# BioSecure Signature Evaluation Campaign 2009 (BSEC'2009): Results

Nesma Houmani, Sonia Garcia-Salicetti, Aurelien Mayoue, and Bernadette Dorizzi
Institut TELECOM; TELECOM & Management SudParis;Dept EPH , Evry, France
July 2009

In this document, we present the participants and report the experimental results on the three Evaluations conducted in BSEC'2009.

We also report the results obtained with BioSecure Reference System (http://share.int-evry.fr/svnview-eph/ref_syst/Signature_GET-INT/).

We only give a limited analysis of results, and in particular we do not compare systems between them. Indeed, we aim at publishing results in a Journal including all participants as authors (no participant was anonymous).

**Participants:** Nine universities have registered to this Evaluation Campaign, and 14 systems have been submitted. We indicate in the third column below to which of the 3 evaluations (numbered as 1, 2, 3) each system has participated.

| University | System | Evals | Matching Algorithm |
|---|---|---|---|
| Escola Universitaria Politecnica de Mataro (Spain) | Sys1 | 1, 2, 3 | Biometric Dispersion Matcher |
| | Sys2 | 1, 2, 3 | Ratio of means and standard deviation of parameters on each piece of signature |
| U1 Research Institute (Hungary) | Sys3 | 1, 2, 3 | DTW |
| Seikei University (Japan) | Sys4 | 1, 2, 3 | DTW |
| Ain Shams University (Egypt) | Sys5 | 1, 2, 3 | DTW |
| University of Valladolid (Spain) | Sys6 | 1, 2, 3 | DTW |
| Sabanci University (Turkey) | Sys7 | 1, 2, 3 | DTW |
| Universidad Autonoma de Madrid (Spain) | Sys8 | 1, 2, 3 | DTW tuned for random forgeries |
| | Sys9 | 1, 2, 3 | DTW tuned for skilled forgeries |
| | Sys10 | 1, 2, 3 | HMM |
| | Sys11 | 1, 2, 3 | Global and Mahalanobis distance |
| | Sys12 | 1, 2, 3 | Fusion of 4 systems: Sys8, Sys9, Sys10,and Sys 11 |
| Waseda University (Japan) | Sys13 | 2 | DTW |
| University of Magdeburg (Germany) | Sys14 | 2 | Biometric Hash Algorithm, Canberra distance |
| Reference System TMSP (France) | | 1, 2, 3 | HMM |

## I. Evaluation 1: Impact of mobility acquisition conditions

In this section, 12 systems were submitted. The systems were tuned on the development dataset DS2-50, and then tested only on the first session of DS2-382 and DS3-382 datasets (DS2-382 and DS3-382 contain the same persons).

The experimental results obtained on DS2-382 and DS3-382, with skilled and random forgeries are displayed by DET curves in Figures 1 and 2, and the Equal Error Rates are reported in Table1.

