Performance Evaluation of Biometric Template Update

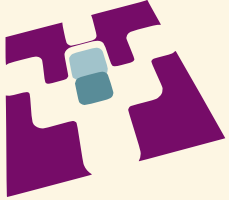
IBPC 2012

R. Giot[1], C. Rosenberger[1], B. Dorizzi[2]

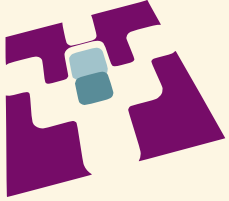
[1] Université de Caen Basse-Normandie, UMR 6072 GREYC
ENSICAEN, UMR 6072 GREYC
CNRS, UMR 6072 GREYC

[2] Institut Télécom; Télécom SudParis
UMR 5157 SAMOVAR

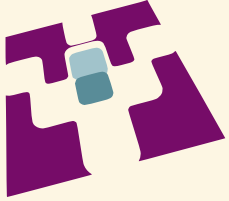
March 8, 2012



Outline



Outline



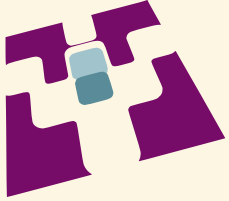
Introduction

Template update

- Allows to take into account intraclass variability through time
- Active field of research
- Experimented on various modalities

Template update evaluation

- Lacks of homogeinity
- Does not allow study comparison



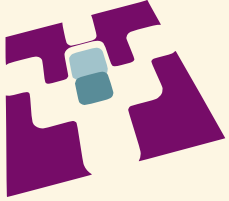
Outline

Available Public Databases

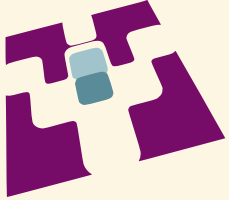
- Several databases are used in the literature
- They concern different modalities
  - Morphological modalities
    - 2D face (*EQUINOX, MORPH, UMIST, AR, VADANA*)
    - 3D face (*FRGC-EXP3*)
    - Fingerprint (*FVC2002*)
  - Behavioral modalities
    - Keystroke dynamics (*GREYC2009, DSL2009*)
    - Handed signature (*MCYT-100*)
- Few of them are specifically designed for template update (cf. next slide)

Differences Among Datasets

Database	# users	# samples	# sessions
2D face			
EQUINOX	40-50	20-100	-
MORPH	14	> 20	-
UMIST	20	25-55	-
AR	120	26	2
YOUTUBE videos	4	1200	1200
VADANA	43	≈53	-
3D face			
FRGC-EXP3	410+270	1-22	-
Fingerprint			
FVC2002	110	8	1
Keystroke dynamics			
GREYC2009	100	60	5
DSN2009	51	400	8
Handwritten signature			
MCYT-100	100	25	5



Outline



Differences Among Studies

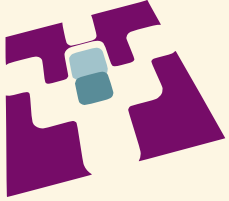
We can find several differences in the:

1. Template update system
   - ⇒ mechanism used to update the biometric reference

   This is not our subject in this presentation
2. Template update scenario
   - ⇒ configuration parameters of the study

   evaluation in a specific context
3. Template update evaluation
   - ⇒ Analysis of the performance of the system

   We will illustrate this point in this presentation



Scenario differences

Sessions Awareness

- Several sessions
- No session separation

Query Presentation Order

- Global
  - Genuine first
  - Impostor first
  - Random presentation
  - Rule (*Seeger et al. 2011*)
- Local (*Seeger et al. 2011*)
  - All random
  - Closest person
  - Closet sample

