1st BioSec Workshop
Barcelona, June 28th, 2004

Independent Evaluation of Biometric System Performance

Prof. Davide Maltoni

BioLab - Biometric Systems Lab
University of Bologna - ITALY
http://bias.csr.unibo.it/research/biolab

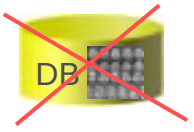
Evaluation of Biometric Systems

- Technology, Scenario and Operational evaluations

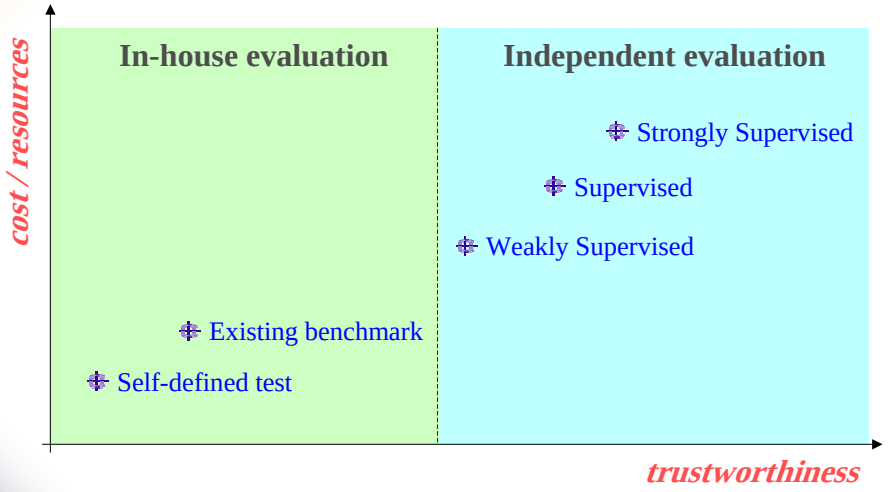
Off-line evaluation

- *comparable*
- *reproducible*

On-line evaluation



How much can we trust an evaluation?



Why independent ?

Judging one's own work is hard,
and judging dispassionately impossible...

For his mother, he is certainly a lovely puppy!

In-house: Self-defined test

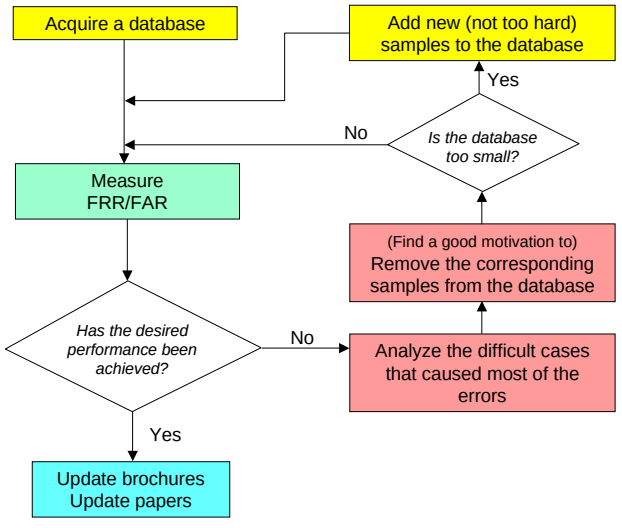
Usually internally collected

Self-defined training and test set
Often not well documented

Not comparable
Not reproducible by a third party

In-house: Self defined test (2)

An iterative approach to achieve a good performance...



In-house: Existing benchmark

Publicly available database

Existing test protocol for the database

Comparable with others obtained using the same protocol on the same database

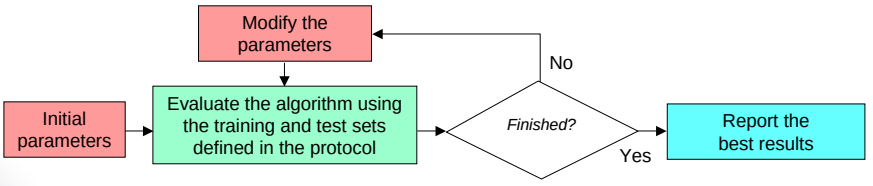
In-house: Existing benchmark (2)

Some examples

- *Most of the recent relevant scientific papers*
- *Face Authentication Contest at ICPR2000 (Part I)*
- *FAC2004: ICBA Face Verification Contest*