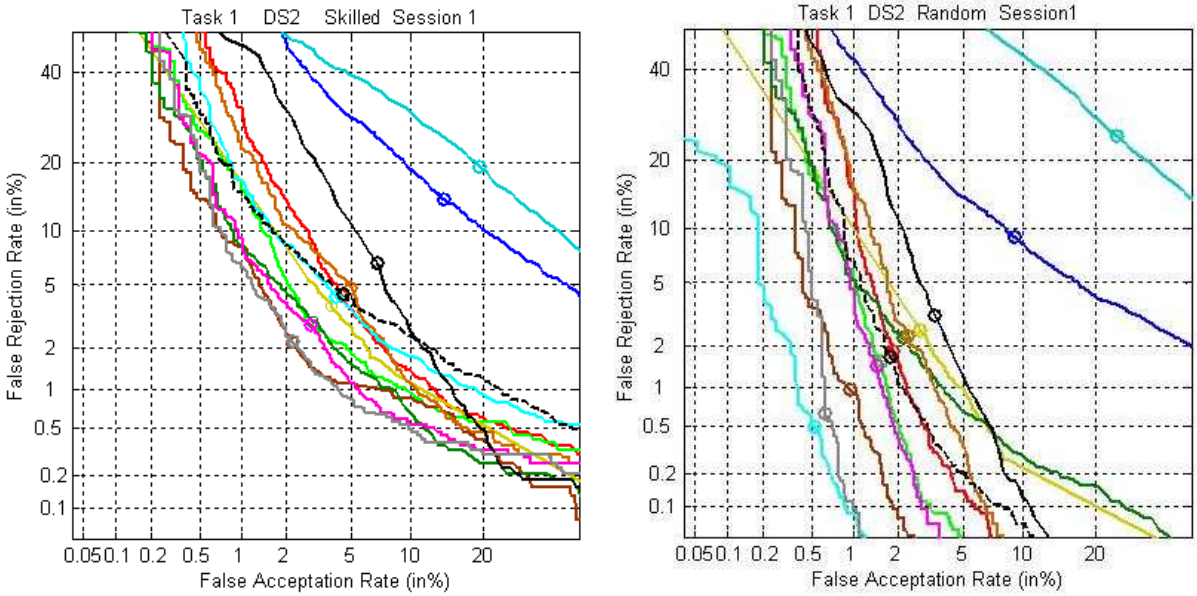


**Figure 1: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 1 of DS2-382 dataset.**

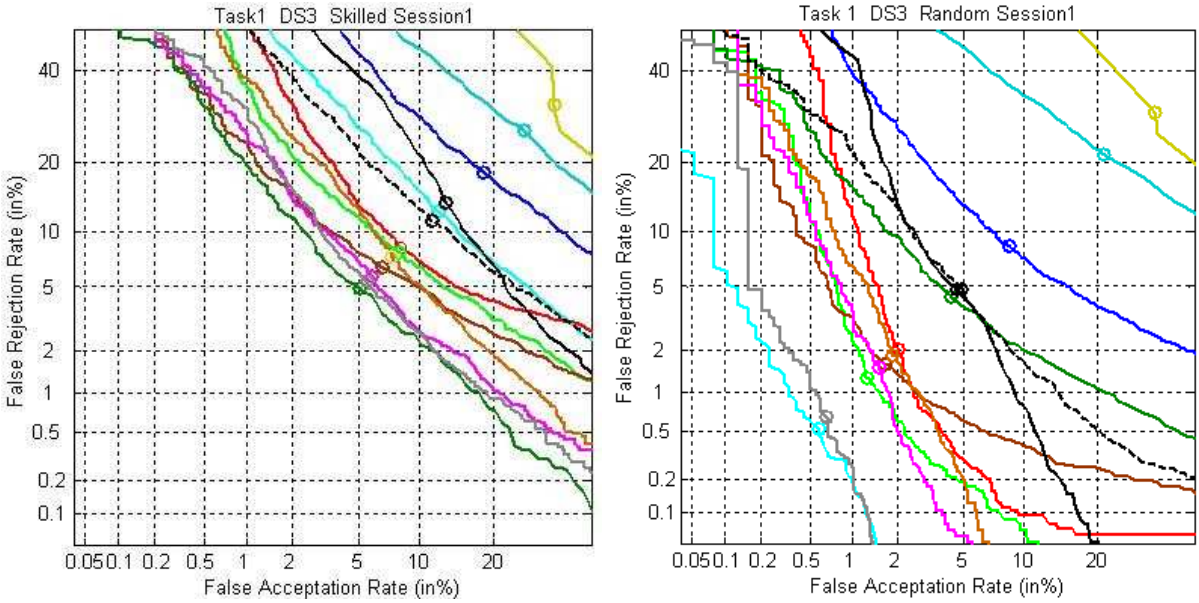


**Figure 2: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 1 of DS3-382 dataset.**

**Table 1: Equal Error Rate on Session 1 of DS2-382 and DS3-382 datasets, with skilled and random forgeries.**

| 13 Systems | DS2-382 | | DS3-382 | |
|---|---|---|---|---|
| | Skilled | Random | Skilled | Random |
| **Sys 1** | 4.4 | 1.85 | 8.18 | 2.05 |
| **Sys 2** | 4.91 | 2.33 | 7.38 | 1.86 |
| **Sys 3** | 13.99 | 8.98 | 18.32 | 8.36 |
| **Sys 4** | 2.88 | 1.58 | 7.87 | 1.29 |
| **Sys 5** | 3.82 | 2.67 | 31.57 | 30.64 |
| **Sys 6** | 2.20 | 0.97 | 6.58 | 1.65 |
| **Sys 7** | 2.98 | 2.23 | 4.99 | 4.32 |
| **Sys 8** | 4.18 | 0.51 | 12.20 | 0.55 |
| **Sys 9** | 2.88 | 1.47 | 5.77 | 1.54 |
| **Sys 10** | 19.23 | 24.14 | 25.85 | 21.34 |
| **Sys 11** | 6.71 | 3.31 | 13.26 | 4.7 |
| **Sys 12** | 2.23 | 0.63 | 5.47 | 0.66 |
| **Ref_Sys -----** | 4.47 | 1.74 | 11.27 | 4.8 |

## Analysis of results

- Globally, as expected, results on DS2-382 are better than on DS3-382. With skilled forgeries, the performance on DS3-382 is degraded roughly by a factor 2 at the EER. With random forgeries, the degradation is less important.

# II. Evaluation 2: Impact of time variability

We evaluate the submitted systems on both sessions (spaced of 4 weeks) separately, according to 3 configurations of input data:

1. only coordinates
2. coordinates and pressure
3. coordinates, pressure, and pen inclination

Note that all the systems were tuned on Session 1.

## 1. Input data: coordinates

### 1.1. On DS2-382 dataset

The 12 submitted systems were tuned on the development dataset DS2-50, then tested on DS2-382 considering both sessions separately.

Note that these systems are those tested in Evaluation 1 on Session1. In this task, the systems are tested on Session 2 of DS2-382. The experimental results obtained on DS2-382, with skilled and random forgeries, are shown by DET curves (Figures 3), and are reported in Table 2.

Figure 3 and Table 2 must be compared to Figure 1 and Table 1 in Evaluation 1.

