Independent Performance Evaluation of Biometric Systems

Davrondzhon Gafurov, Bian Yang, Patrick Bours and Christoph Busch
Gjøvik University College,
{Firstname.Lastname}@hig.no

I. Introduction

Often in the development of a biometric product the evaluator of the system is the same institution who developed the algorithm. Furthermore, usually the test data set is also collected by the same developer/evaluator and in most cases such database will not be public. Consequently, test results cannot be verified by independent institutions. Although this can be justifiable (e.g. in the optimization phase of an algorithm), from the perspective of a potential customer it reduces trustworthiness of the developed system and reported performances. Therefore, for performance evaluation the availability of independent databases and desirably independent evaluators are very important. It is also essential that algorithm developers do not have access to the testing database and thus the risk of tuned algorithms is minimized.

Pseudonymous identifiers (PI) are complementary to image- or minutiae- based references and provide a level that is both more privacy protective and more efficient than symmetric or asymmetric encryption of a biometric reference image or minutiae template record [1]. With PI an individual retains complete control of its biometric data as multiple PI can be generated from a single biometric characteristic without any risk that these can be linked together. At the same time any of these identifiers can be cancelled and replaced by a new one if needed. Research on such PI is a core objective of the TURBINE project [2].

This work presents the biometric performance test report on a fingerprint performance evaluation that has been generated in the context of the TURBINE project [2]. It is worth to emphasize that this performance report is results of the first testing round which are not final results of the project. For the second and final testing round, project partners will submit their improved algorithms, and its results will be available in year 2011. Furthermore, here we only report the ”biometric performance” per se of the algorithms while the ”security performance” of the PI algorithms is evaluated by others in TURBINE.

II. Performance metrics and data sets

The main error types associated with any biometric performance are FMR versus FNMR, and FAR versus FRR. In this work we will refer to the former and latter pairs as algorithm and system performances (or errors), respectively. They are related to each other according to the below formulas:

$$FAR = FMR{*}(1{-}FTA), FRR = FTA{+}FNMR{*}(1{-}FTA)$$

In our tests, we define FTA using formula below.

$$FTA = FTC + FTX * (1 - FTC)$$

where FTC (Failure To Capture) and FTX (Failure To eXtract) are estimated as follow

$$FTC = \frac{\text{(\# terminated capture attempts) + (\# not sufficient quality images)}}{\text{Total \# capture attempts}}$$

$$FTX = \frac{\text{(\# not encoded templates)}}{\text{Total \# images submitted to the template encoder}}$$

(”#” stands for ”number of”). It is worth noting that FTA computation incorporates both hardware (in FTC) and software (in FTX) related failures.

The (binary) fingerprint verification algorithms are provided by project partners, in particular Sagem Securite, Precise Biometrics, Philips Research Europe and University of Twente. An external fingerprint verification package by Neurotechnology (VeriFinger [3]) is bought and also included in the testing. The submitted PI software for the first benchmark encompasses software which simulates the effect of a physical protection layer obtained when implementing encoding and comparison within a smartcard (on-card-comparison techniques).

As a test database we use the GUC100 multi-scanner fingerprint database which consisted of fingerprint images of all 10 fingers from 100 subjects (almost 72000 fingerprint images in total) [4]. Neither project partners nor external parties had access to the GUC100 database or were involved in the testing activity. Performance evaluations were carried out solely by the GUC research team as an independent and neutral academic party in the project.

III. Performance results

The focus of this work is not on comparing individual performances of algorithms or scanners but rather emphasizing characteristics of biometric performance evaluation and observing the potential performance degradation in the transition to the PI level. Therefore, the names of scanner (denoted as S1, ..., S6) and algorithm suppliers (except Neurotechnology) are anonymized. Test results are given in terms of the DET-curves. At the minutiae level curves x-axis are plotted in logarithmic scale.

A. *Results at the minutiae level*

Figure 1 shows results of algorithm and system performance evaluations using algorithms from the aforementioned suppliers on the GUC100 database at minutiae level, which is *unprotected* from a data privacy perspective. Figure 2 presents results of same performance evaluations (as in Figure 1) but now by taking image quality into account. We used NFIQ algorithm [5] for estimating fingerprint image quality. We selected NFIQ value 3 as a quality threshold, i.e. fingerprint images with NFIQ score 4 or 5 are considered as having insufficient quality and counted in FTC.