Main drawback

- *Overfitting (partially mitigated in case of a very large test set)*



Independent: Weakly Supervised

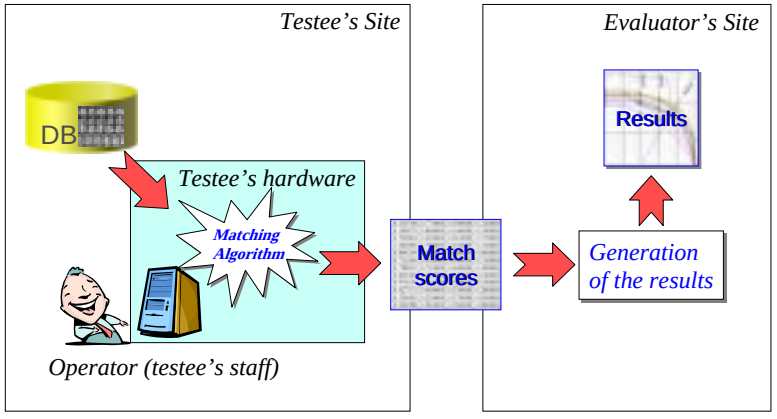
Sequestered data made available only during the test
Data is unlabeled (images with random filenames)

Test executed at testee's site
Time constraints

Results generated by the evaluator from the matching scores obtained during the test

Independent: Weakly Supervised (2)

Testing procedure



Independent: Weakly Supervised (3)

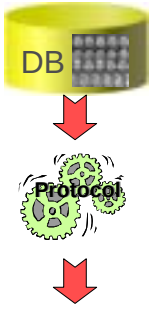
Some examples

- *FERET (1996)*
  - *3813 images cross-matched*
  - *time limit: 72 hours*
- *FRVT2000*
  - *13872 images cross-matched (>192 millions comparisons)*
  - *time limit: 72 hours (>740 matches per second)*

Main drawback

- *Cannot avoid human intervention, result editing ...*
  which could be in principle carried out with huge resources

Independent: Supervised



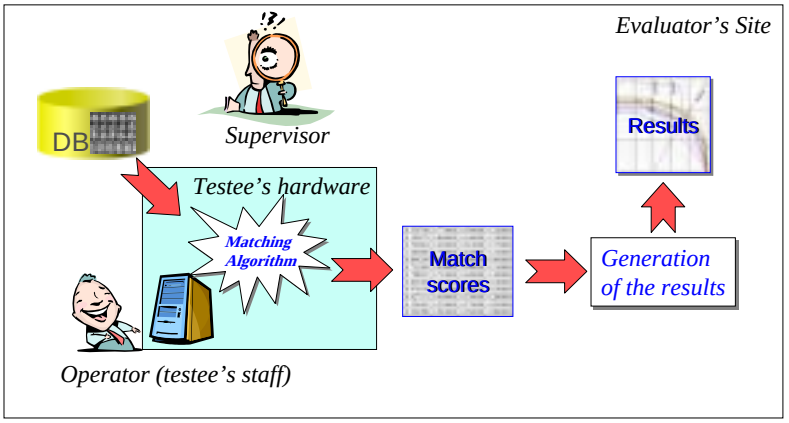
Sequestered data made available only during the test
Data is unlabeled (images with random filenames)

Test executed at evaluator's site, on testee's hardware
Time constraints

Results generated by the evaluator from the matching scores obtained during the test

Independent: Supervised (2)

Testing procedure



Independent: Supervised (3)

Some examples

- *FRVT2002*

FACE RECOGNITION VENDOR TEST

- *FpVTE2003*

FpVTE

Drawbacks

- *No way to compare efficiency (different hardware can be used)*
- *Other interesting statistics (template size, memory usage) cannot be obtained*
- *Cannot avoid "score normalization", "template consolidation"*

Independent: Strongly Supervised

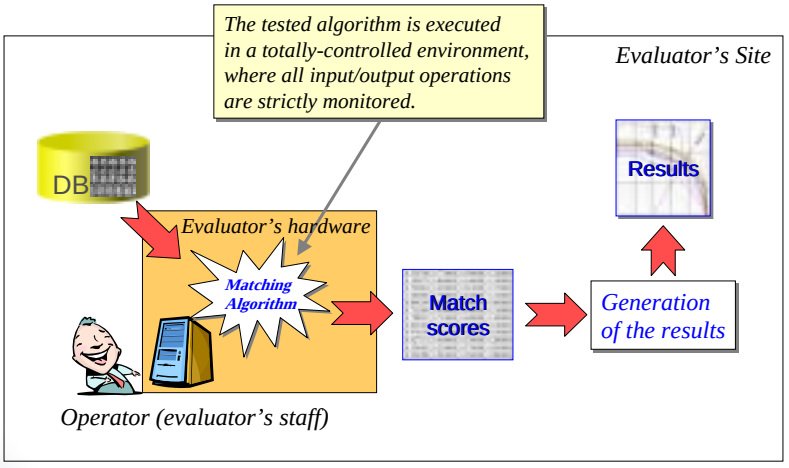
Sequestered data not made available during the test

