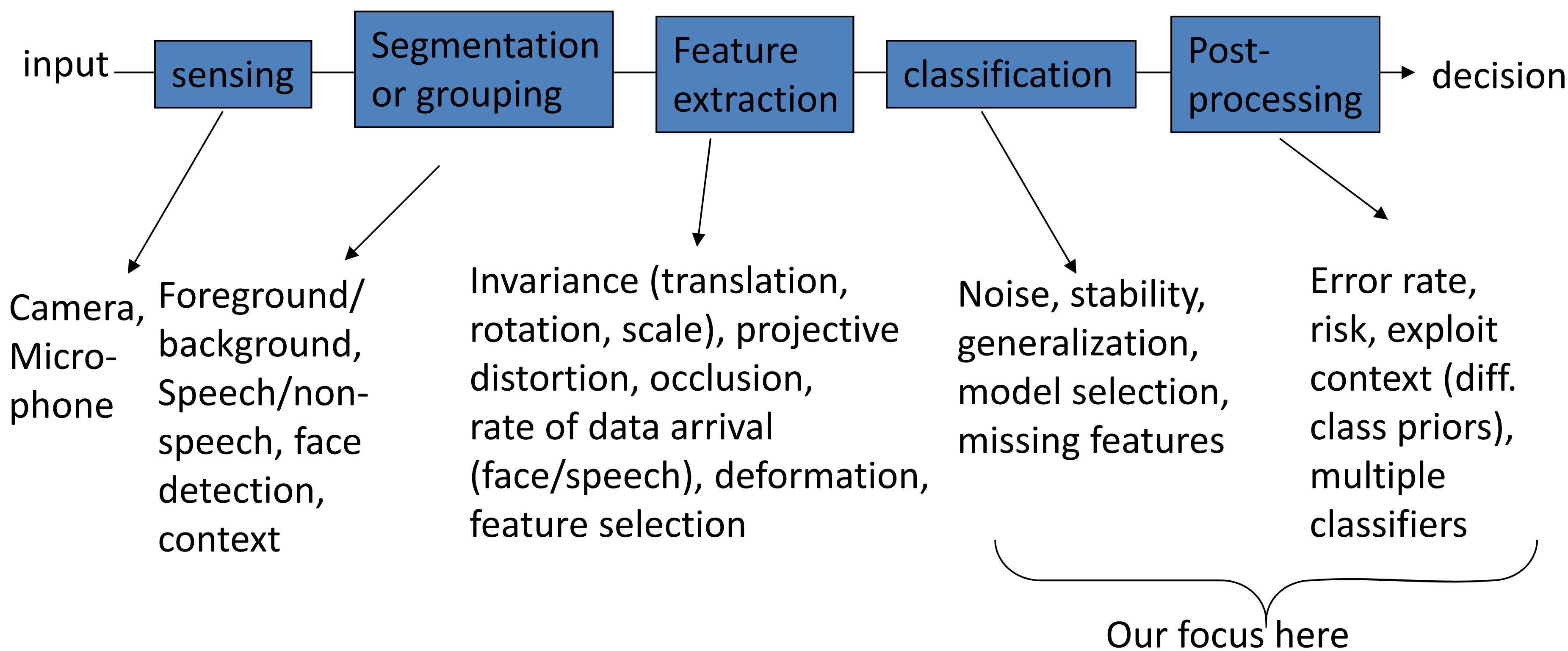
Talk Outline

- Part I: Bayesian classifiers and decision theory
- Part II: Sources of auxiliary information
  - Biometric sample quality
  - Cohort information
  - User-specific information
- Part III: Hetergoneous information fusion

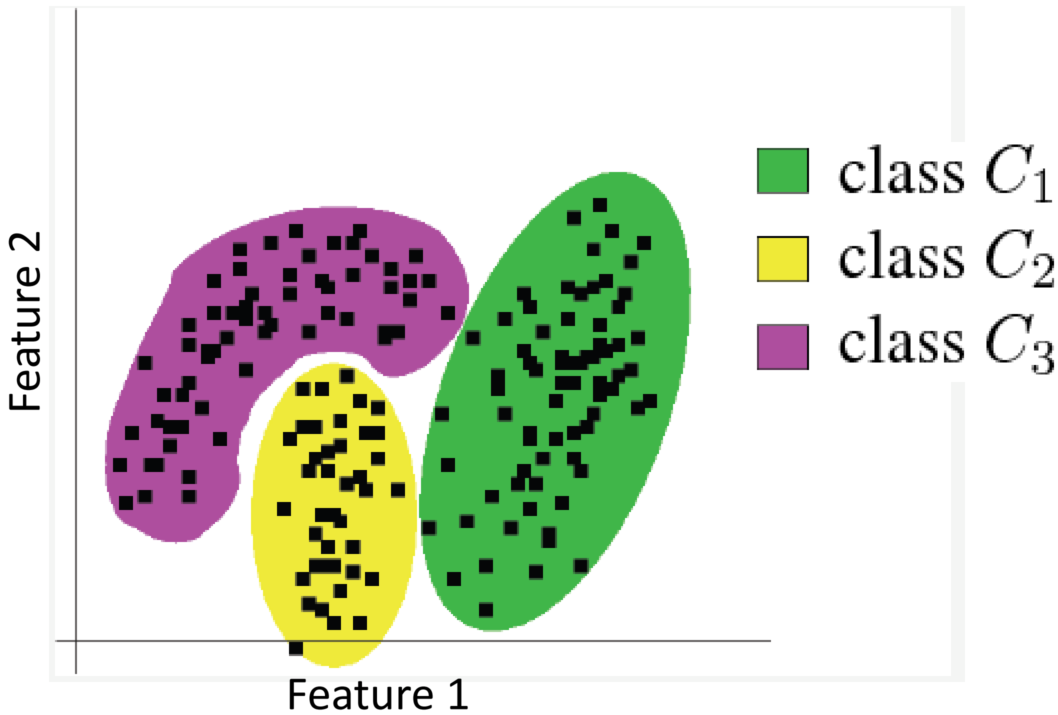
PART I

- Part I-A:
  - Bayesican classifier
  - Bayesian decision theory
  - Bayes error vs EER
- Part I-B:
  - Parametric form of error

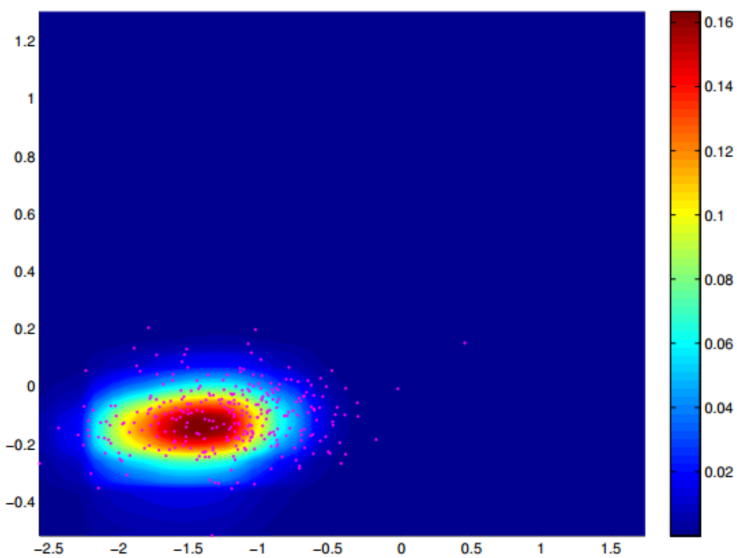
Part I-A: A pattern recognition system



Distribution of features

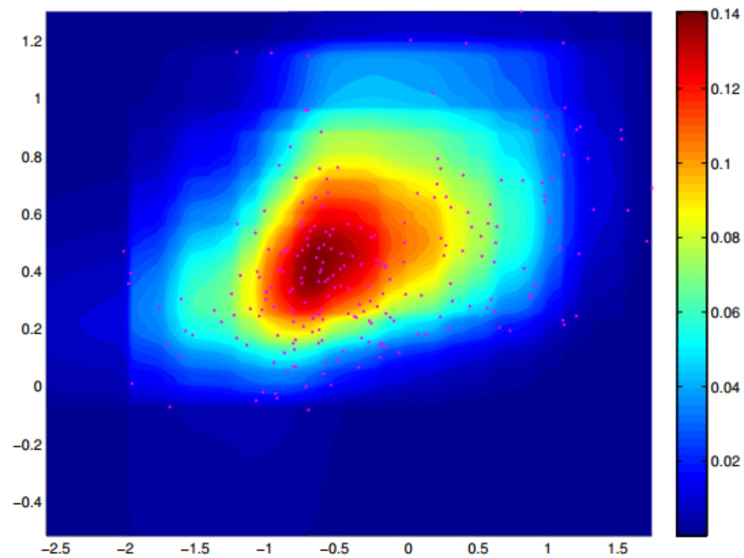


$p(E|H_1)$



The joint density of a negative class

$p(E|H_0)$

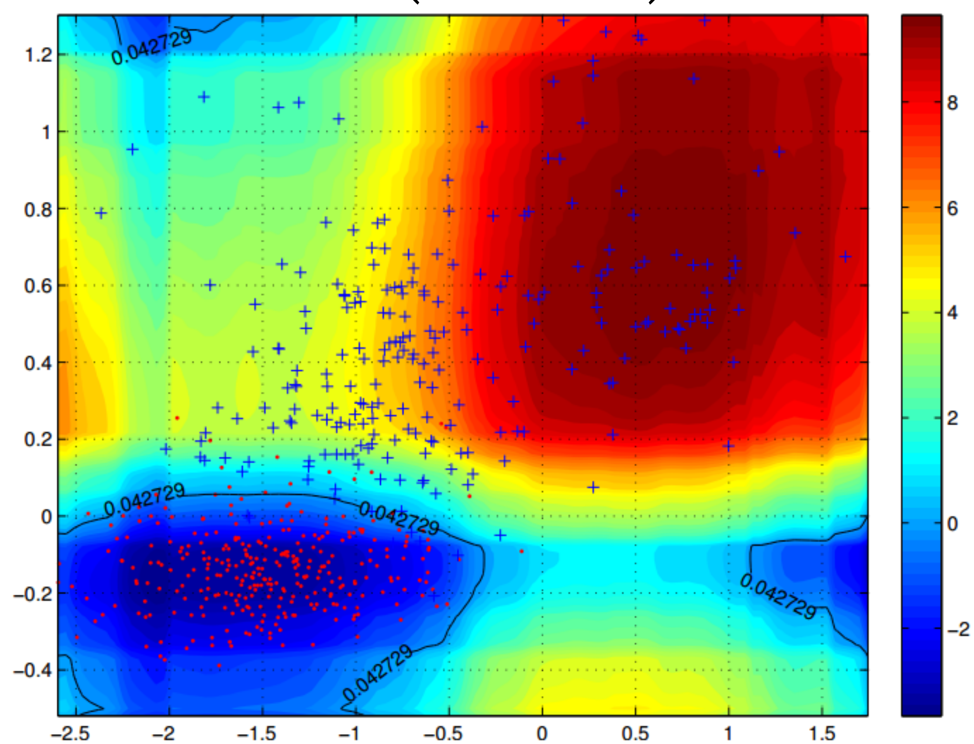


The joint density of a positive class

Log-likelihood map

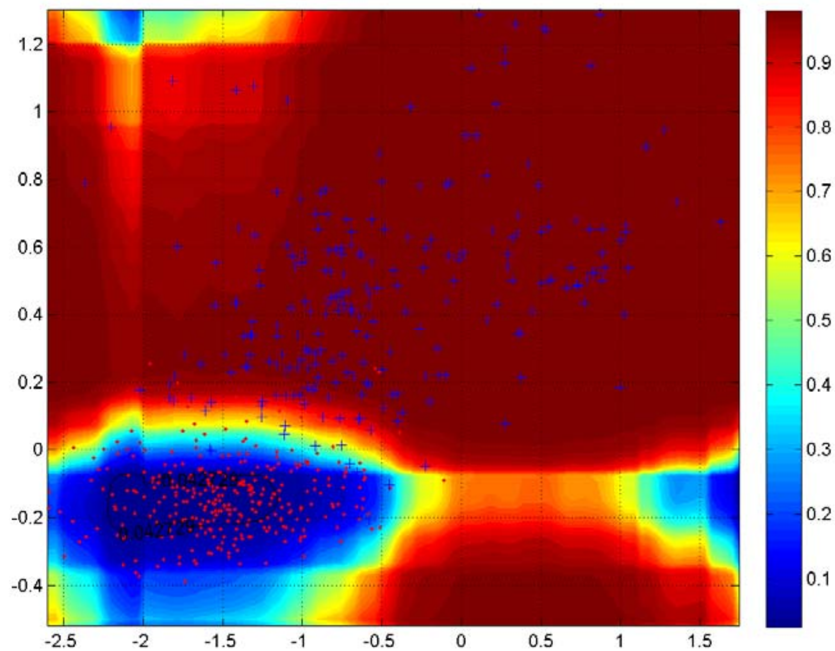
A possible decision boundary

$$\log\left(\frac{p(E|H_0)}{p(E|H_1)}\right)$$



Posterior probability map

$$P(H_o|E) = \left(\frac{p(E|H_0)P(H_0)}{\sum_{\omega} p(E|H_{\omega})P(H_{\omega})}\right)$$



What you need to know

- Sum rule: $\sum_a P(a,b) = P(b)$ (discrete) $\int_a p(a,b)\,da = p(b)$ (continuous)
- Product rule:

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

Important terms

Likelihood (density estimator), e.g., GMM, kernel density, histogram, "vector quantization"

Prior (probability table)

posterior

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})}$$

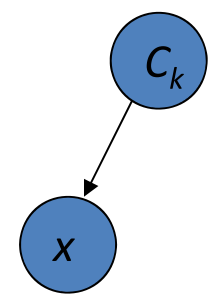
evidence

The most important lesson:

$$\boxed{\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}}$$

$\mathbf{x}$: Observation

$\mathcal{C}_k$: Class label



A graphical model (Bayesian network)

"equal (class) prior probability": 0.5 for client; 0.5 for impostor

Note: GMM representation is similar.