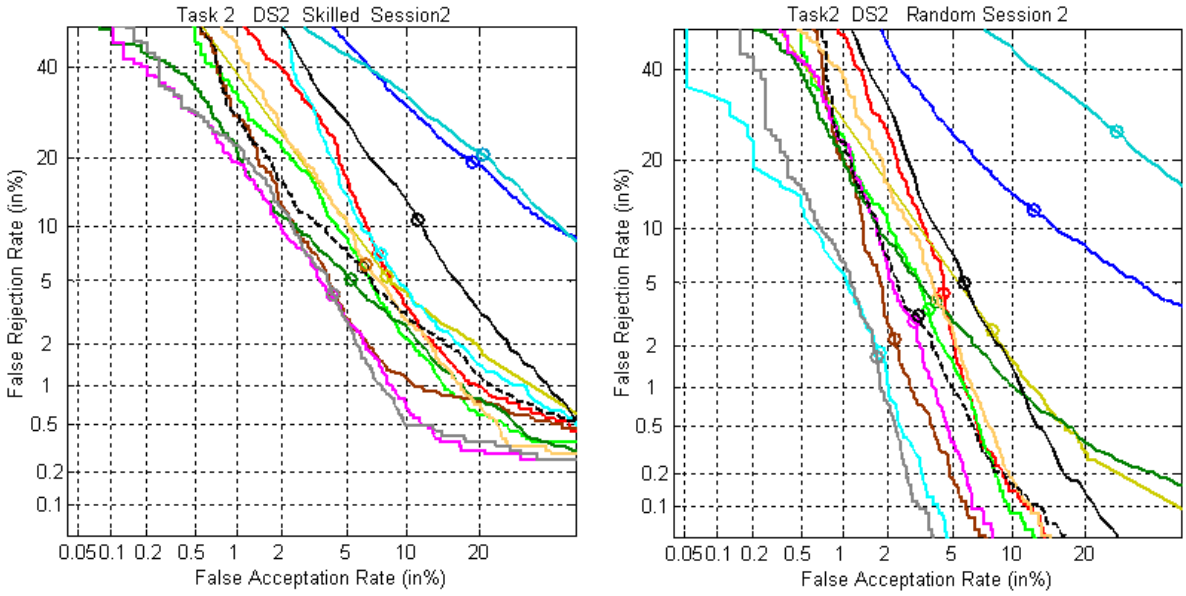


**Figure 3: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 2 of DS2-382.**

**Table 2: Equal Error Rate on Session 2 of DS2-382 dataset, with skilled and random forgeries.**

| 13 Systems | DS2-382 on Session2 | |
|---|---|---|
| | Skilled | Random |
| **Sys 1** | 7.15 | 4.40 |
| **Sys 2** | 6.20 | 4.02 |
| **Sys 3** | 19.03 | 12.29 |
| **Sys 4** | 5.99 | 3.55 |
| **Sys 5** | 6.61 | 5.23 |
| **Sys 6** | 4.21 | 2.24 |
| **Sys 7** | 5.13 | 3.96 |
| **Sys 8** | 7.26 | 1.80 |
| **Sys 9** | 4.08 | 2.93 |
| **Sys 10** | 20.47 | 25.62 |
| **Sys 11** | 10.92 | 5.37 |
| **Sys 12** | 4.18 | 1.70 |
| **Ref_Sys -----** | 5.99 | 3.16 |

## Analysis of results

- Given that all systems were tuned on Session 1, we notice as expected that results are degraded in presence of time variability (test on Session 2), and that independently of the type of forgeries. However, the ranking of the systems does not change in presence of time variability.

### 1.2. On DS3-382 dataset

In this task, 13 systems were submitted. The systems were tuned on the development dataset DS3-50, and tested on DS3-382 considering both sessions separately.
The experimental results obtained on DS3-382, with skilled and random forgeries are reported for both sessions. The DET curves in Figures 4 and 5 show system performance considering Session1 and Session 2, respectively. Equal error Rates are reported in Table 3.