Query Chronology

- No respect to chronology
- Respect to chronology

Input Size

- More impostors
- More genuine
- Equal size

Illustration Of The Complexity

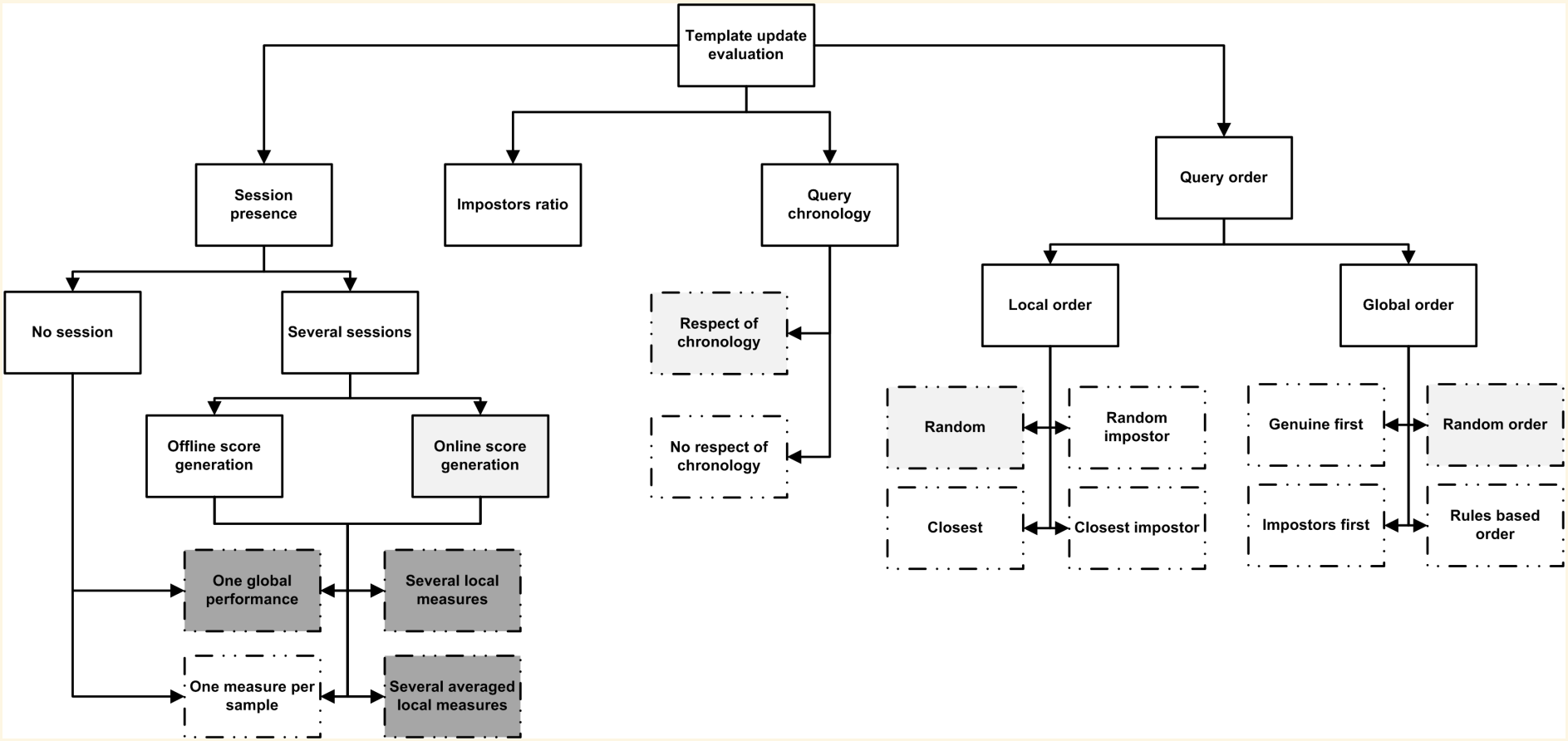
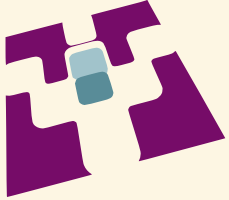


Figure: Summary of all the possible variabilities in a template update evaluation. Dotted nodes represent the possible configuration values, while nodes with a straight line represent the configuration types.