Building a Bayes Classifier

There are two variables: $\mathbf{x}$ and $\mathcal{C}_k$

We will use the Bayes (product) rule to relate their joint probability

$$p(\mathbf{x}, \mathcal{C}_k) = p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}) = P(\mathcal{C}_k|\mathbf{x})P(\mathbf{x})$$

The sum rule

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}, \mathcal{C}_k)$$

$$= \sum_k p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)$$

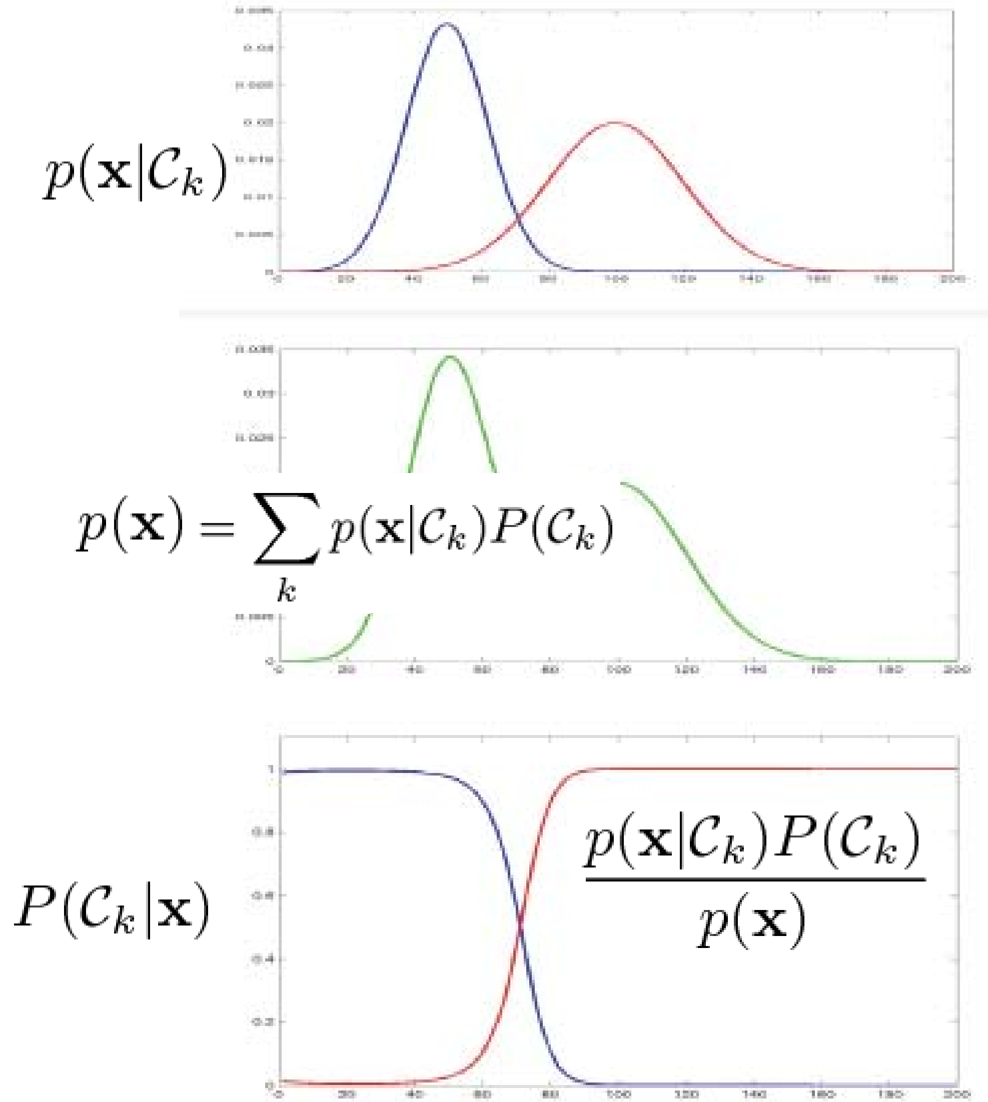
Rearranging, we get:

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})}$$

[Duda, Hart and Stork, 2001; PRML, Bishop 2005]

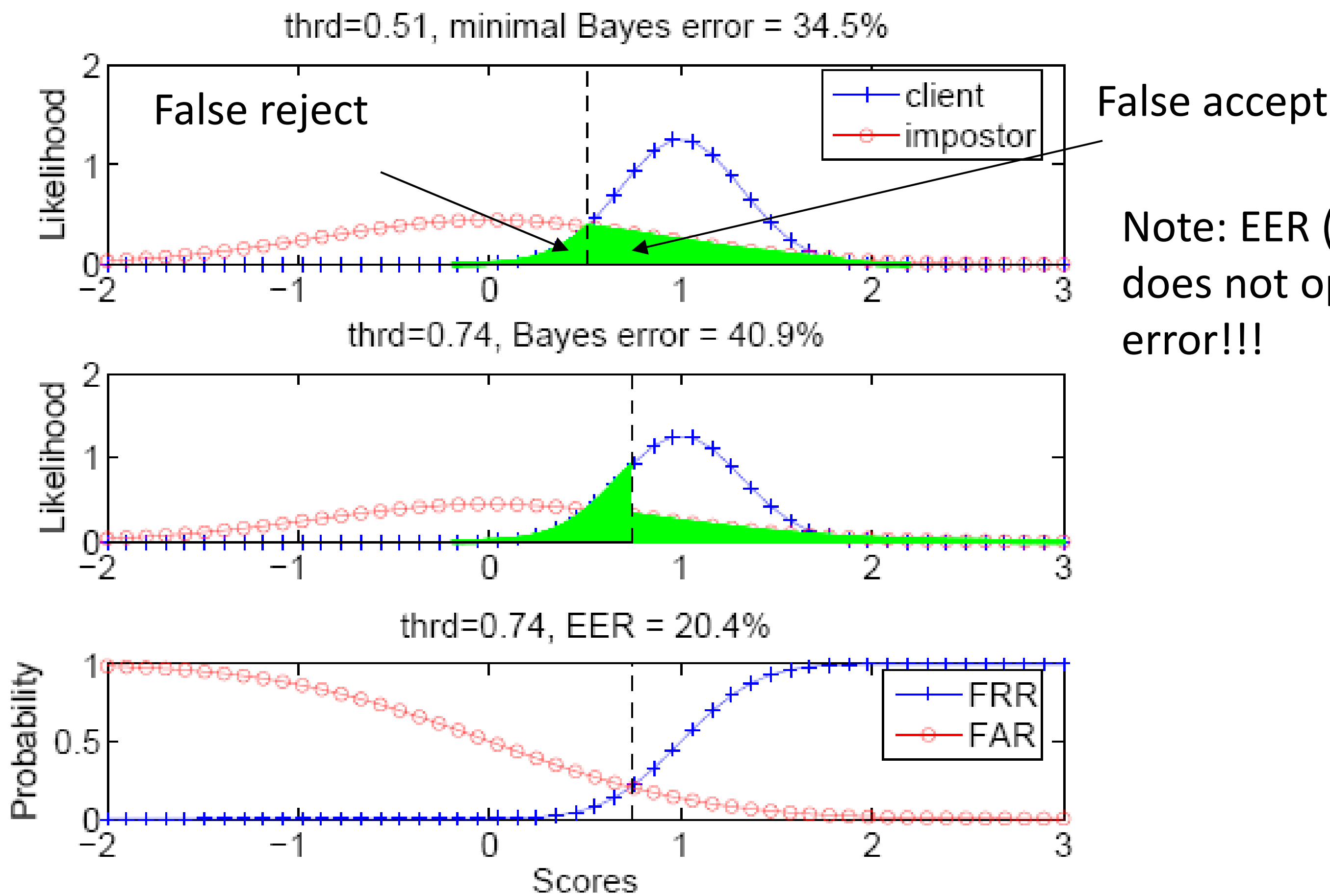
The sum/product rules are all you need to manipulate a Bayesian Network/graphical model

A plot of likelihoods, unconditional density (evidence) and posterior probability



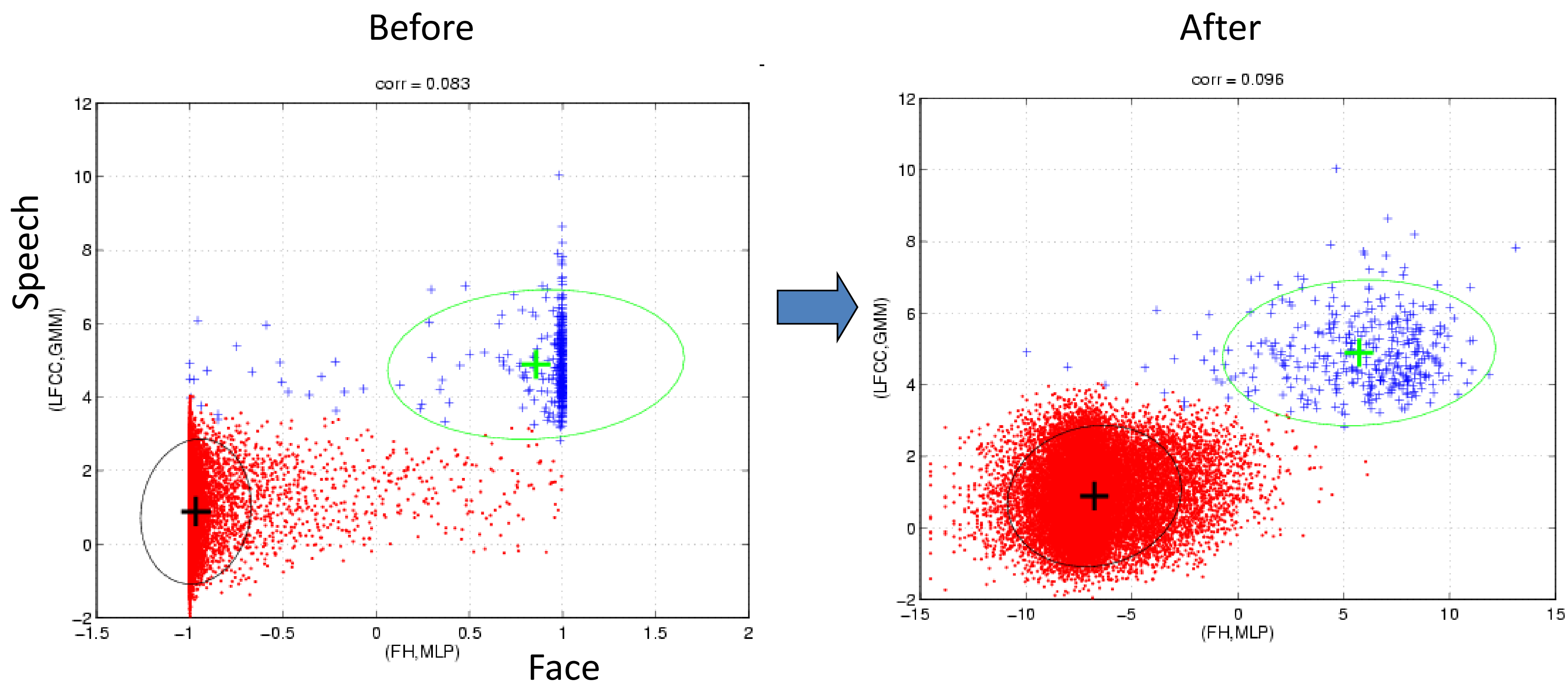
Minimal bayes error vs EER

What's the difference between the two?



Note: EER (Equal error rate) does not optimize the Bayes error!!!