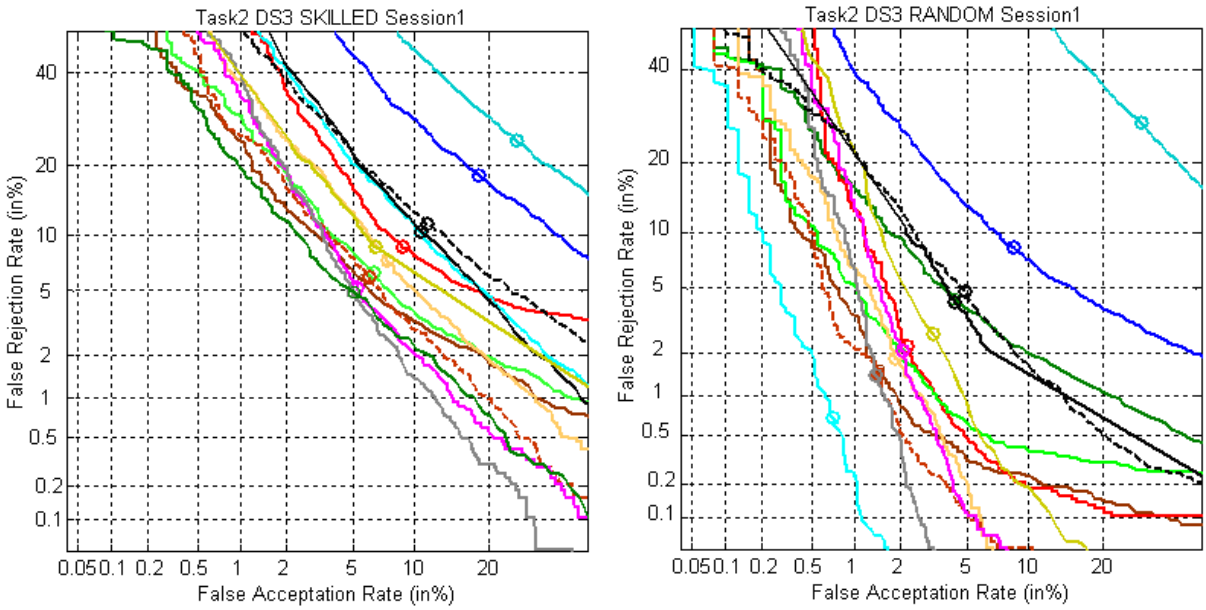


**Figure 4: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 1 of DS3-382 dataset.**

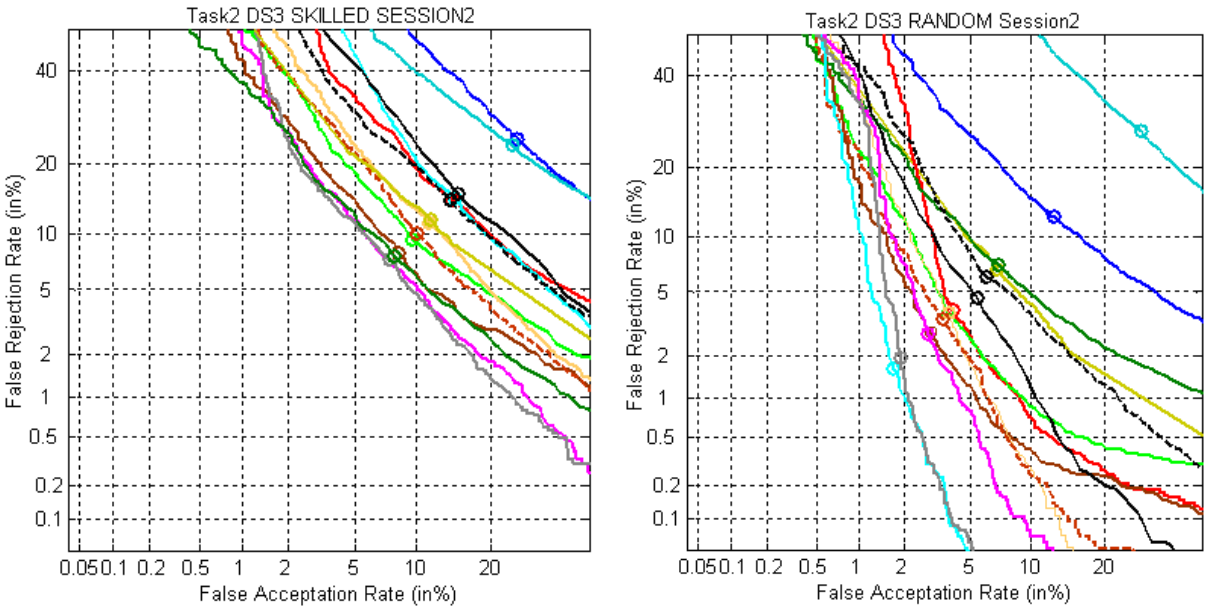


**Figure 5: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 2 of DS3-382 dataset.**

**Table 3: Equal Error Rate on DS3-382 dataset, with skilled and random forgeries, considering Session 1 and Session2 separately.**

| 14 Systems | DS3-382 | | | |
|---|---|---|---|---|
| | Session 1 | | Session 2 | |
| | Skilled | Random | Skilled | Random |
| Sys 1 | 8.71 | 2.22 | 14.24 | 3.94 |
| Sys 2 | 7.38 | 1.85 | 11.25 | 3.76 |
| Sys 3 | 18.32 | 8.36 | 24.68 | 12.40 |
| Sys 4 | 6.37 | 2 | 9.43 | 3.72 |
| Sys 5 | 7.55 | 4.24 | 11.51 | 6.78 |
| Sys 6 | 5.69 | 1.5 | 8.06 | 2.90 |
| Sys 7 | 4.98 | 4.31 | 7.69 | 7.02 |
| Sys 8 | 10.40 | 0.7 | 14.51 | 1.67 |
| Sys 9 | 5.24 | 2.09 | 7.42 | 2.83 |
| Sys 10 | 24.79 | 27.29 | 23.52 | 26.81 |
| Sys 11 | 10.49 | 2.93 | 15 | 5.01 |
| Sys 12 | 4.93 | 1.41 | 7.42 | 1.93 |
| Sys 13 ---- | 5.98 | 1.44 | 9.93 | 3.48 |
| Ref_Sys ----- | 11.27 | 4.8 | 14.03 | 6.06 |

## Analysis of results

- As for DS2-382, given that all systems were tuned on Session 1, we notice as expected that results are degraded in presence of time variability (test on Session 2), and that independently of the type of forgeries.

- When comparing systems developed on DS2-50 and then tested on DS3-382, with systems developed on DS3-50 and then tested on DS3-382 (Figure 4 vs. Figure 2), we notice that some systems are enhanced when the Development Set contains data of the same nature than in the Test Set.

## 2. Input data: coordinates and pressure

For this task, 11 systems were submitted. The experimental results with skilled and random forgeries are shown by DET curves (Figures 6 and 7), and Equal Error Rates are reported in Table 4.

