Assessing Methodologies for Operational Testing and Evaluation on Biometric Black Boxes

Antoine CABANA
ELITT & GREYC

Christophe CHARRIER
GREYC

Alain LOUIS
ELITT
Spread of biometric authentication solution

- Grant access to various applications/data
  - Critical data
    - Banking information
    - E-mails
    - Privacy

- Necessity of an evaluation process
  - Performed by third parties
    - Manufacturer’s cooperation?
    - On operational devices?

- Black boxes evaluation
Presentation of the problematic

Presentation of the two experiments
  - Common protocol
  - Nomad evaluation
  - Static evaluation

Outcomes confrontation

Conclusion
Biometric black boxes
- e.g. smartphones, tablets…
- No access to intermediate data
  - Biometric samples & templates
  - Comparison scores
- Access to final decision only

Constraints on the evaluation
- Availability for test of the users along the whole evaluation
- Manually performed

Objectives?
- Estimate error rate
- Ensure an upper bound
How to perform a biometric evaluation considering these constraints?

- Optimized way?
  - Industrial perspectives: time gain, reasonable costs, relevant results
  - Methodology conform to the state of the art, and standards (ISO 19795)

- Which information could be collected during the evaluation?
  - e.g. To reproduce some observed error case
  - to exploit a possible vulnerability, in security test

- How to achieve the expected upper bound?
Objectives

- e.g. $10^{-4}$ ~ PIN entropy

Time estimation

- 60h
- 10 working days

Population size

- Rule of 3 (cf. ISO 19795)
- Settings :
  - 30 users
  - 8 fingers
  - 5 presentations per finger
Standards

- ISO 19795: biometric performance testing and reporting
- ISO 29197: Evaluation methodology for environmental influence in biometric system performance
Two experiments:
- Nomad experiment
- Static experiment
**Round 1:**

Reference of the current round

**Round i:**

Users involved in impostor attempts (FAR)

**Round n:**

Genuine attempts (FRR)

Test 1: \( u_1 \) \( \vdots \) \( u_n \)

Genuine attempts, in order to allow the user to familiarize with the system

Test n: \( u_1 \) \( \vdots \) \( u_n \)

n: population size

\( u_i \): \( i \)th user
Unit under test
- smartphone

Presentation setting
- 4 fingers per hand
- Both hands
- Little finger discarded
  - Capture issues

Reference setting
- Thumb and index of preferred hand
Test populations

- Similar size: 12 and 13 people
- Similar gender representation
- Similar age representation
Base idea: trade-off between
- User’s scarce availability for testing (not a dedicated population)
- Time consumed
- Execution of the evaluation

Specific protocol
- Supervisor records results (fillable forms)
- Various offices/rooms (same building)

Limitations
- No environmental recording (conformity to ISO 19795)
- Manual processing of the results (error rate computation)
- Users’ interaction hardly observable
Results

Average time per session
- First round: 6 min 30 s
- Last rounds: ~ 2 min 30 s

Accommodation effect along the evaluation
- Familiarization with the system
- Habituation to the evaluation process

Observations
- Few information collected
- No FAR error case

Time consumption
- 3 days
- 5000 comparisons results thus FAR < 6.10^-4
**Base idea**

- Assisted evaluation
  - Test tool: records results in a dedicated database

**Specific protocol**

- Users proceed to record results
- Same location/office
- Supervisor has time to observe interactions
  - Collect system’s feedback messages
  - Determine failures’ conditions

**Limitations**

- Less convenient for users
TEST TOOL
USER INTERFACE

Explanations on Error classification

Strict rejection

FTA-type error

Uncollected error-type

TEST INSTRUCTIONS

ACCEPTED

REJECTED
TEST TOOL
USER INTERFACE

TEST INSTRUCTIONS

ACCEPTED

REJECTED
Explanations on Error classification

FTA-type error

Strict rejection

Uncollected error-type
 Results

- Time consumption
  - Average time: 6 min 30 s
  - Min: 3 min 30 s
  - Max: 11 min

- Observations
  - Information collected
  - No FAR error case

- Time consumption
  - 3 days (tests stopped)
  - 2700 comparisons results FAR < 1,1 \times 10^{-3}
<table>
<thead>
<tr>
<th>Methodology</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nomad evaluation</td>
<td>Low time consumption</td>
<td>Lack of observations</td>
</tr>
<tr>
<td></td>
<td>Little constraining</td>
<td>No environmental conditions recording</td>
</tr>
<tr>
<td>Static evaluation</td>
<td>Possible observations:</td>
<td>Slower</td>
</tr>
<tr>
<td></td>
<td>Interactions</td>
<td>Constraining for users</td>
</tr>
<tr>
<td></td>
<td>Modality condition</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Efficient system’s feedback</td>
<td></td>
</tr>
<tr>
<td></td>
<td>message recording</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

- Estimation of the required time
- Deterninate difficulties
  - Test crew presence and availability for test
  - Planning the evaluation
- Improve the test tool
  - Camera recording

Next step

- “gray box”
  - Parallel analysis of the modality
  - Ground truth (similarity score on a reference system)
- Security part
  - Spoofing & black boxes