Preprocess the matching scores



For this example, apply inverse tanh to the face output; in general, we can apply the "generalized logit transform" $y' = \log\left(\frac{y-a}{b-y}\right)$ y=[a,b]

Types of performance prediction

- Unimodal systems [our focus]
  - F-ratio, d-prime [ICASSP’04]
  - Client/user-specific error [BioSym’08]
- Multimodal systems [Skip]
  - F-ratio
    - Predict EER given a linear decision boundary [IEEE TSP’05]
  - Chernoff/Bhattacharya bounds
    - Upperbound the Bayes error (HTER) assuming a quadratic discriminant classifier [ICPR’08]

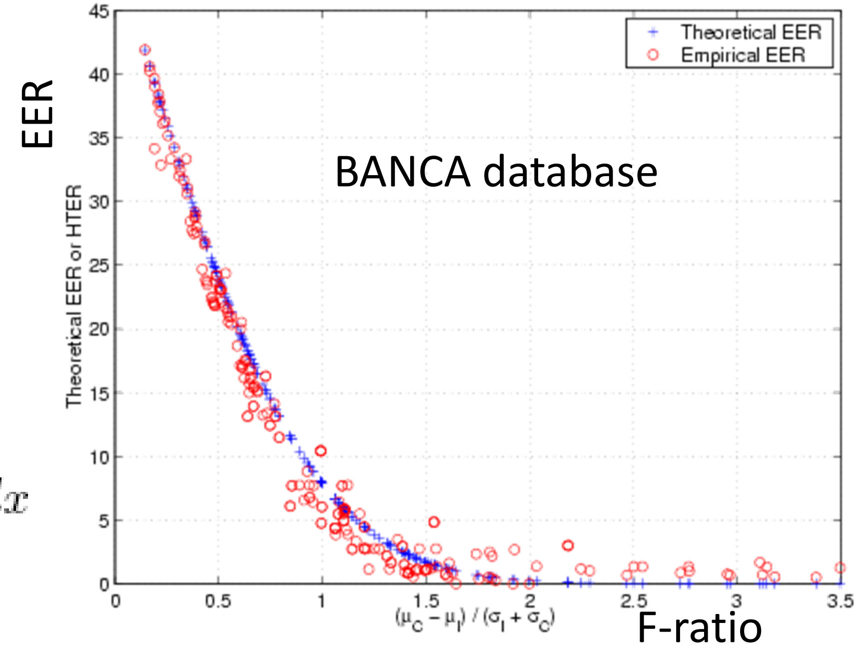
The F-ratio

- Compare the theoretical EER and the empirical one

$$\mathrm{EER}_j = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right)$$

$$\text{F-ratio} = \frac{\mu^{\mathrm{C}} - \mu^{\mathrm{I}}}{\sigma^{\mathrm{C}} + \sigma^{\mathrm{I}}}$$

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-x^2\right] dx$$



[Poh, IEEE Trans. SP, 2006]

Other measures of separability

$$\text{Fisher-ratio} = \frac{(\mu^{\mathtt{C}} - \mu^{\mathtt{I}})^2}{(\sigma^{\mathtt{C}})^2 + (\sigma^{\mathtt{I}})^2}$$

[Duda, Hart, Stork, 2001]

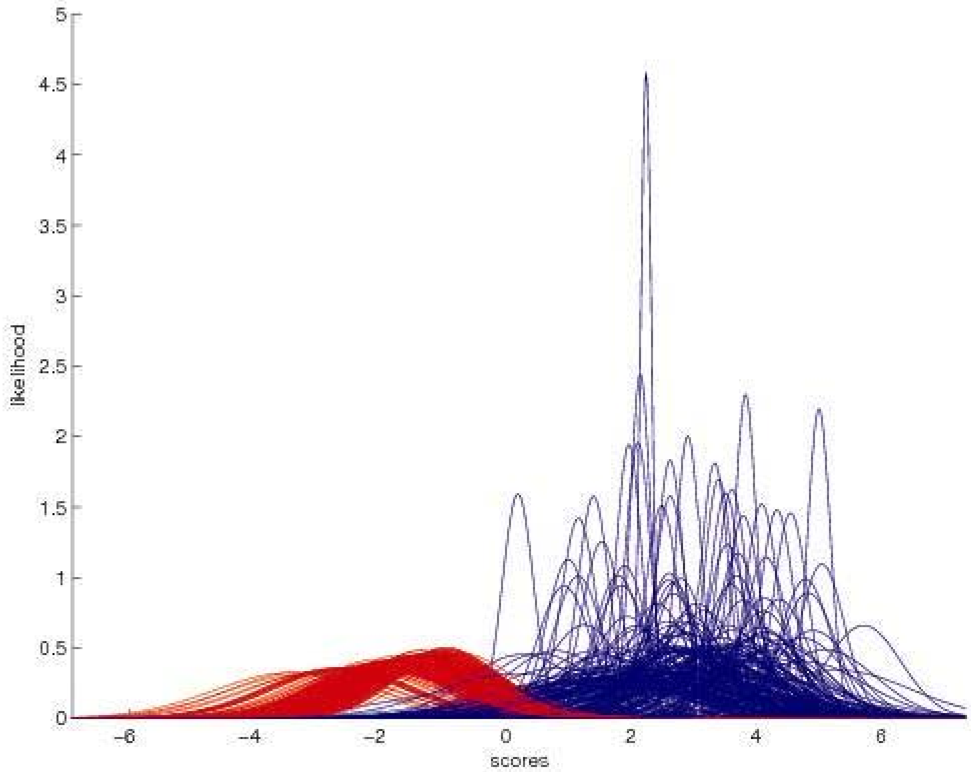
$$d' = \frac{|\mu^{\mathtt{C}} - \mu^{\mathtt{I}}|}{\sqrt{\frac{1}{2}(\sigma^{\mathtt{C}})^2 + \frac{1}{2}(\sigma^{\mathtt{I}})^2}}$$

[Daugman, 2000]

$$J_1 = \frac{\mu^{\mathtt{C}}}{\mu^{\mathtt{I}}},\, J_2 = \frac{(\mu^{\mathtt{C}} - \mu^{\mathtt{I}})^2}{\mu^{\mathtt{C}}\mu^{\mathtt{I}}},\ \text{and } J_3 = \frac{(\mu^{\mathtt{C}} - \mu^{\mathtt{I}})^2}{(\sigma^{\mathtt{C}})^2 + (\sigma^{\mathtt{I}})^2}$$

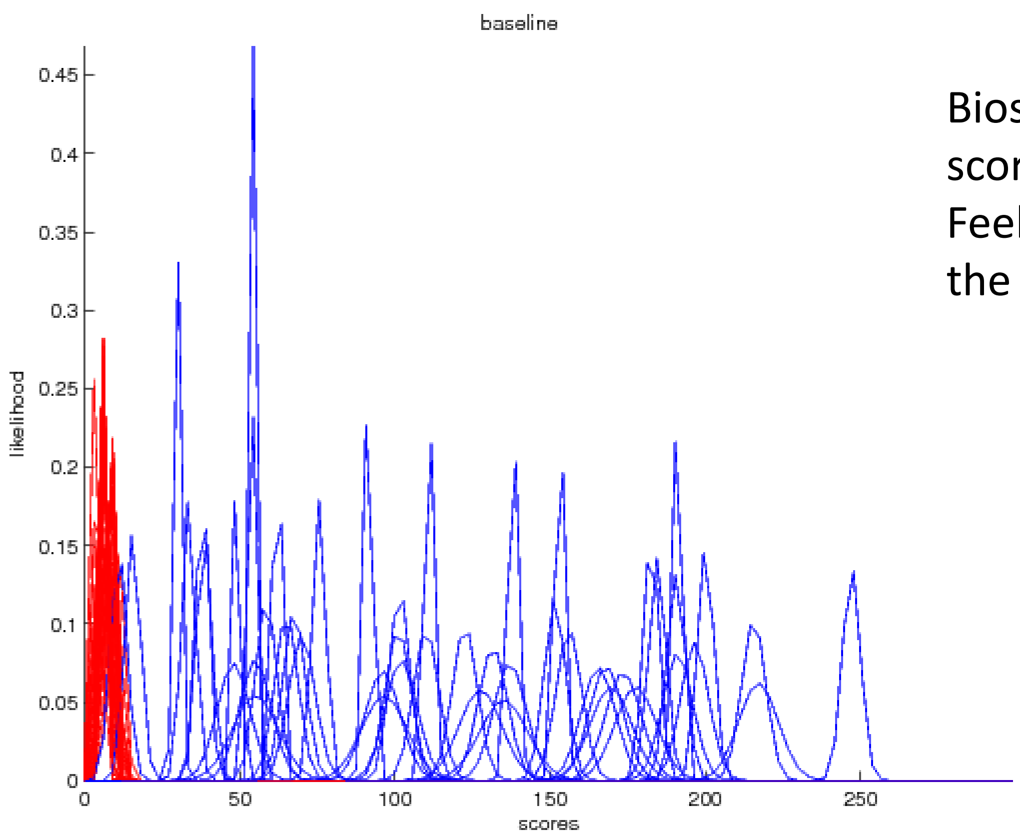
[Kumar and Zhang 2003]

Case study: face (and speech)



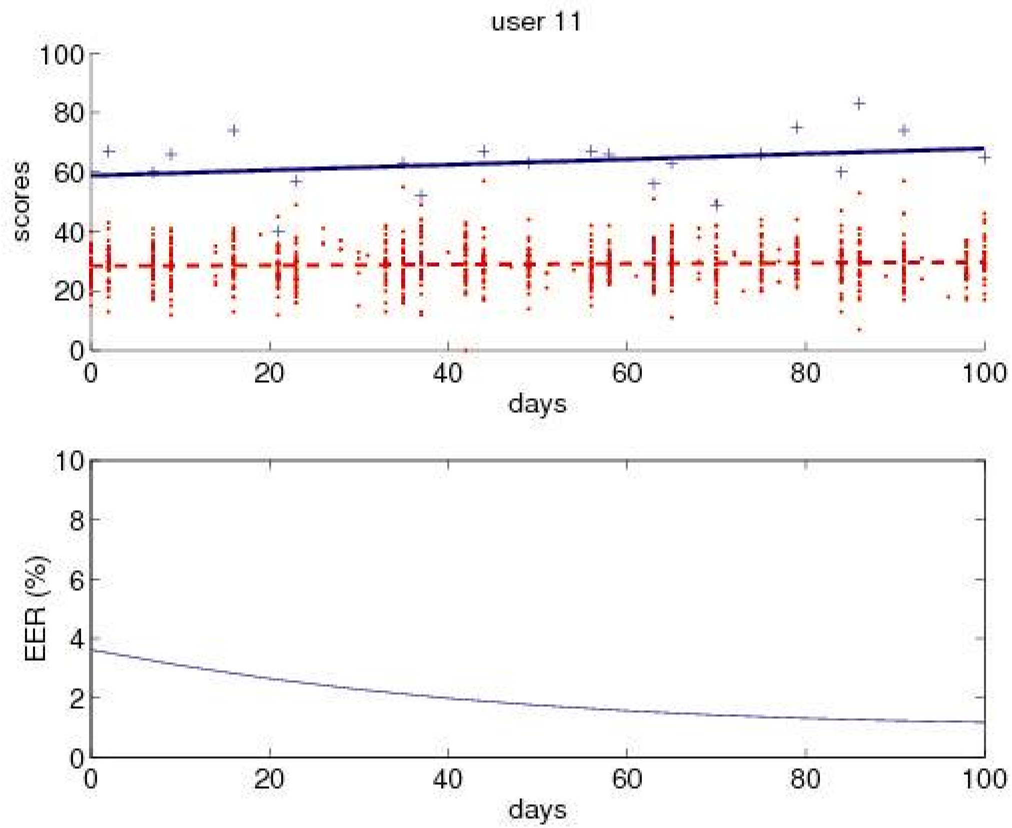
- XM2VTS face system (DCTmod2,GMM)
- 200 users
- 3 genuine scores per user
- 400 impostor scores per user

Case study: fingerprint



Biosecure DS2 score+quality data set. Feel free to download the scores

EER prediction over time



Inha university (Korea) fingerprint database

- 41 users
- Collected over one semester (aprox. 100 days)
- Look for sign of performance degradation over time

Part II: Sources of auxiliary information

- Motivation
- Part II-A : user-specific normalization
- Part II-B : Cohort normalization
- Part II-C : quality normalization
- Part II-D : combination of the different schemes above

Part II-A: Why biometric systems should be adaptive ?

- Each user (reference/target model) is different, I.e., every one is unique
  - → user/client-specific score normalization
  - → user/client-specific threshold
- Signal quality may change, due to
  - the user interaction
  - the environment
  - the sensor

Same [IEEE TASLP'08]

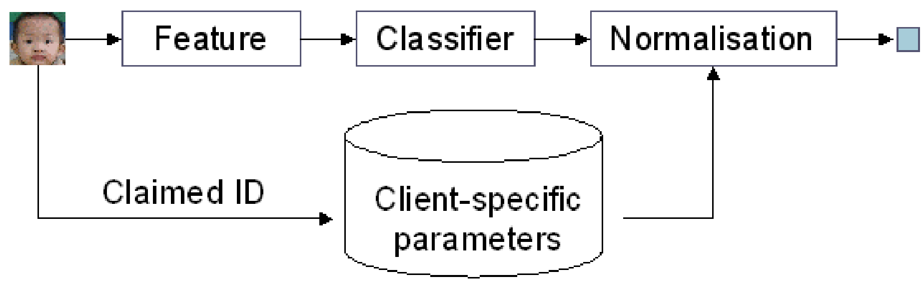
→Quality-based normalization

→Cohort-based normalization

- Biometric traits change [skip]
  - Eg, due to use of drugs and ageing
  - → semi-supervised learning (co-training/self-training)

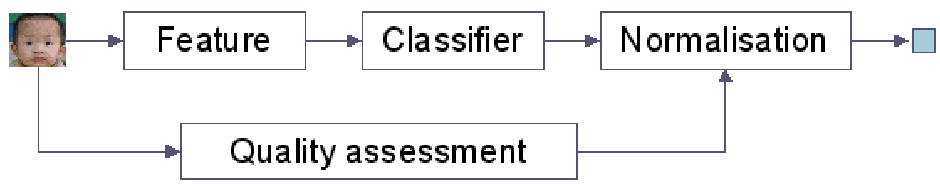
Information sources

Client/user-specific normalization (offline)



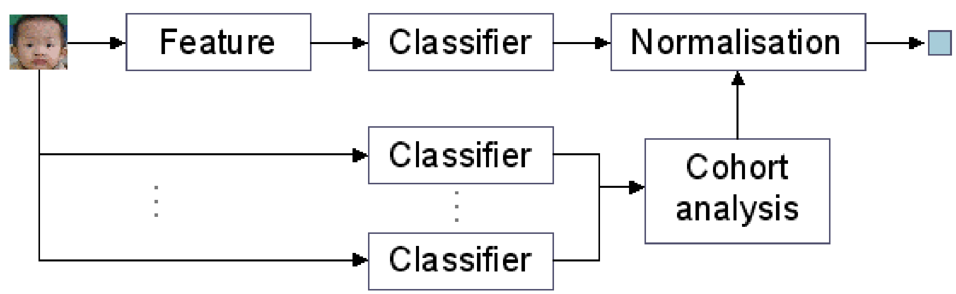
User-dependent score characteristics

Quality-based normalization



Changing signal quality

Cohort-based normalization (online)