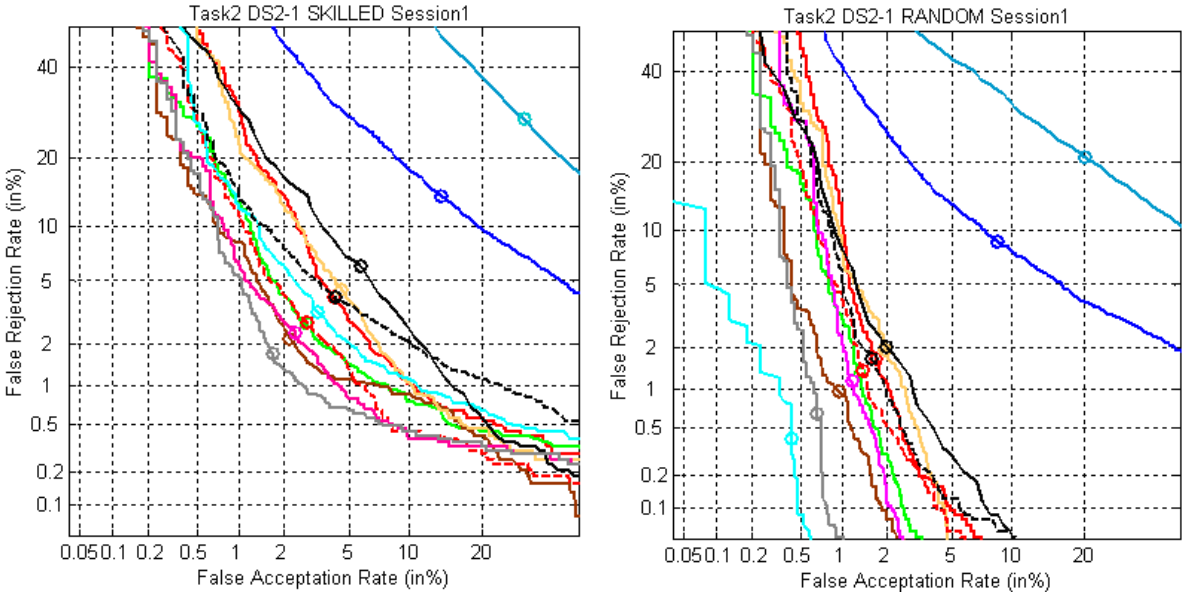


**Figure 6: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 1 of DS2-382 dataset.**

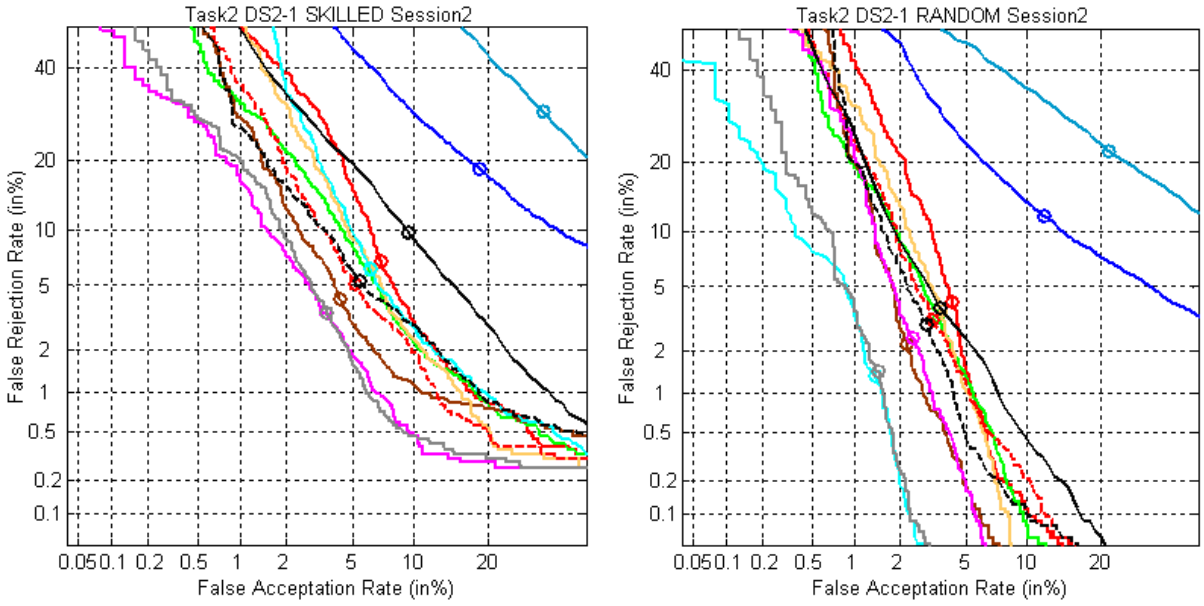


**Figure 7: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 2 of DS2-382 dataset.**

**Table 4: Equal Error Rate on DS2-382 dataset, with skilled and random forgeries, considering Session 1 and Session2 separately.**

| 12 Systems | DS2-382 | | | |
|---|---|---|---|---|
| | Session 1 | | Session 2 | |
| | Skilled | Random | Skilled | Random |
| **Sys 1** | 4.03 | 1.70 | 6.88 | 4.16 |
| **Sys 2** | 4.5 | 1.96 | 6.28 | 3.61 |
| **Sys 3** | 13.69 | 8.62 | 18.54 | 11.81 |
| **Sys 4** | 2.76 | 1.33 | 6.07 | 3.42 |
| **Sys 6** | 2.19 | 0.97 | 4.21 | 2.23 |
| **Sys 8** | 3.26 | 0.42 | 6.21 | 1.37 |
| **Sys 9** | 2.38 | 1.17 | 3.48 | 2.46 |
| **Sys 10** | 27.76 | 20.51 | 30.13 | 21.61 |
| **Sys 11** | 5.9 | 2.02 | 9.52 | 3.65 |
| **Sys 12** | 1.71 | 0.65 | 3.49 | 1.46 |
| **Sys 13 - - - -** | 2.84 | 1.38 | 5.1 | 3.19 |
| **Ref_Sys - - - - -** | 4.07 | 1.65 | 5.32 | 2.96 |