Outline

Experimental Protocol - Presentation

- We compute the scores of a biometric template update system
  - One set of score per session
  - Online evaluation
- We evaluate its performance in three different ways

Template Update System

(Giot et al. 2011)

Modality	Keystroke Dynamics
Authentication method	Distance computing
Update decision	Double-threshold semi-supervised online

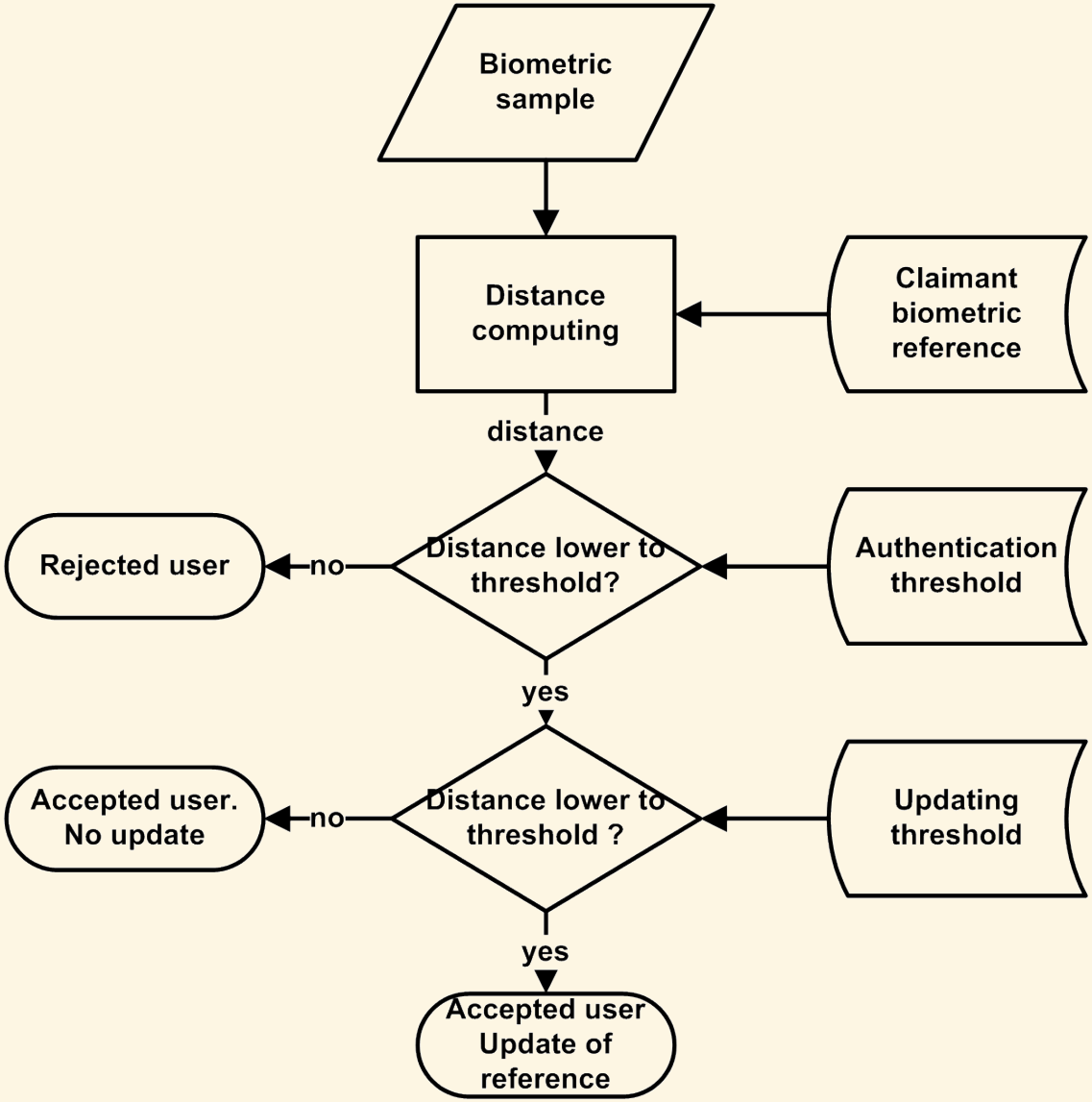


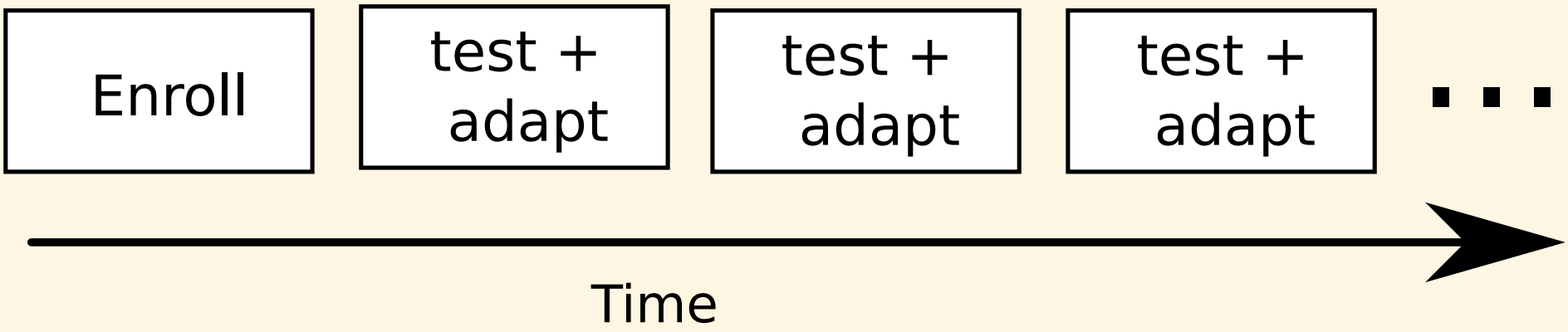
Figure: Explanation of double threshold authentication

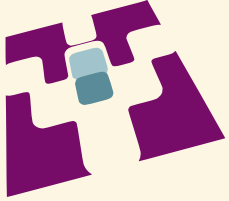
Experimental Protocol - Fixed parameters

Template Update Evaluation Parameters (*Giot et al. 2011*)

Sessions	yes (8 sessions)
Evaluation	online
Respect to chronology	yes
Impostor rate	30%
Presentation orders	random
Evaluation metric	Equal Error Rate (Variation of the acceptance threshold)

Experimental Protocol - Scores Computation





Experimental Protocol - 1st evaluation

Evaluation A

(Giot et al. 2011)

- Session performance is computed with the scores computed at this session:

$$\boldsymbol{A_i} = \mathrm{EER}\left(\boldsymbol{scores_i}\right), \quad \forall \boldsymbol{i}, 2 \leq \boldsymbol{i} \leq \#\ \text{sessions} \tag{1}$$

- We have one EER per validation session:

$$\mathbf{A} = [\boldsymbol{A}_2, \ldots, \boldsymbol{A}_{\#\boldsymbol{sessions}}] \tag{2}$$

Experimental Protocol - 2nd evaluation

Evaluation B

(Rattani et al. 2011)

- Session performance is computed by the mean of all the previous sessions' performance (including the current one).

$$\begin{aligned} B_i &= \frac{1}{i-1}\sum_{j=2}^{i} \text{EER}\left(scores_j\right), \quad \forall i, 2 \leq i \leq \#\text{ sessions} \\ &= \frac{1}{i-1}\sum_{j=2}^{i} A_i \end{aligned} \tag{3}$$

- We have one EER per validation session:

$$\mathrm{B} = [B_2, \ldots, B_{\#sessions}] \tag{4}$$

Experimental Protocol - 3rd evaluation

Evaluation C

(Seeger et al. 2011)

- One global performance measure is computed (***i.e.***, all scores of all sessions are merged):

$$C = \text{EER}\left(\bigcup_{i=2}^{\#\text{ sessions}} \textit{scores}_i\right) \tag{5}$$

- We have one EER for the whole interval:

$$\mathbf{C} = [\underbrace{C, \ldots, C}_{\#\textit{sessions}-1}] \tag{6}$$

Results - One threshold configuration

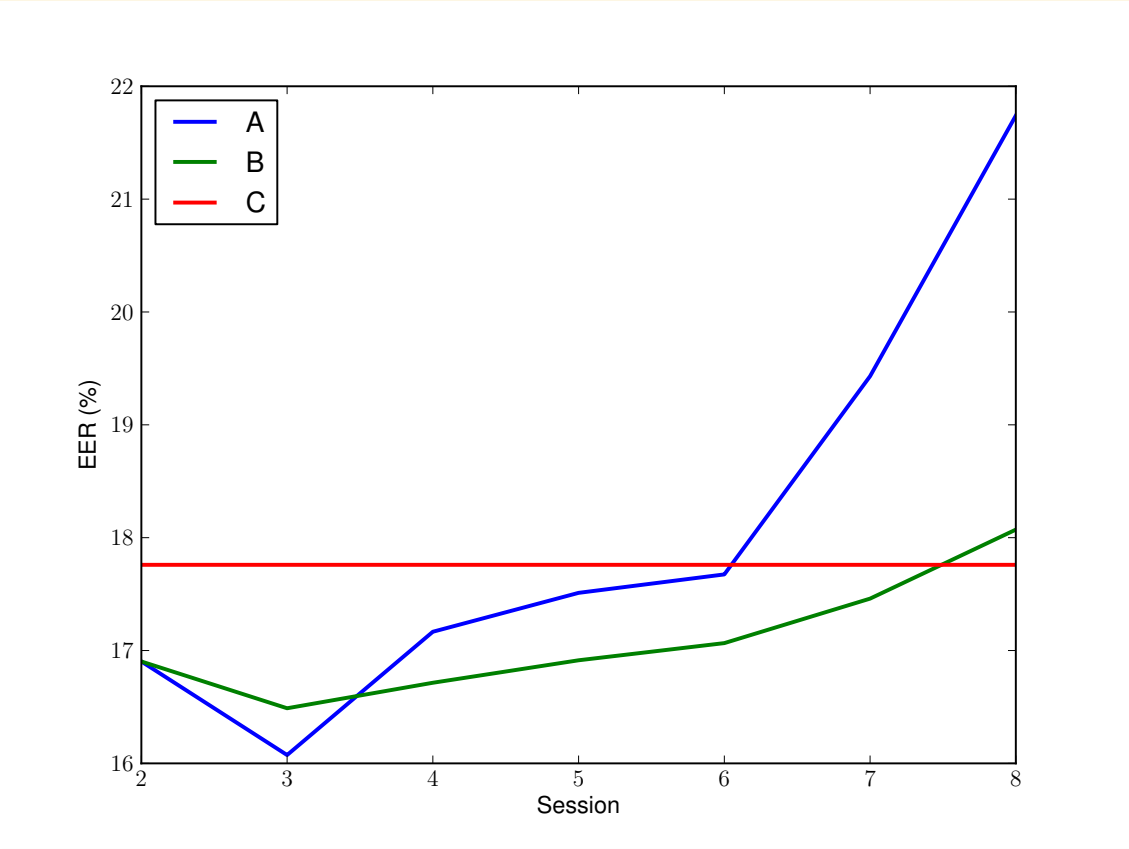


Figure: EER value per session, for one update threshold.

Results - Another threshold configuration

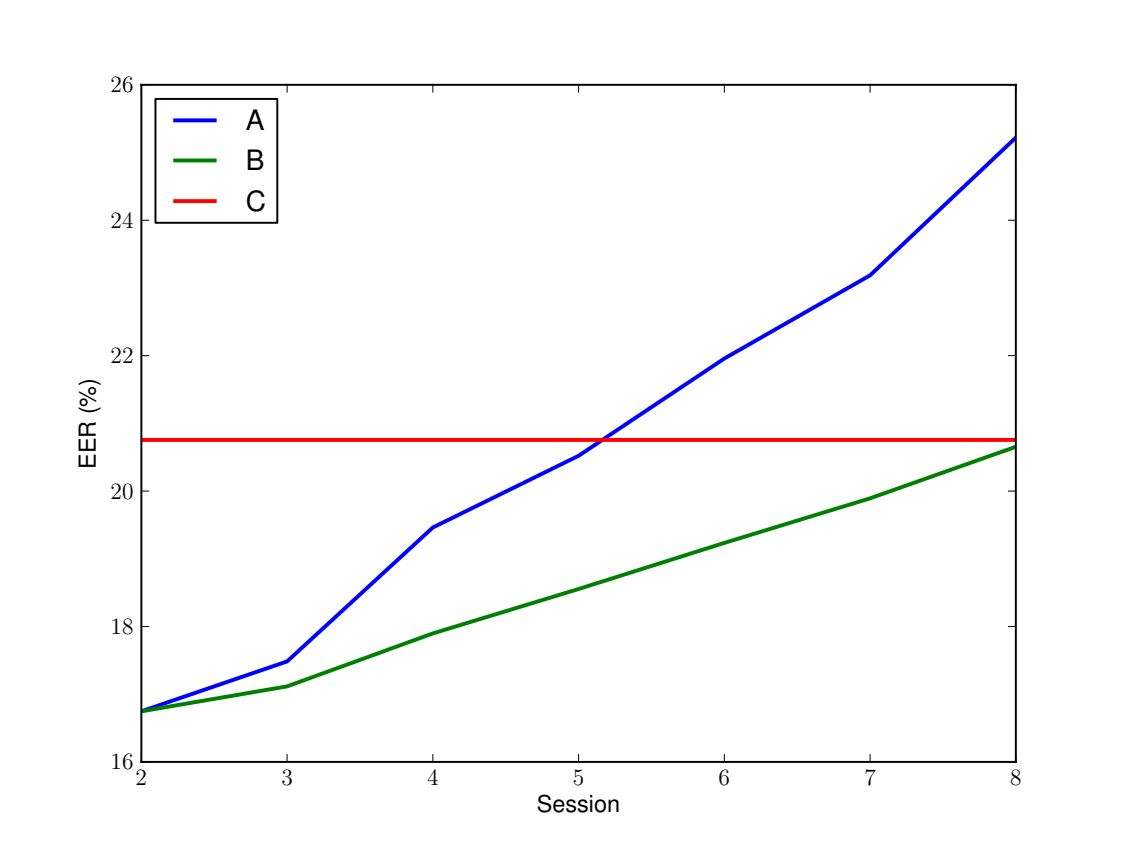
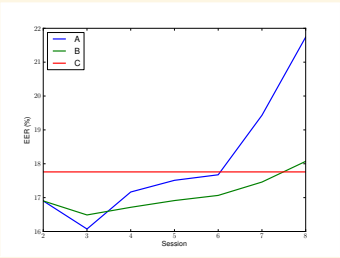
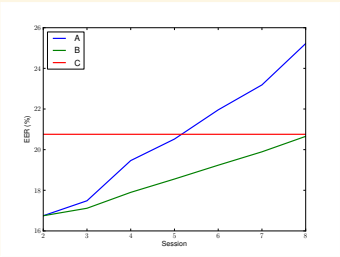