Changing signal quality

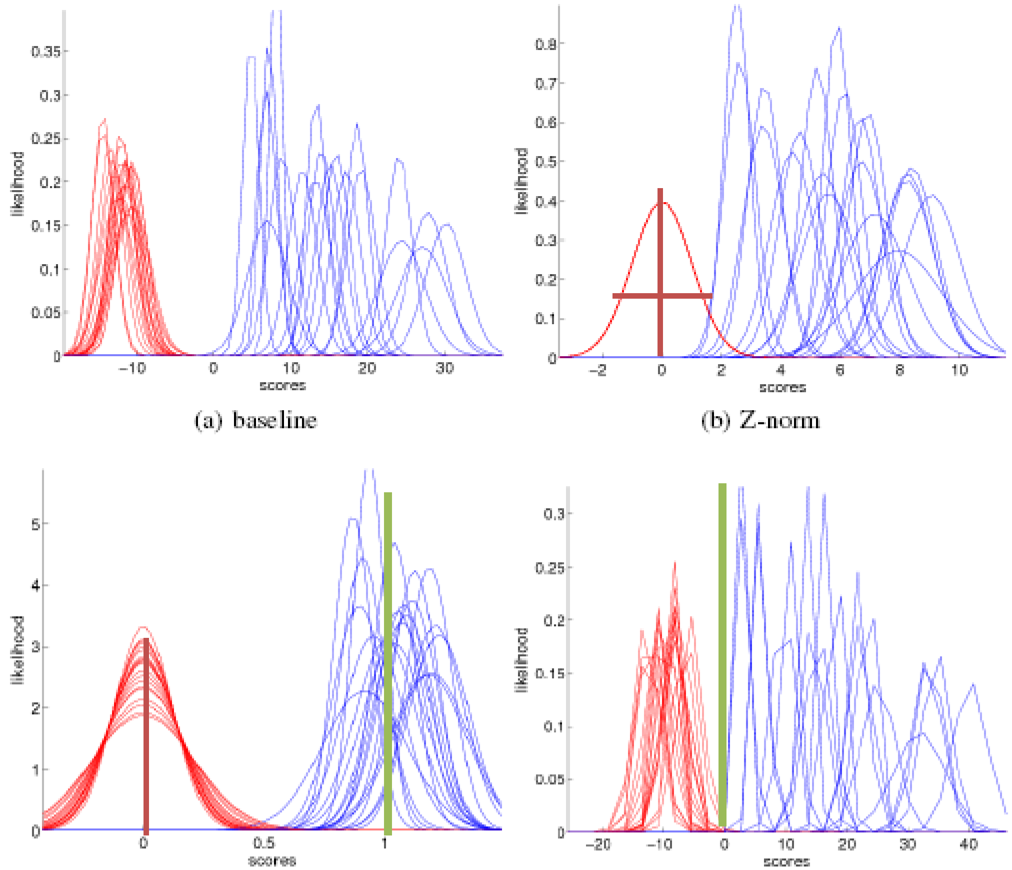
Part II-B: Effects of user-specific score normalization

Original matching scores

Z-norm

F-norm

Bayesian classifier (with log-likelihood ratio)



(a) baseline

(b) Z-norm

(c) F-norm

(d) likelihood ratio norm

The properties of user-specific score normalization

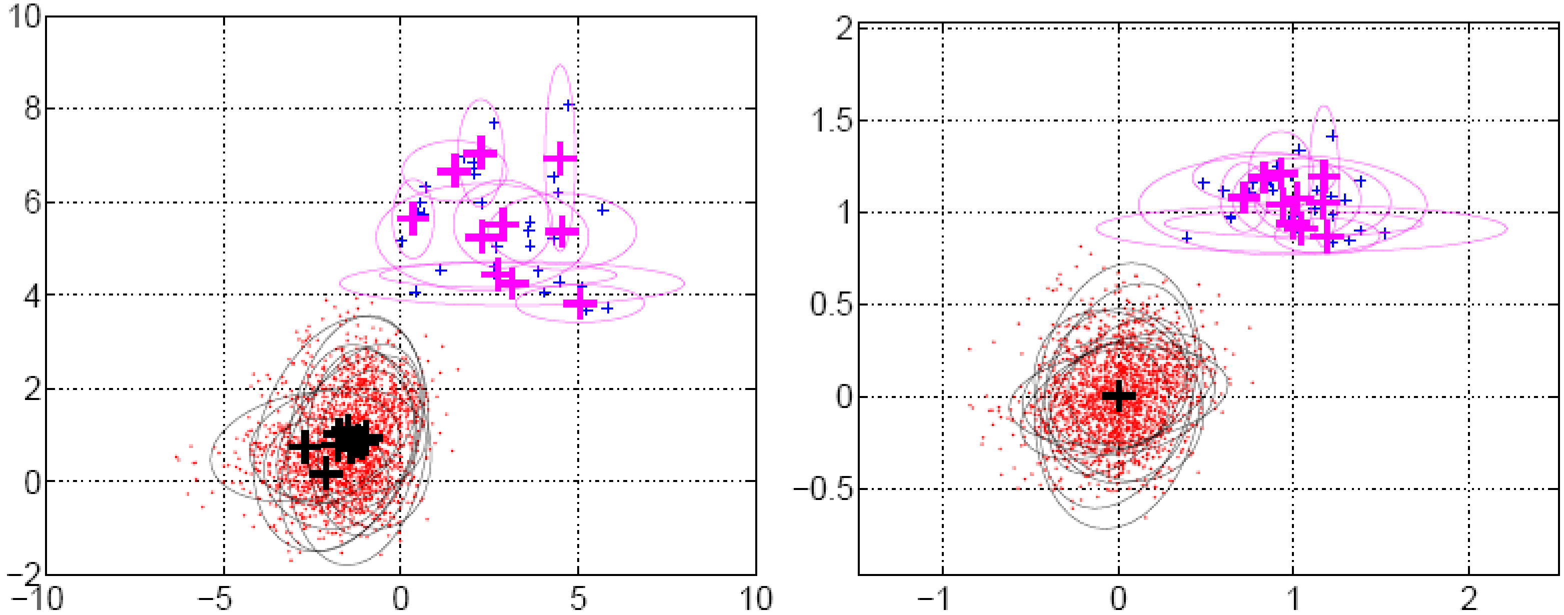
Procedures	Formulas	Properties
Z-norm	$y_j^Z = \frac{y-\mu_j^{\mathtt{I}}}{\sigma_j^{\mathtt{I}}}$	$E_j[y_j^Z \mid \mathtt{I}] = 0$ and $var_j[y_j^Z \mid \mathtt{I}] = 1$
F-norm	$y_j^F = \frac{y-\mu_j^{\mathtt{I}}}{\gamma\mu_j^{\mathtt{C}}+(1-\gamma)\mu^{\mathtt{C}}-\mu_j^{\mathtt{I}}}$	$E_j[y_j^F \mid \mathtt{I}] = 0$ and $E_j[y_j^F \mid \mathtt{C}] = 1$
EER-norm	$y^{EER} = y - \Delta_j$	$y_j^{EER} > 0$ is an optimal decision function (at EER) for all j
MS-LLR norm	$y^{llr} = \log \frac{p(y \mid \mathtt{C}, j)}{p(y \mid \mathtt{I}, j)}$	$y_j^{llr} > 0$ is an optimal decision function (at EER) for all j

$$\mu_j^{F,\mathtt{C}} \equiv E[y_j^F \mid \mathtt{C}] = \frac{E[y_j^{\mathtt{C}}] - \mu_j^{\mathtt{I}}}{\mu_j^{\mathtt{C}} - \mu_j^{\mathtt{I}}} = 1, \text{ for all } j$$

$$\mu_j^{F,\mathtt{I}} \equiv E[y_j^F \mid \mathtt{I}] = \frac{E[y_j^{\mathtt{I}}] - \mu_j^{\mathtt{I}}}{\mu_j^{\mathtt{C}} - \mu_j^{\mathtt{I}}} = 0, \text{ for all } j$$

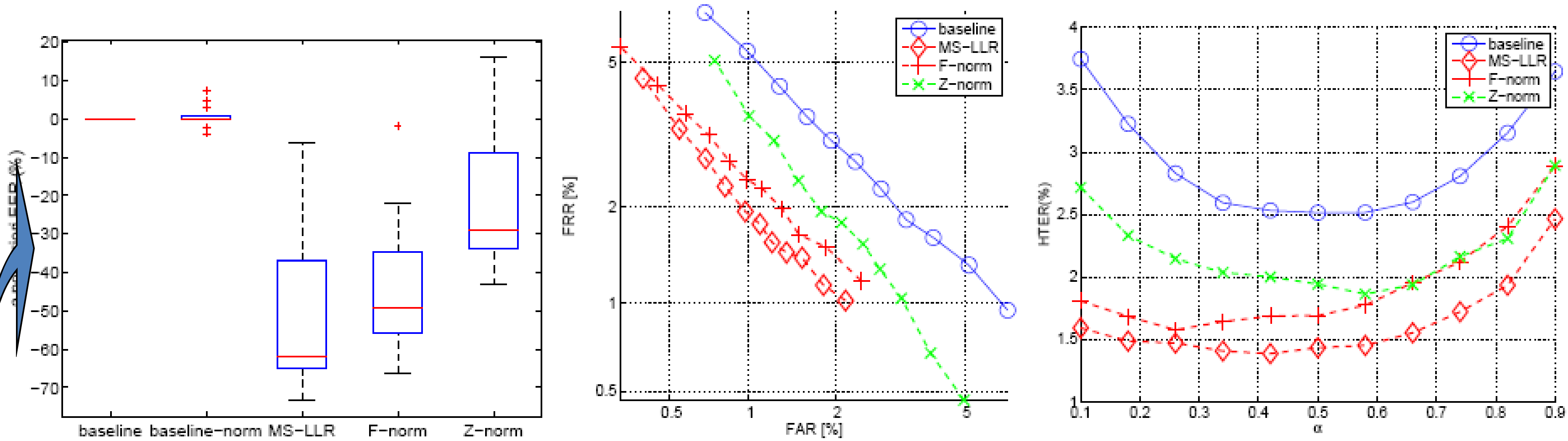
[IEEE TASLP'08]

User-specific score normalization for multi-system fusion



(a) Before F-norm (b) After F-norm

Results on the XM2VTS

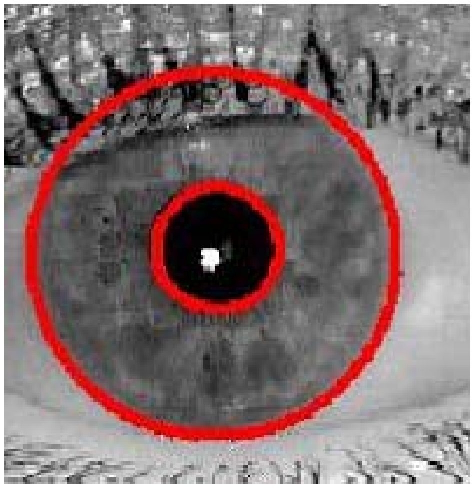


(a) relative change of EER (b) DET (c) EPC

$$\frac{\mathrm{EER}_{norm}}{\mathrm{EER}_{baseline}} - 1$$

1. EPC: expected performance curve
2. DET: decision error trade-off
3. Relative change of EER
4. Pooled DET curve

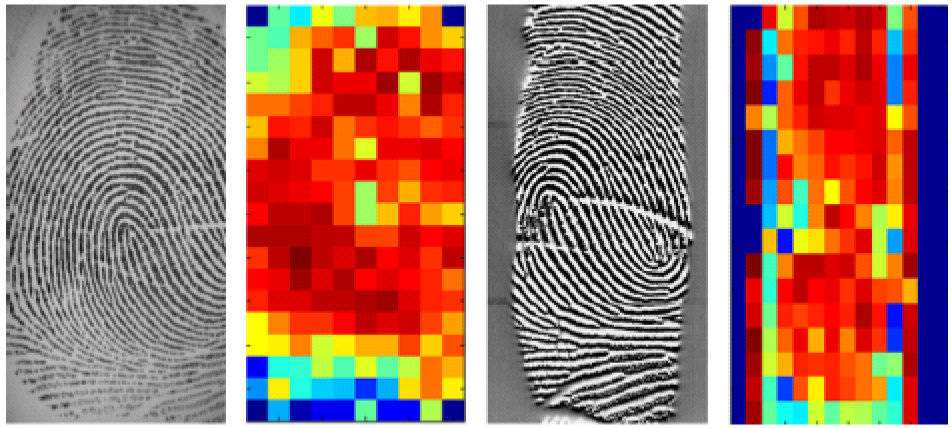
Part II-B: Biometric sample quality



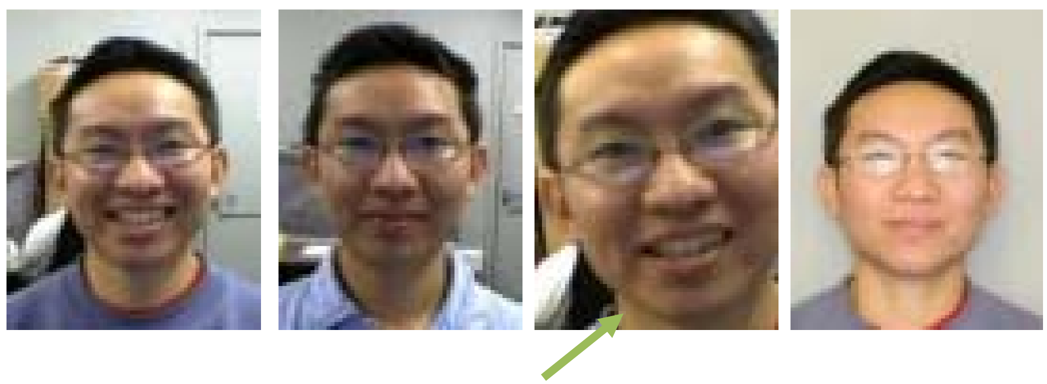
- What is a quality measure?
  - Information content
  - Predictor of system performance
  - Context measurements (clean vs noisy)
  - The definition we use: an array of measurements quantifying the degree of excellence or conformance of biometric samples to some predefined criteria known to influence the system performance
    - The definition is *algorithm-dependent*
    - Comes from the *prior knowledge* of the system designer
- Can quality predict the system performance?
- How to incorporate quality into an existing system?