## Analysis of results

- When comparing Figure 6 to Figure 1, adding pressure to coordinates as inputs to the systems, improves performance when no time variability is present.

- Using pressure at the input of the systems in addition to coordinates, does not improve the robustness of the systems to time variability.

## 3. Input data: coordinates, pressure and pen inclination

In this section, 6 systems were submitted. Pen coordinates, pen pressure and pen inclination are considered as input data. The experimental results are shown by DET curves (Figures 8 and 9), and Equal Error Rates are reported in Table 5.

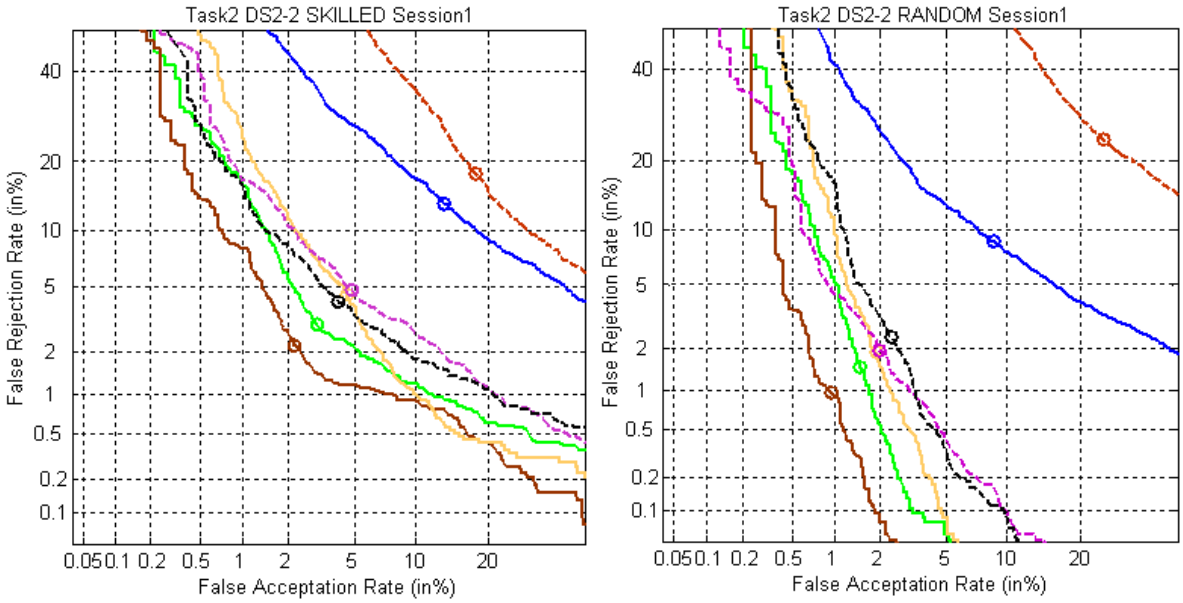


**Figure 8: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 1 of DS2-382 dataset.**

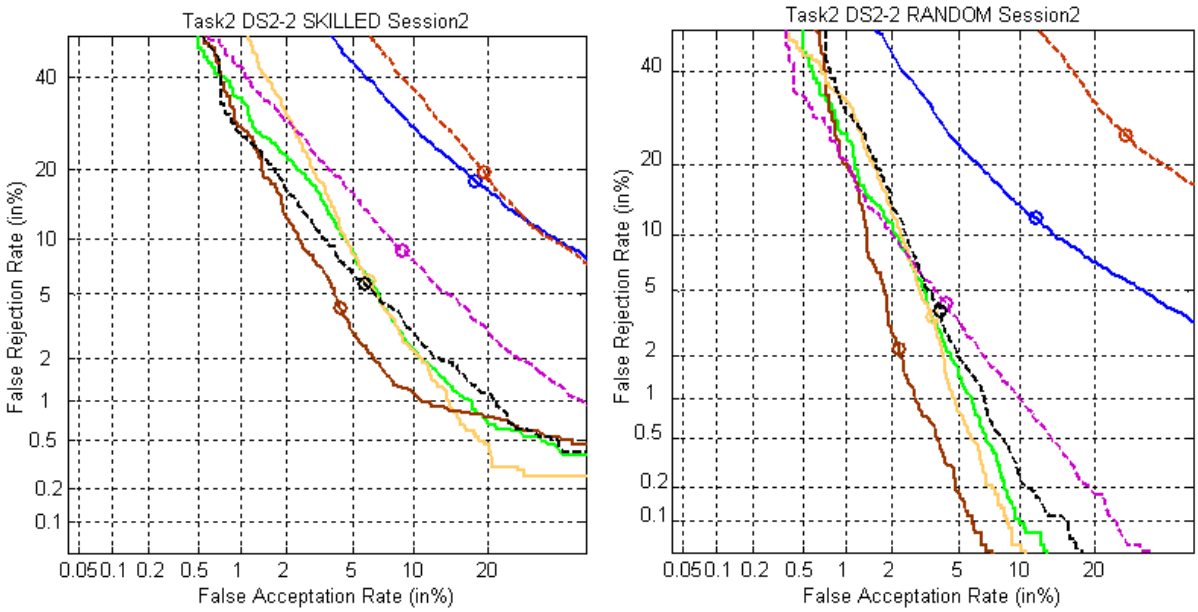


**Figure 9: DET-Curves considering skilled forgeries (left) and random forgeries (right), on Session 2 of DS2-382 dataset.**

**Table 5: Equal Error Rate on DS2-382 dataset, with skilled and random forgeries, considering Session 1 and Session2 separately.**

| 7 Systems | DS2-2 | | | |
|---|---|---|---|---|
| | Session 1 | | Session 2 | |
| | Skilled | Random | Skilled | Random |
| Sys 2 | 4.52 | 1.91 | 5.99 | 3.53 |
| Sys 3 | 13.41 | 8.63 | 17.91 | 11.77 |
| Sys 4 | 3.02 | 1.49 | 6.02 | 3.52 |
| Sys 6 | 2.19 | 0.97 | 4.21 | 2.23 |
| Sys 13 ---- | 17.94 | 24.06 | 19.34 | 25.61 |
| Sys 14 ---- | 4.82 | 1.98 | 8.73 | 4.24 |
| Ref_Sys ----- | 4.07 | 2.39 | 5.72 | 3.87 |

## Analysis of results

- When comparing Figure 8 and Figure 6, we notice that for all systems, adding pen inclination to coordinates and pressure as inputs to the systems, does not improve the performance when no time variability is present.