Figure: EER value per session, for one update threshold.

Interpretation

Interpretation is different depending on the evaluation scheme



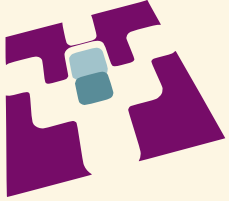


A The template update system does not perform well.

B The template update system is not too bad.

C Performance is averaged, but we cannot know if it is because of template ageing, because of a poor algorithm or because of a bad dataset.

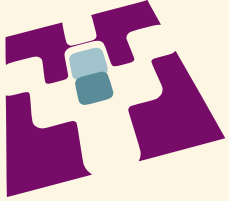
In the three schemes, the scores are strictly the sames.



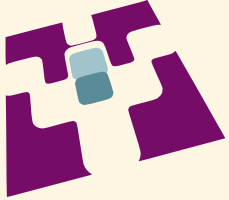
Interpretation

Discussion

- This difference of interpretation is problematic
- We cannot fairly compare the existing studies
- Which of these three methods is more appropriate ?



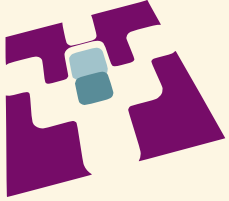
Outline



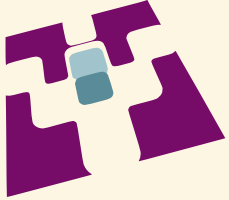
Open Questions

In addition to these problems, we can raise additional questions:

- What are the characteristics of an interesting dataset for such kind of studies?
- What is the best evaluation procedure in order to easily compare the systems without doing each time all the previous experiments from scratch ?
- Is it more informative to work with datasets separated in several sessions, or with datasets captured in a longer period without more information ?



Outline

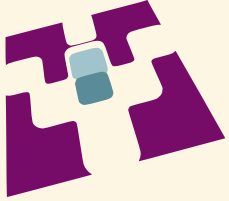


Conclusion

Template update is an active field of research.
However, there is no common:

1. Way of evaluation or template update systems
   - We have shown that the way of evaluating a system can change its perception.
2. Method to create and characterize useful datasets
   - Most datasets are not specifically designed for template update
3. Specific vocabulary
   - First try in keystroke dynamics (*Seeger et al. 2011*)

We think that these three points must be answered in the future in order to ease the work on such subject (especially the first one).



Thank you for your attention

Any questions ?