Measuring "quality"

Optical sensor Thermal sensor



[Biosecure] an EU-funded project



Quality measure is system-dependent. If a module (face detection) fails to segment a sample or a matching module produces lower matching score (a smiley face vs neutral face), then the sample quality is low, even though we have no problem recognizing the face.

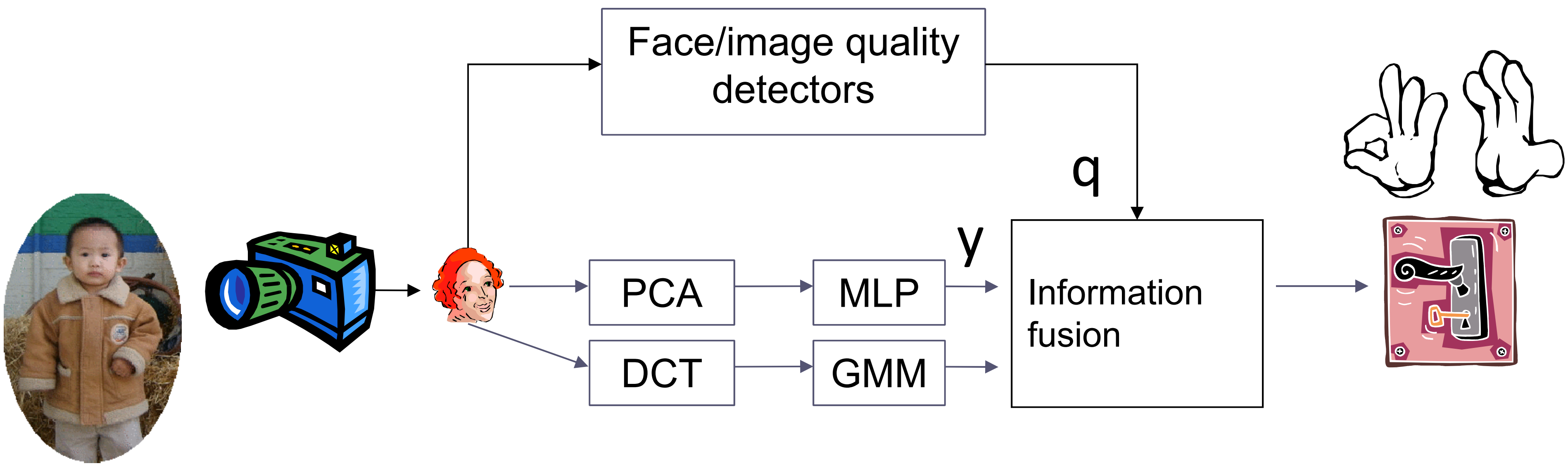
There is a still a gap between subjective quality assessment (human judgement) vs the objective one.

Face quality measures

- Face
  - Frontal quality
  - Illumination
  - Rotation
  - Reflection
  - Spatial resolution
  - Bit per pixel
  - Focus
  - Brightness
  - Background uniformity
  - Glasses

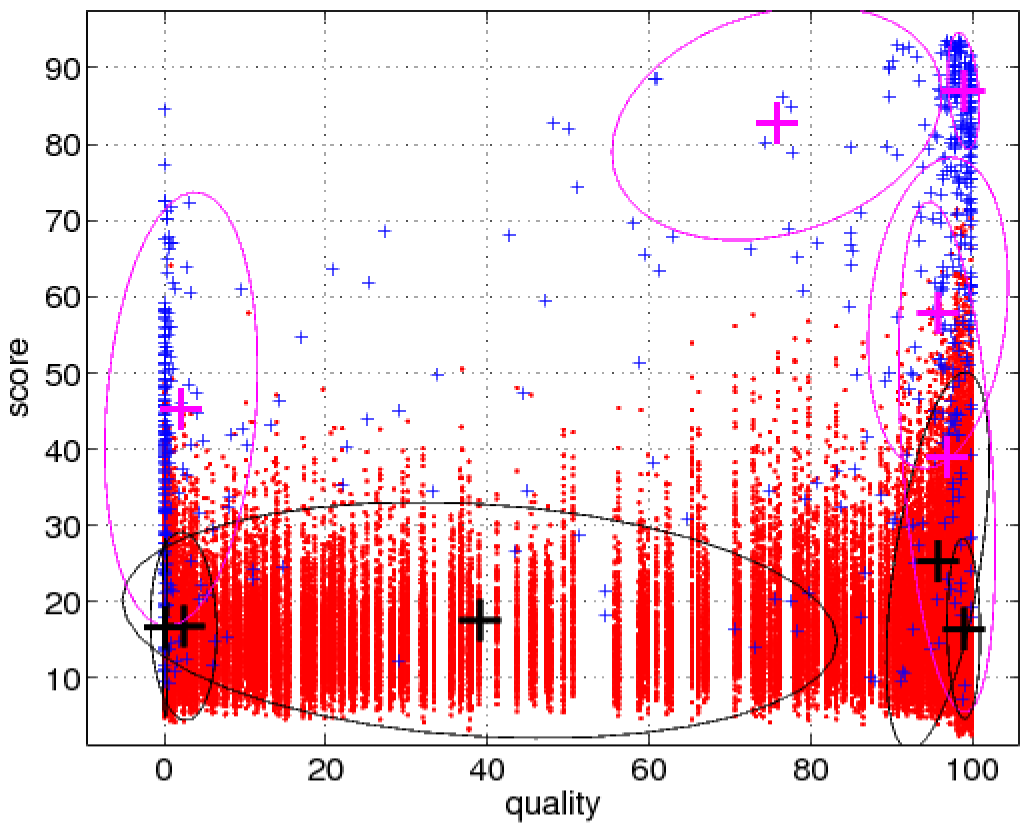
Well illuminated	Side illuminated
	
Glass=89%	Glass=15%
Illum.=100%	Illum=56%

Enhancing a system with quality measures



Build a classifier with [y,q] as observations
Problem: q is not discriminative and worse, it's dimension can be large for a given modality

How do (y,q) look like?



Strong correlation for the genuine class

Weak correlation for the impostor class

p(y,q|k)

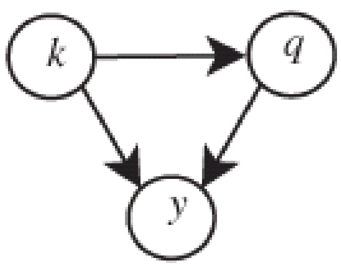
A learning problem

Approach 1

- train a classifier with [y,q]

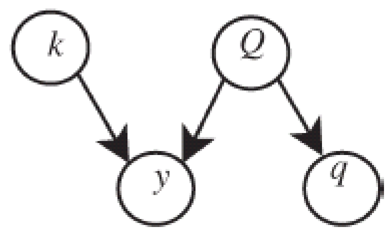
Approach 2

- cluster q into Q clusters. For each cluster, train a classifier using [y] as observations



Feature-based

p(y,q,k)p(q|k)=p(y,q|k)



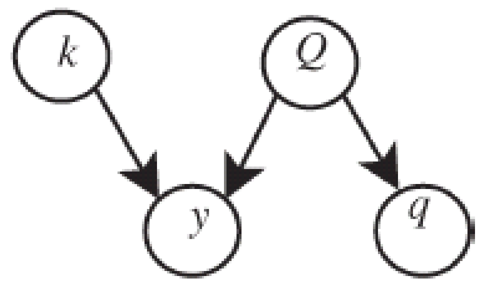
Cluster-based

p(y|k,Q)

p(q|Q)

y: score
q: quality measures
Q: quality cluster
k: class label

A note



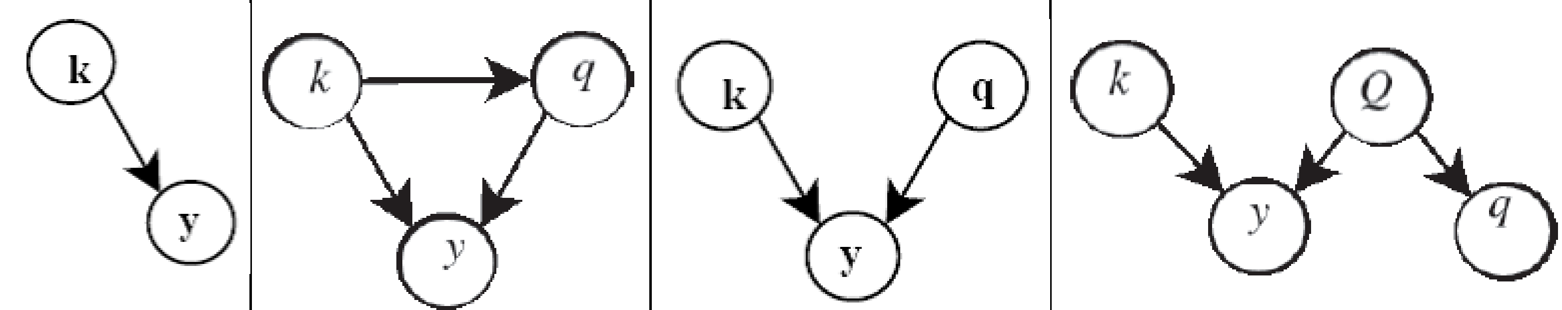
- If we know Q, the learning the parameters becomes straight forward:
  - Divide q into a number of clusters
  - For each cluster Q, learn p(y|k,Q)

Details [skip]

$k \in \{\mathtt{C}, \mathtt{I}\}$	Class label (*unobserved* in test)
$y \in \mathbb{R}^N$	Vector of scores (could be a scalar)
$q \in \mathbb{R}^{\mathtt{N}_q}$	Vector of quality measures
$Q \in \{1, \ldots, \mathtt{N}_Q\}$	Quality states (*unobserved* in test)

[IEEE T SMCA'10]

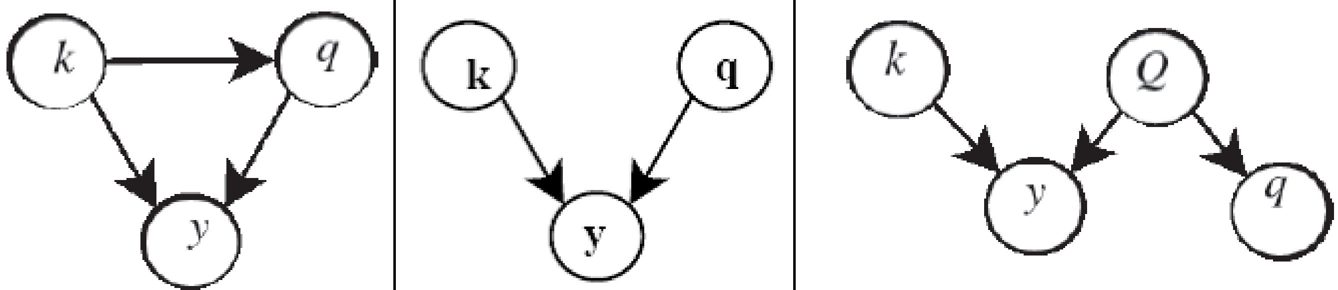
Models



Conditional densities	$p(y\|k)$	$p(y\|k,q)p(q\|k) = p(y,q\|k)$	$p(y\|k,q)$	$p(y\|k,q) = \sum_Q p(y\|k,Q)P(Q\|q)$

$$P(Q|\mathbf{q}) = \frac{p(\mathbf{q}|Q)P(Q)}{p(\mathbf{q})} \qquad p(\mathbf{q}) = \sum_Q p(\mathbf{q}|Q)P(Q)$$

Details [skip]



$$p(y, q|k)$$

$$p(y|k, q)$$

$$p(y|k, q) = \sum_{Q} p(y|k, Q)P(Q|q)$$

$$P(k|\mathbf{y}, \mathbf{q}) = \frac{p(\mathbf{y}, \mathbf{q}|k)P(k)}{\sum_{k'} p(\mathbf{y}, \mathbf{q}|k')P(k')}$$