- Using pen inclination in addition to coordinates and pen pressure, does not improve the robustness of the systems to time variability.

## III. Evaluation 3: Impact of information content

### On Session 1 of DS2-382

In order to study the impact of information content in signatures on algorithms' performance, all systems submitted to Evaluation 1 are tested on DS2-382 dataset on different categories of users depending on the quality of their signatures. Such quality is quantized using our Personal Entropy Measure, defined in [1, 2].

[1] Bernadette Dorizzi, Raffaele Cappelli, Matteo Ferrara, Dario Maio, Davide Maltoni, Nesma Houmani, Sonia Garcia-Salicetti and Aurélien Mayoue, "Fingerprint and On-Line Signature Verification Competitions at ICB 2009", In Proceedings of International Conference on Biometrics (ICB 2009), LNCS Vol. 5558, Springer, p.725-732, June 2009.

[2] S. Garcia-Salicetti, N. Houmani, B. Dorizzi, "A Novel Criterion for Writer Enrolment based on a Time-Normalized Signature Sample Entropy Measure", EURASIP Journal on Advances in Signal Processing , 2009

For this evaluation, we consider the two extreme categories: High Personal Entropy category and Low Personal Entropy category. The first one corresponds to signatures of low quality, and contains signatures of 60 persons. The second category corresponds to signatures of high quality and contains signatures of 161 persons.

For more information about this measure, see our published paper in EURASIP Journal. To this end, go to http://www.hindawi.com/journals/asp/aip.964746.html

The experimental results obtained on Session 1 of DS2-382, on the two extreme categories, are shown by DET curves (Figures 11 and 12), and Equal Error Rates are reported in Table 6.