Software components compliant to a given input/output protocol are tested on the evaluator's hardware

Results generated by the evaluator from the matching scores obtained during the test

Independent: Strongly Supervised (2)

Testing procedure



Independent: Strongly Supervised (3)

Some examples

- *FVC2000*
- *FVC2002*
- *FVC2004*

- *SVC2004*

Main drawback

- *Very time- and resource-consuming*

…but in practice

Are Independent Strongly Supervised Evaluations always practicable ?

- High cost
  - Database acquisition and tuning (appropriate size and difficulty)
  - Hardware and logistic
  - Personnel (supervising test, solving compatibility problems)
- Most of the biometric companies cannot afford such costs
- Biometric systems are characterized by frequent updates, this requires continuous performance evaluations

Performance evaluations vs Comparative evaluations

Performance evaluation

- aimed at measuring *absolute* performance of a system
- large database, representative population, …

Comparative evaluation

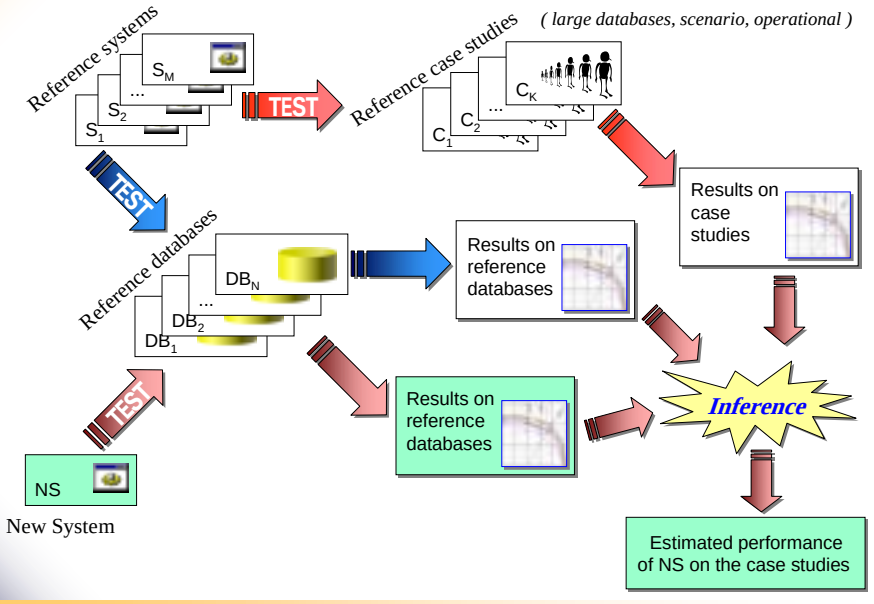
- aimed at measuring *relative* performance of a system
- less expensive, easy to reproduce for a given DB

Comparative evaluation + Inference

?

Estimation of real performance

Inferring performance



Conclusions

- Avoid Self-defined In-house evaluation
  - when standard databases and protocols are available
  - otherwise, provide your database to the community!
- Strongly Supervised Independent evaluations are the most trustable but also the most expensive.
- To what extent can we use comparative evaluation + inference to estimate real performances?

Ranking difference in FVC2002

Database 1	Database 2	Database 3	Database 4
1st - *PA15*	1st - *PA27*	1st - *PA15*	1st - *PA15*
2nd - *PB27*	2nd - *PA15*	2nd - *PA27*	2nd - *PB27*
3rd - *PA27*	3rd - *PB27*	3rd - *PB15*	3rd - *PA27*
4th - *PB05*	4th - *PA08*	4th - *PB27*	4th - *PA02*
...	...	...	...
31st - *PA03*	31st - *PA03*	31st - *PA03*	31st - *PA03*

FVC2002
Average Ranking Difference = 2.84
(St.Dev = 2.51)

Ranking Difference for PB27:
(|2-3|+|2-4|+|2-2|+|3-4|+|3-2|+|4-2|)/6=1.17