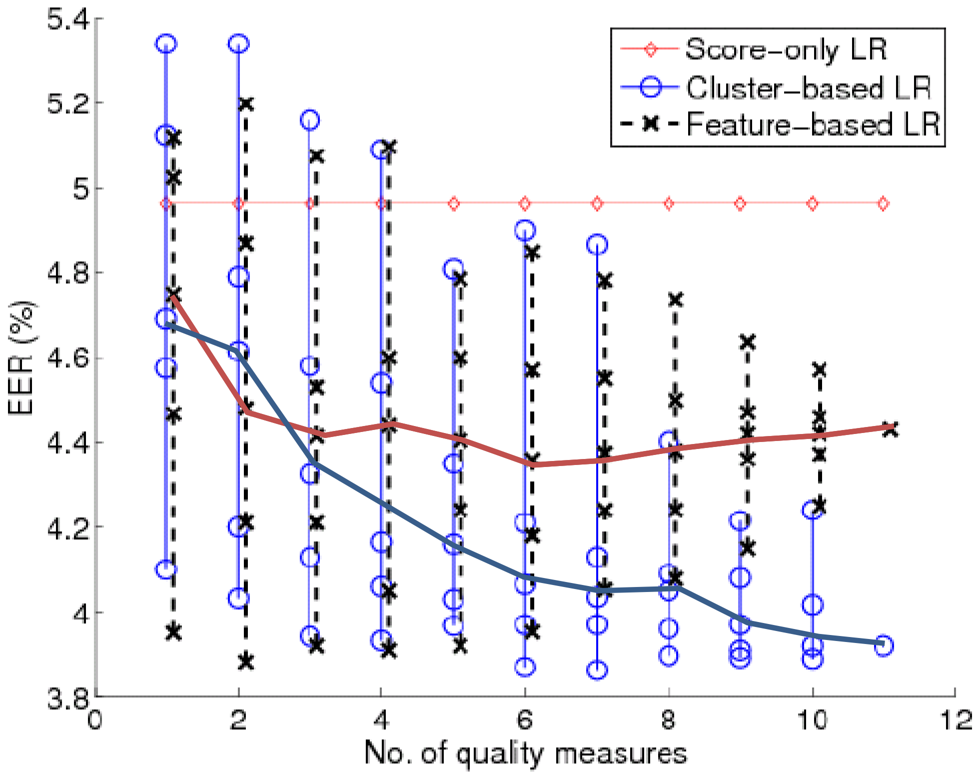
?

$$P(k|\mathbf{y}, \mathbf{q}) = \frac{p(\mathbf{y}|\mathbf{q}, k)P(k)}{\sum_{k'} p(\mathbf{y}|\mathbf{q}, k')P(k')}$$

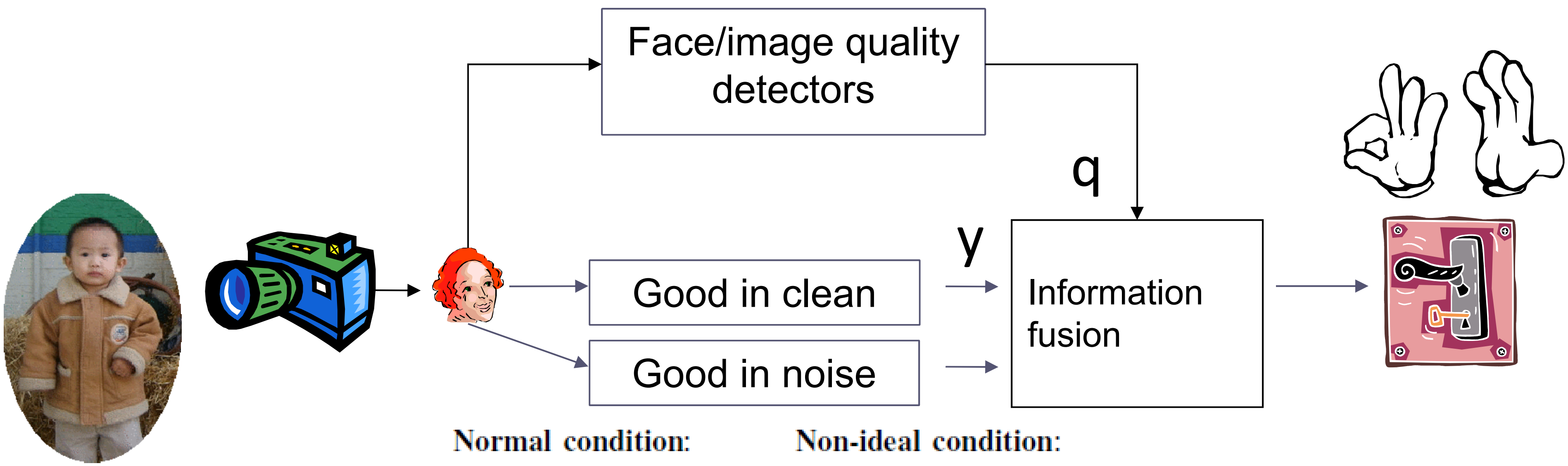
This is nothing but a Bayesian classifier taking y and q as observations

We just apply the Bayes rule here!

Effect of large dimensions in *q*



Exploit diversity of experts competency in fusion



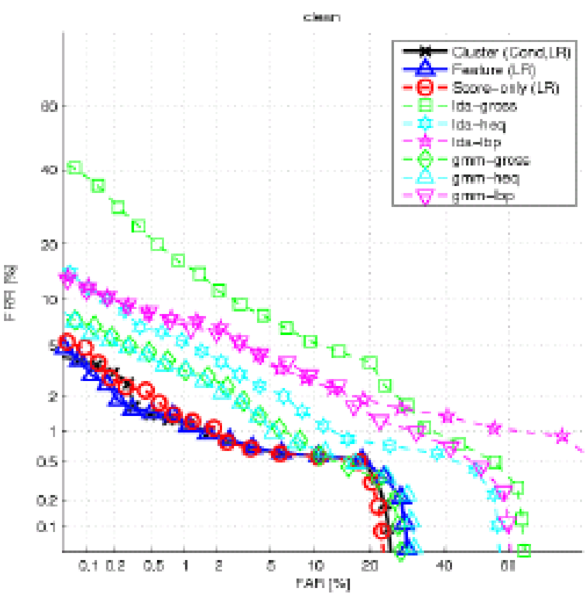
Normal condition:

1) gmm-heq
2) gmm-gross
3) lda-heq
4) lda-lbp
5) gmm-lbp
6) lda-gross

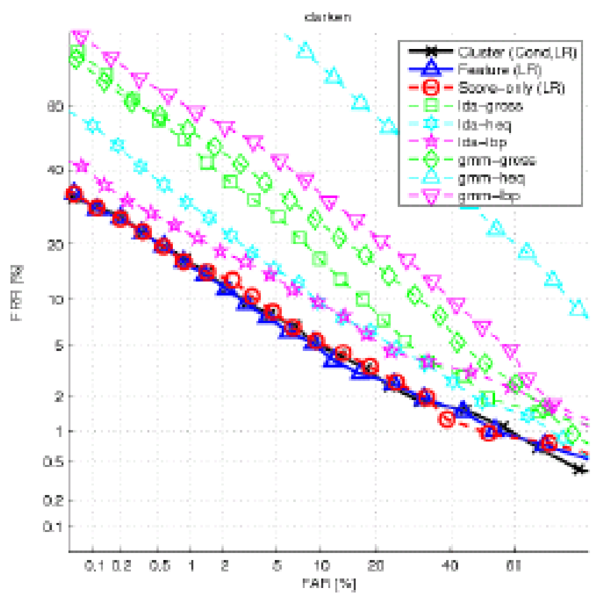
Non-ideal condition:

1) lda-lbp
2) lda-heq
3) lda-gross
4) gmm-gross
5) gmm-lbp
6) gmm-heq

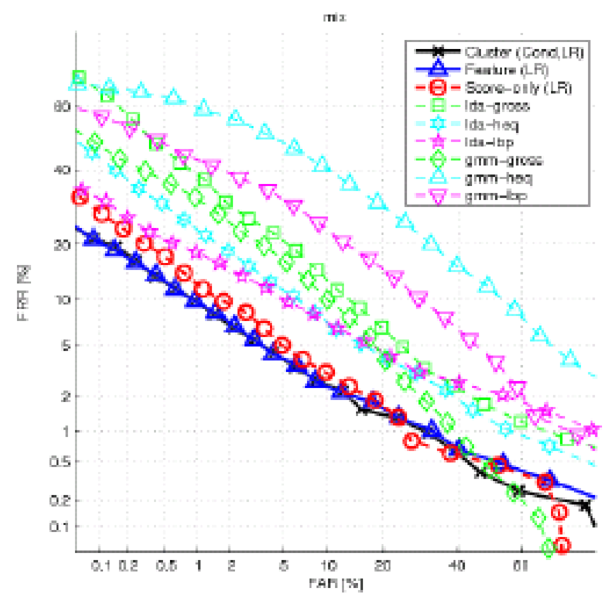
Experimental evidence



clean



noisy



mixed=clean+noisy

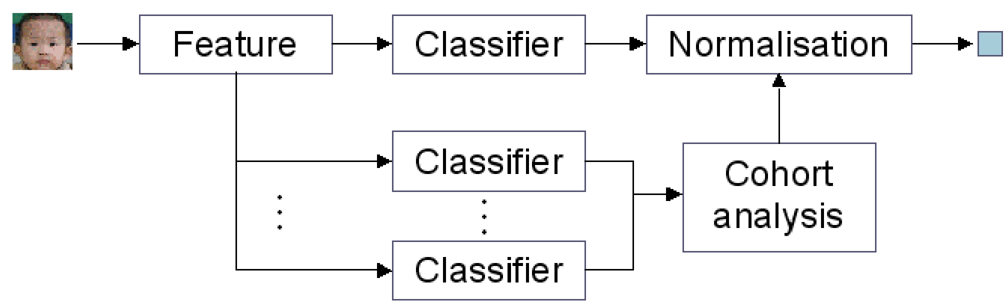
Part II-C: Cohort normalization

- T-norm – a well-established method, commonly used in speaker verification
- Impostor scores parameters are computed online for each query (computationally expensive) and at the same time adaptive to test access

$$y_T = \frac{y - \mu^c}{\sigma^c}$$

$$\mu^c = \mathrm{E}[y]$$

$$(\sigma^c)^2 = \mathrm{E}\left[(y^c - \mu^c)^2\right]$$



Other Cohort-based Normalisation

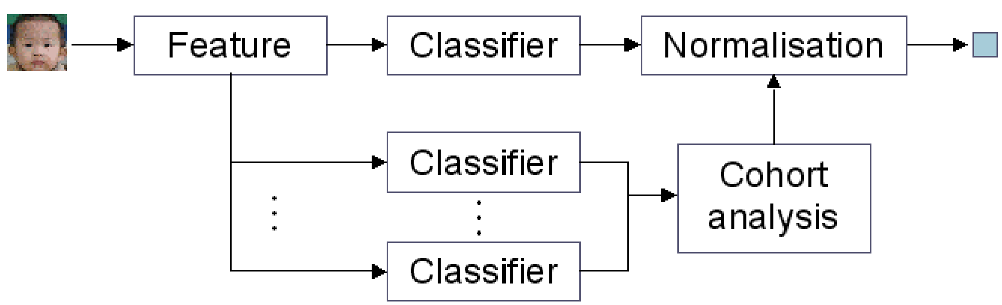
- Tulyakov's approach

$$y_{Tul} = P(\mathrm{C} | y, \max_{y^c \in \mathscr{Y}^c} \{y^c\}),$$

A probability function estimated using logistic regression or neural network

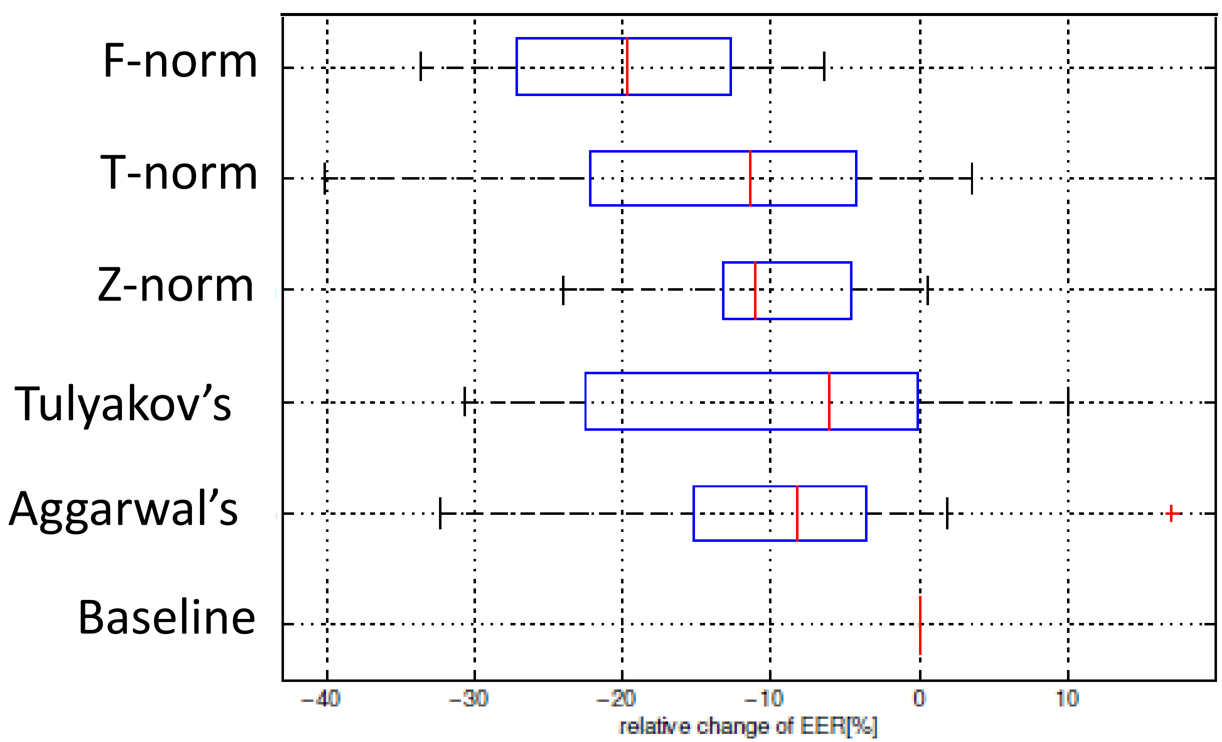
- Aggrawal's approach

$$y_{Ag} = \frac{y}{\max_{y^c \in \mathscr{Y}^c} \{y^c\}}$$



Comparison of different schemes

Box plot of relative change of EER for fingerprint modality over 2 devices and 6 fingers