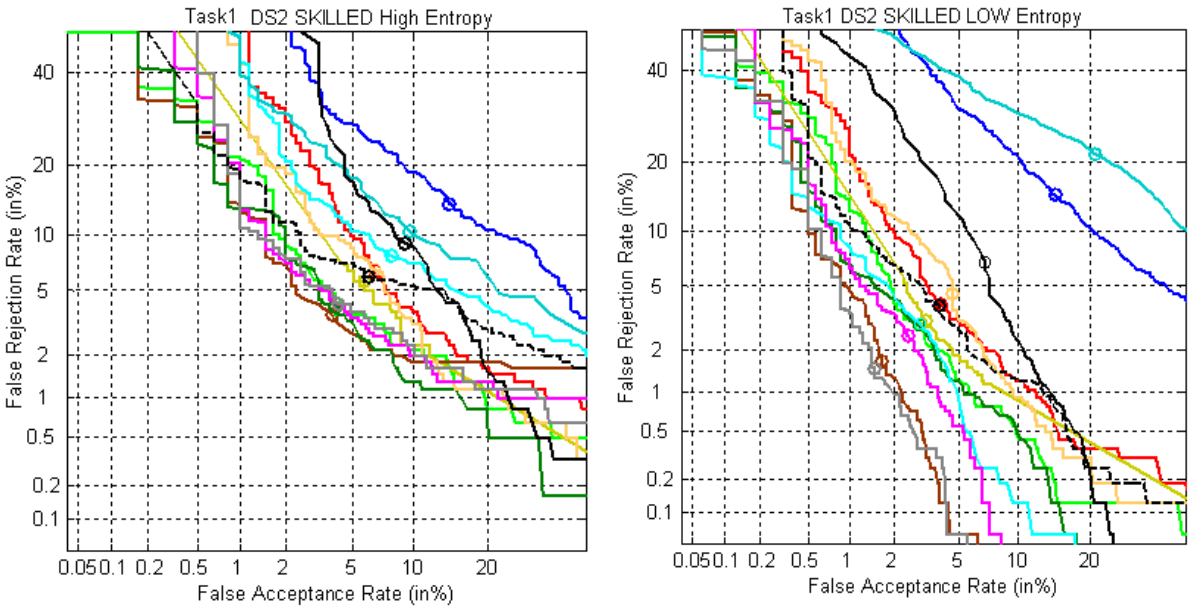


**Figure 11: DET-Curves on each writer category of DS2-382 dataset: High Personal Entropy Category (left) and Low Personal Entropy Category (right), with skilled forgeries considering only Session1.**

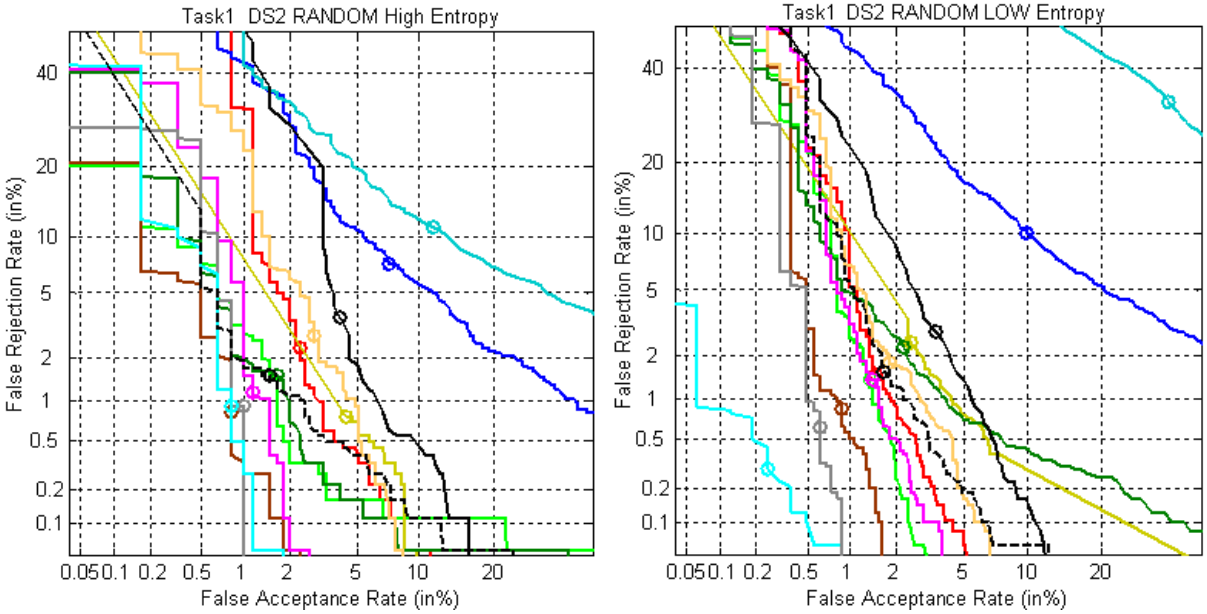


**Figure 12: DET-Curves on each writer category of DS2-382 dataset: High Personal Entropy Category (left) and Low Personal Entropy Category (right), with random forgeries considering only Session1.**

**Table 6: Equal Error Rate on each writer category of DS2-382 dataset, considering only Session 1, and both skilled and random forgeries.**

| 13 Systems | DS2-382 | | | |
|---|---|---|---|---|
| | High Entropy | | Low Entropy | |
| | Skilled | Random | Skilled | Random |
| **Sys 1** | 6.50 | 2.33 | 3.94 | 1.50 |
| **Sys 2** | 6.58 | 2.83 | 4.57 | 1.80 |
| **Sys 3** | 14.00 | 7.22 | 14.5 | 9.98 |
| **Sys 4** | 4.08 | 1.52 | 2.92 | 1.38 |
| **Sys 5** | 5.67 | 2.55 | 3.14 | 2.47 |
| **Sys 6** | 3.75 | 0.83 | 1.68 | 0.87 |
| **Sys 7** | 4.00 | 1.61 | 2.89 | 2.27 |
| **Sys 8** | 7.83 | 0.8 | 2.95 | 0.27 |
| **Sys 9** | 4.17 | 1.19 | 2.48 | 1.42 |
| **Sys 10** | 9.92 | 11.27 | 21.18 | 32.43 |
| **Sys 11** | 9.00 | 3.83 | 6.83 | 3.14 |
| **Sys 12** | 4.17 | 0.91 | 1.49 | 0.62 |
| **Ref_Sys** | 6.00 | 1.52 | 3.81 | 1.62 |

## Analysis of results

- We notice that there is a significant difference in classifiers' performance between the two extreme categories, for both skilled and random forgeries: all systems give the best performance on writers belonging to the category of lowest Personal Entropy, containing the longest, most complex and most stable signatures. At the opposite, all systems give the worst performance on writers belonging to the highest Personal Entropy category, containing the shortest, simplest and most variable signatures.

- Some systems are more robust than others to this quality degradation.

- The ranking of the systems changes according to the category of Personal Entropy used for test.