Biosecure DS2
6 fingers x 2 devices

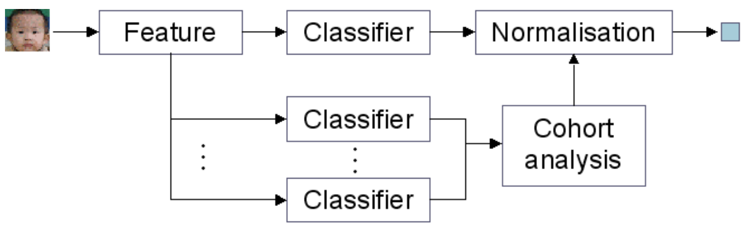
[BTAS'09]

$$\frac{EER_{new} - EER_{bline}}{EER_{bline}}$$

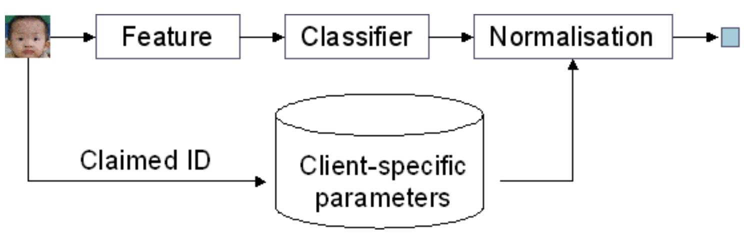
Part II-D: Combination of different information sources

- Cohort, client-specific and quality information is not mutually exclusive
- We will show the benefits of:
  - Case I: Cohort+client-specific information
  - Case II: Cohort+quality information

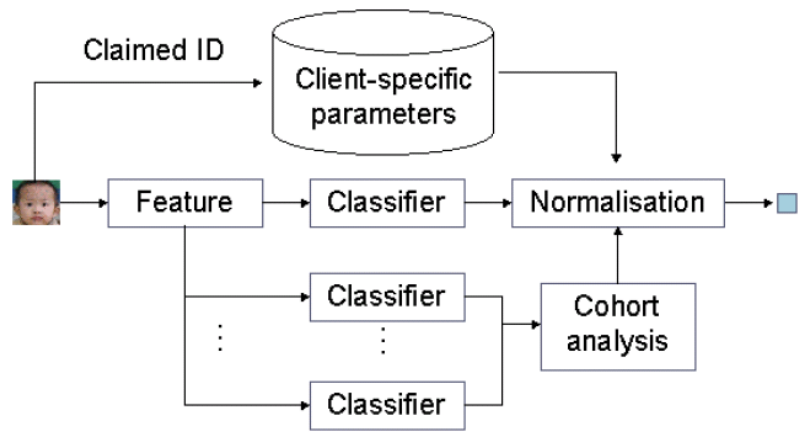
Case I: A client-specific+cohort normalization



Cohort normalization



Client-specific normalization



An example: Adaptive F-norm

Our proposal is to combine these two pieces of information, called, Adaptive F-norm:

- It uses cohort scores
- And user-specific parameters

$$y_j^{AF} = \frac{y - \mu^{\mathtt{c}}}{\tilde{\mu}(\gamma, j) - \mu^{\mathtt{c}}}$$

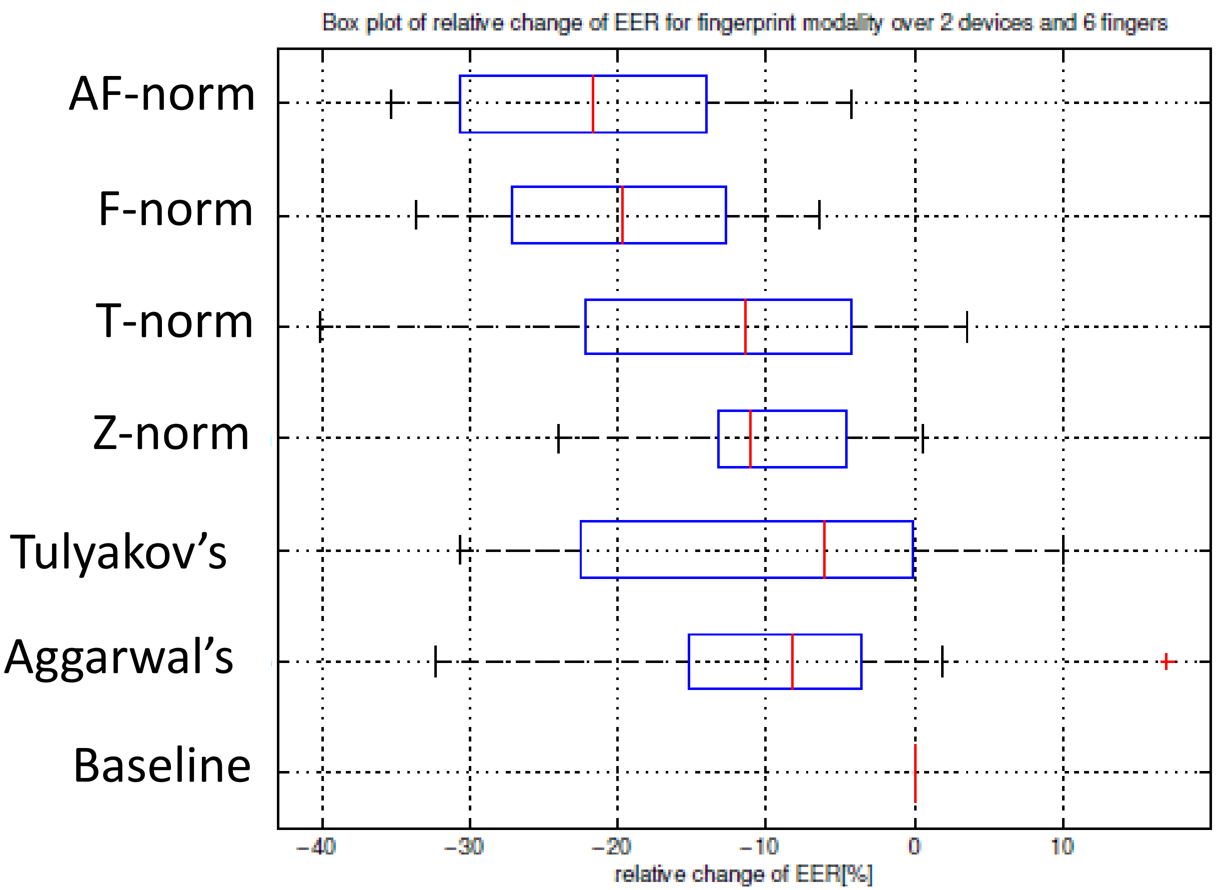
where $\tilde{\mu}(\gamma, j) = \gamma \mu^{\mathtt{d}}_{\mathtt{G},j} + (1-\gamma)\mu^{\mathtt{d}}_{\mathtt{G}}$ and $\gamma \in [0, 1]$

Client-specific mean (offline)

Global client mean:

$$\mu^{\mathtt{d}}_{\mathtt{G}} = \mathbb{E}_{j \in [1,\ldots,J]}[\mu^{\mathtt{d}}_{\mathtt{G},j}]$$

Fingerprint experiments

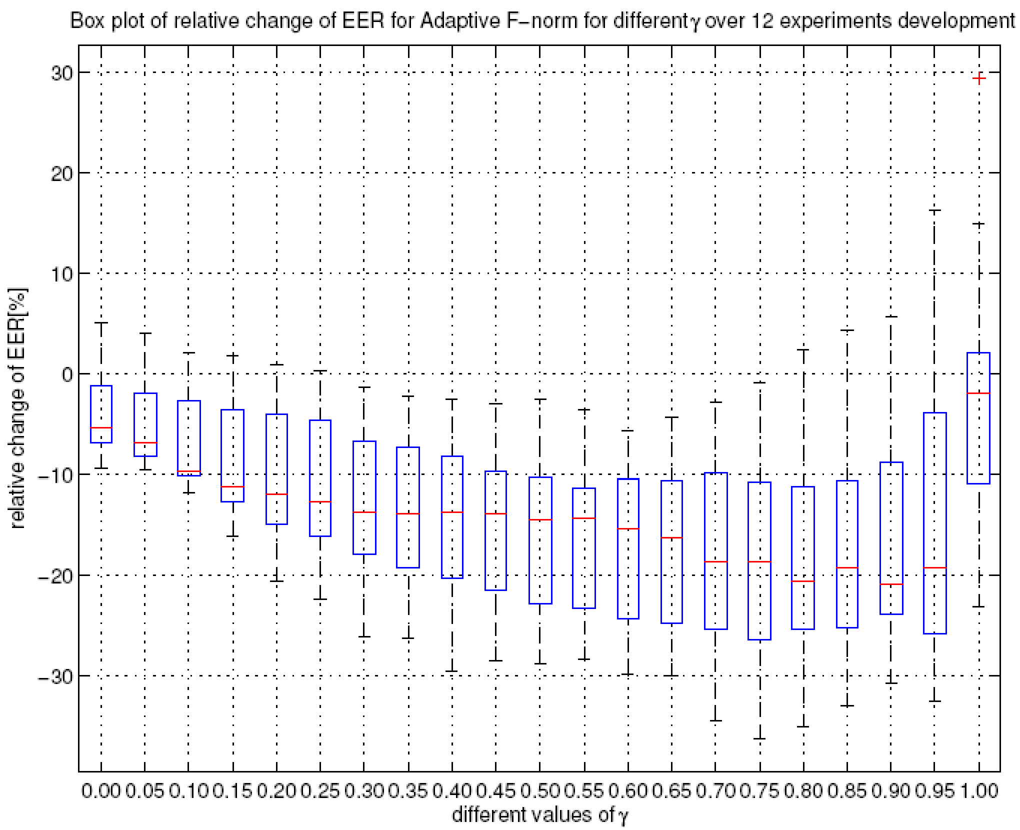


$$\frac{EER_{new} - EER_{bline}}{EER_{bline}}$$

Biosecure DS2
6 fingers x 2 devices

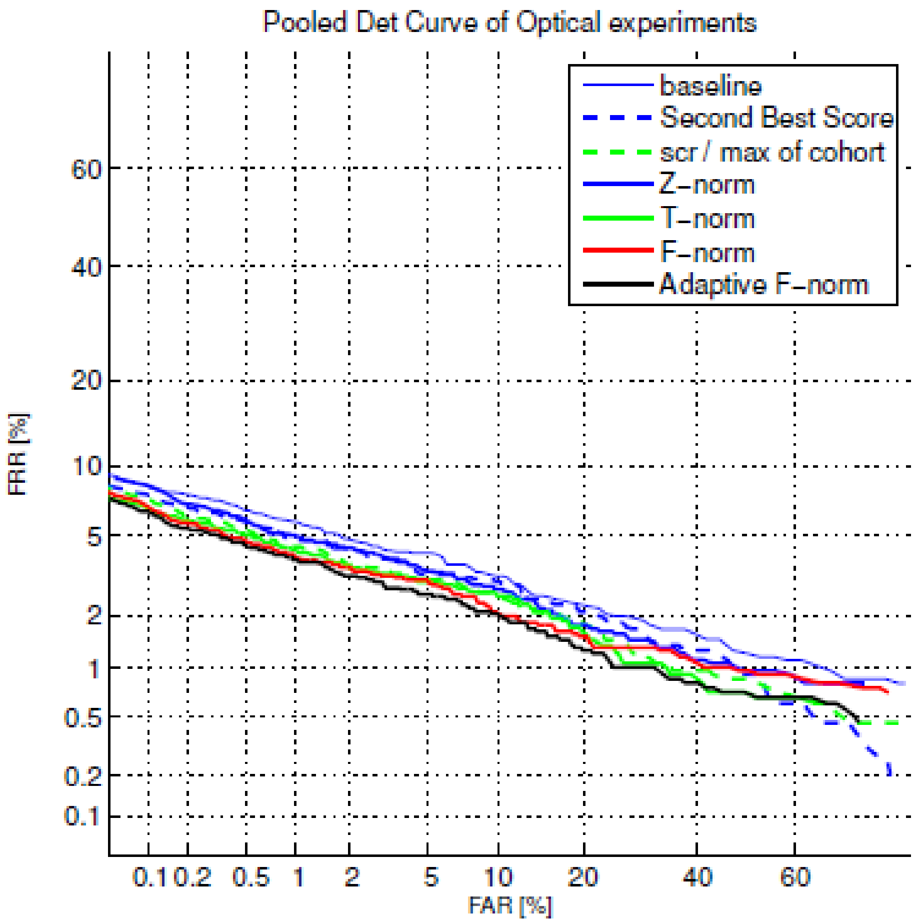
[BTAS'09]

Effect of the gamma parameter



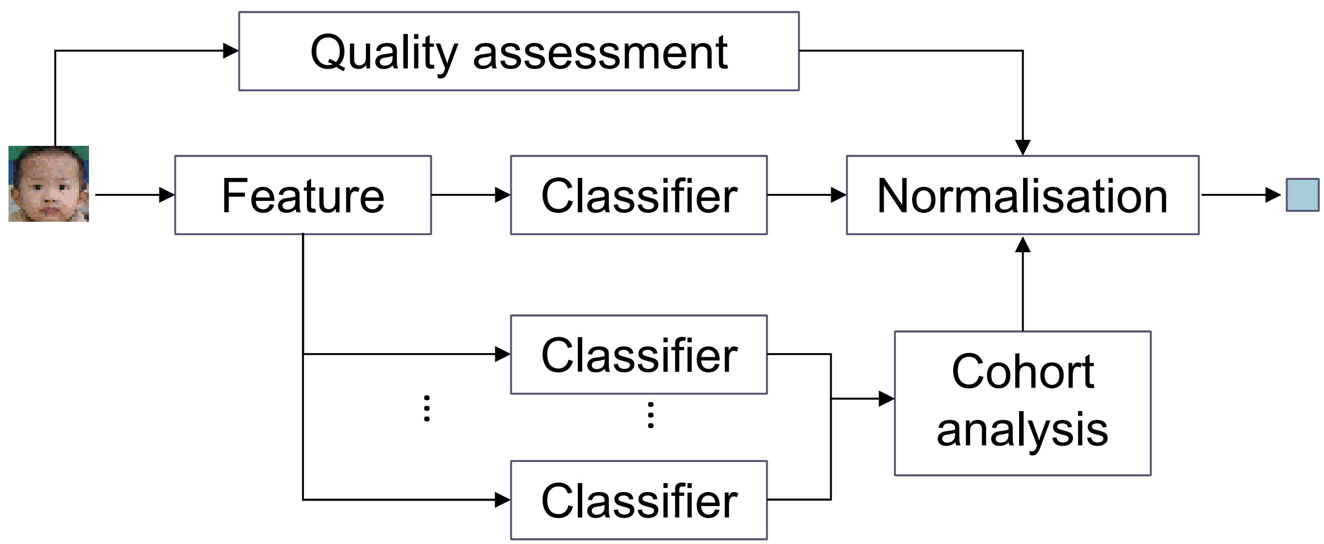
Box plot of relative change of EER (%) versus γ, assessed on the development set (and not the evaluation set).

Recommendation:Set gamma=0.5 when there is only one genuine score to adapt; and higher if there are more training samples

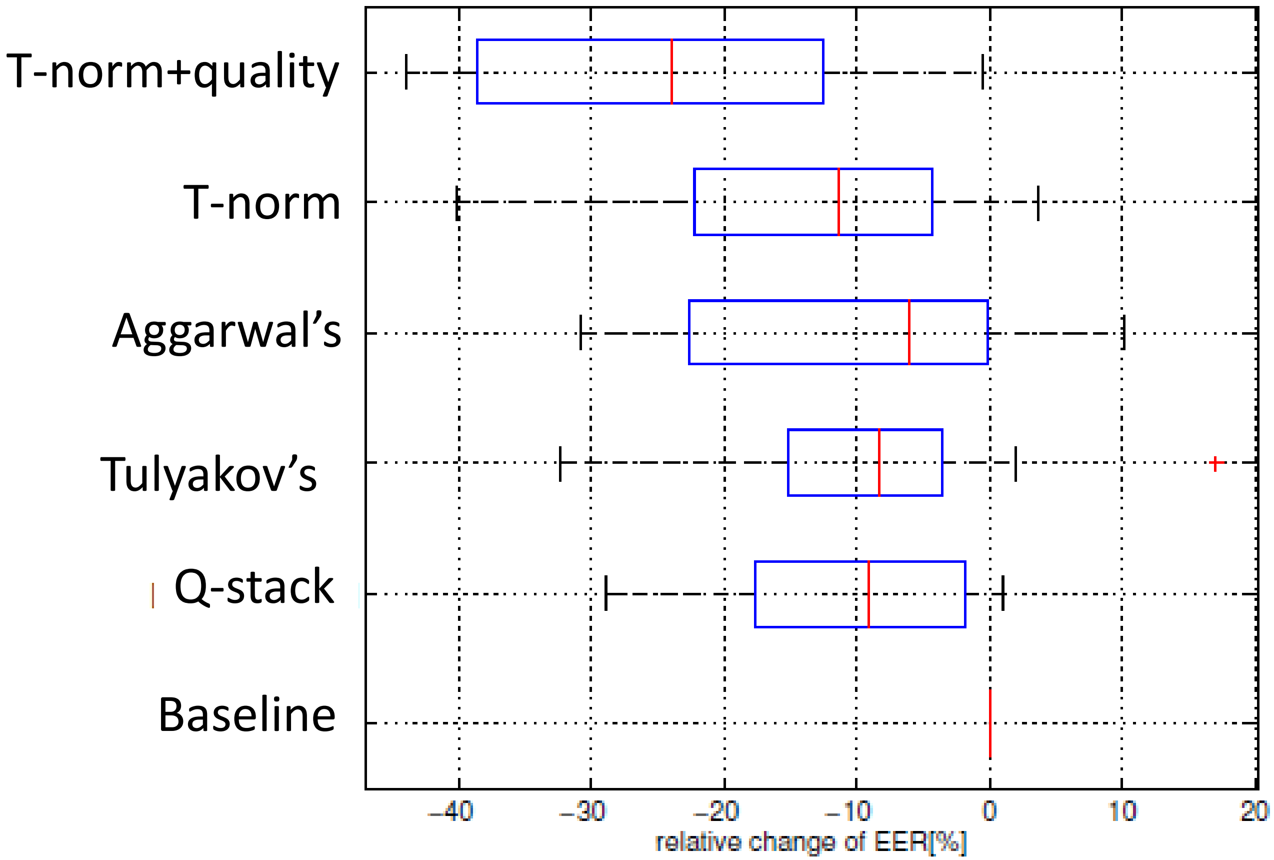


Pooled Det Curve of Optical experiments
baseline
Second Best Score
scr / max of cohort
Z−norm
T−norm
F−norm
Adaptive F−norm
FRR [%]
FAR [%]
60
40
20
10
5
2
1
0.5
0.2
0.1
0.1
0.2
0.5
1
2
5
10
20
40
60

Case II: Cohort + quality information



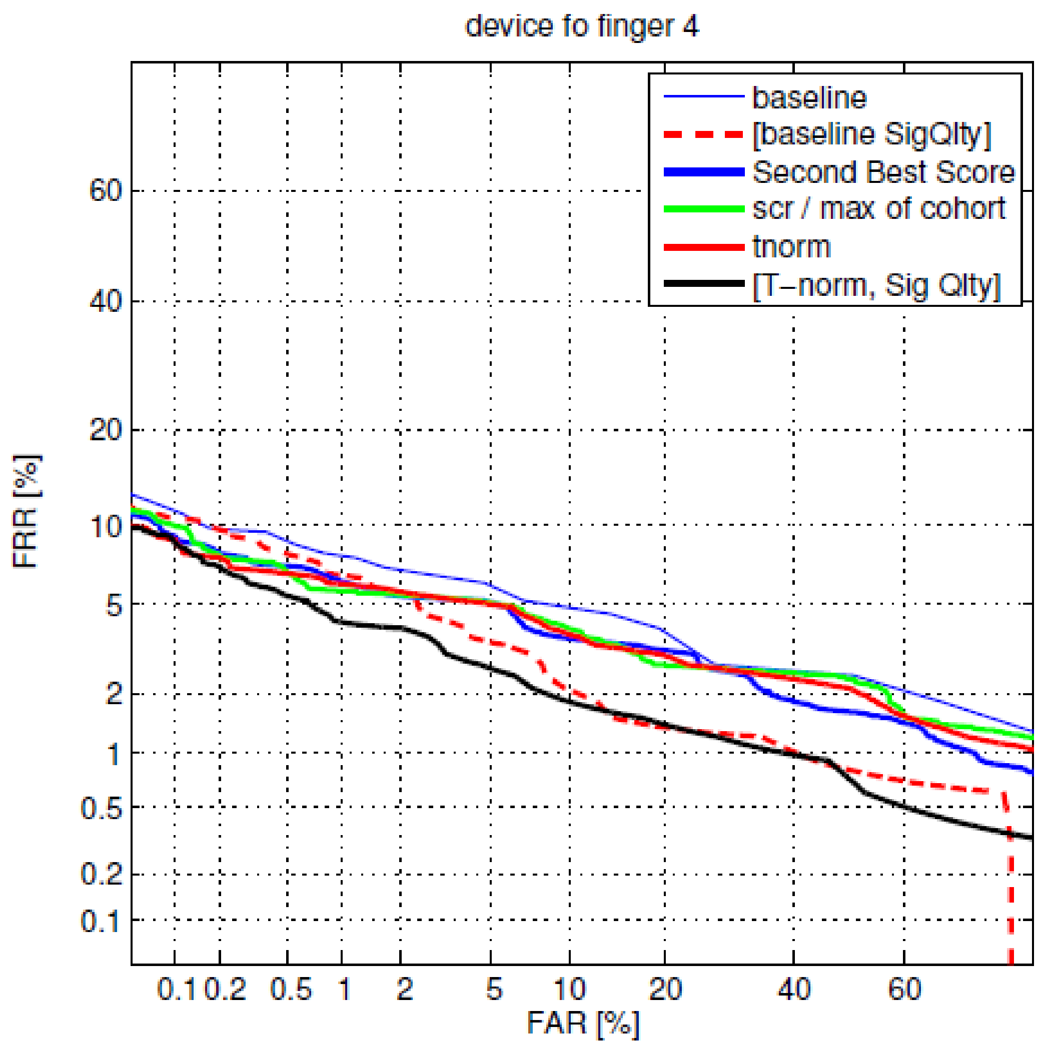
Fingerprint experiments

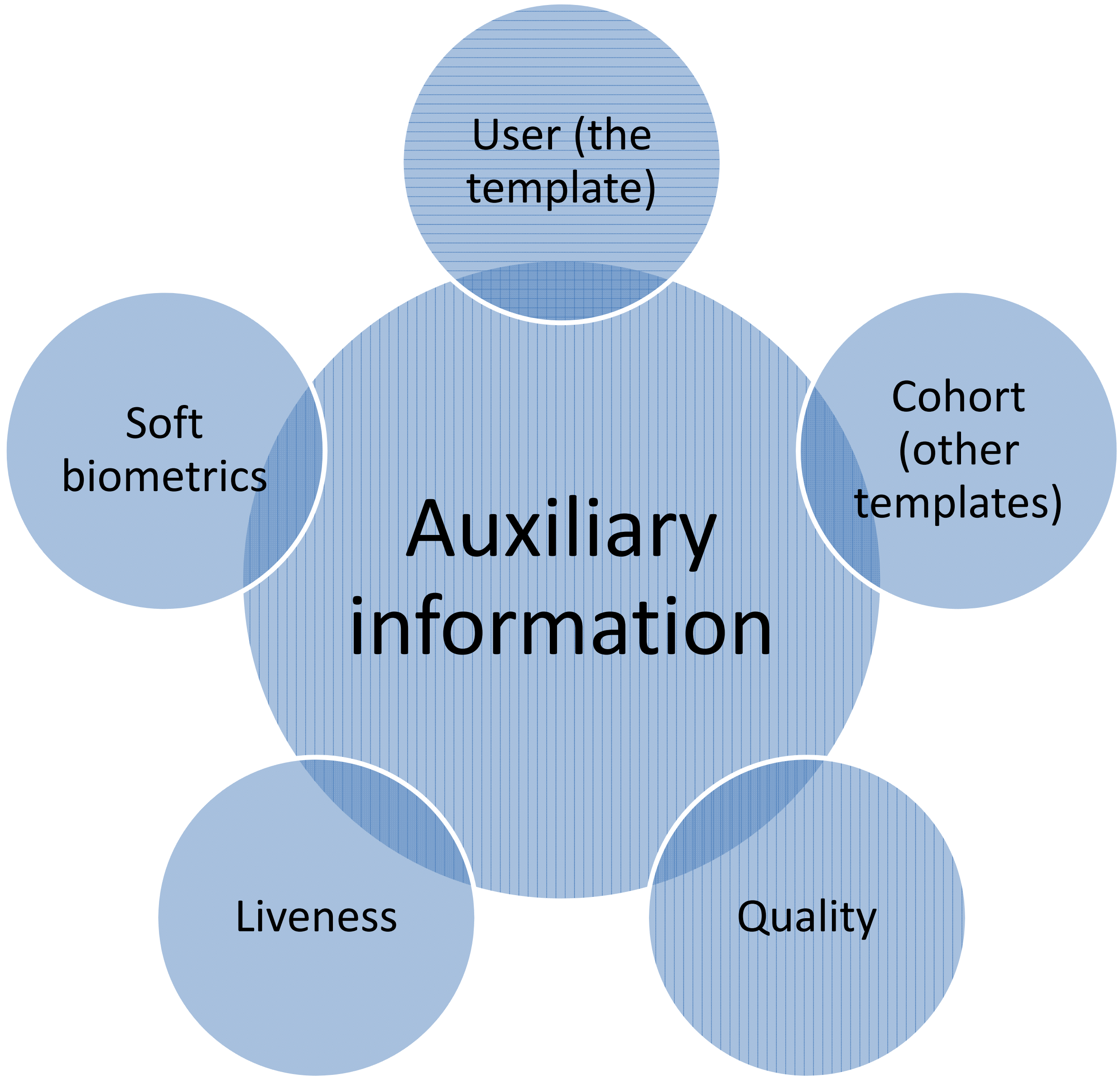


"Confidence Interval" derived from 12 experiments

[EUSIPCO'09]

$$\frac{EER_{new} - EER_{bline}}{EER_{bline}}$$


device fo finger 4
baseline
[baseline SigQlty]
Second Best Score
scr / max of cohort
tnorm
[T−norm, Sig Qlty]
FRR [%]
FAR [%]
60
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2
0.5
1
2
5
10
20
40
60


User (the template)
Soft biometrics
Cohort (other templates)
Auxiliary information
Liveness
Quality

References

- http://info.ee.surrey.ac.uk/Personal/Norman.Poh/publications.php?submenu=2