As one can observe from Figures 1 and 2 the difference between algorithm and system performances is insignificant when ignoring image quality but very significant when taking into account image quality (see in Figures 2).

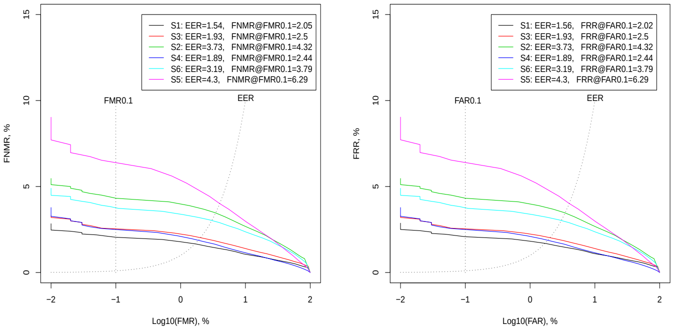

Figure 1. Algorithm and system performances without considering image quality.

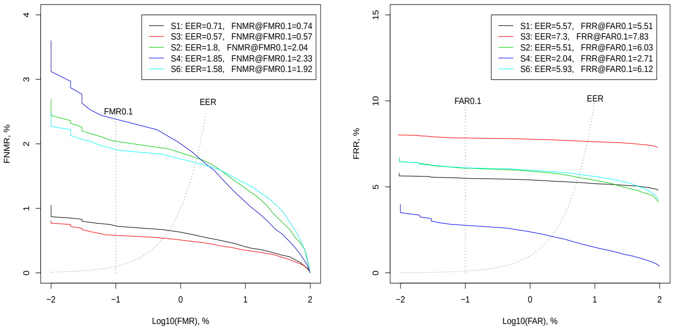

Figure 2. Algorithm and system performances with considering image quality.

B. *Results at the PI level*

Figure 3 shows the first phase performance evaluation results at the pseudonymous identifier level only using images from 3 scanners. At this level performance tests were conducted without considering the image quality metric of a processed sample. The figures indicate that performance deterioration can occur from minutiae level to PI level.

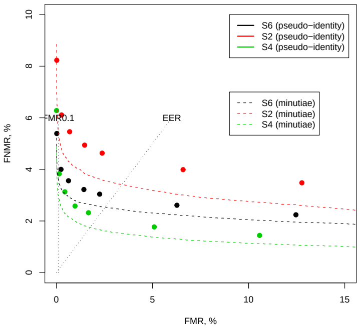

Figure 3. Supplier A - algorithm performance at the minutiae (unprotected) and PI (protected) levels.

For each test scenario at the PI level the testing time is significantly higher than at the minutiae level. Thus, the number of points in PI level DET curves are limited (seven in Figure 3). Also because of small number of points, the single number performance indicators (e.g. EER) are not estimated neither.

IV. Conclusion

In this work we presented an independent report on the first phase performance evaluation of fingerprint verification algorithms in the context of the TURBINE project. Performance testing was conducted both at the minutiae and pseudonymous identifier levels using GUC100 database which consisted of almost 72000 fingerprint images from 100 individuals. Algorithm developers did not have access to this database or were involved in the testing activity. All the performance evaluation tests were conducted independently by a neutral academic party in the project.

For increasing the trustworthiness of biometric performance report for potential customers, it is recommended/desired that evaluations to be conducted by an independent third party.

References

[1] J. Breebaart, C. Busch, J. Grave, and E. Kindt. A reference architecture for biometric template protection based on pseudo identities. In *BIOSIG 2008*, 2008.

[2] TURBINE project - TrUsted Revocable Biometric IdeNtitiEs. http://www.turbine-project.eu/index.php.

[3] Neurotechnology's verifinger 6.0. http://www.neurotechnology.com/. Last visit: 14.10.2009.

[4] GUC100 multi-scanner fingerprint database for in-house (semi-public) performance and interoperability evaluation. http://www.nislab.no/guc100.

[5] M.D. Garris, E. Tabassi, and C.L. Wilson. NIST Fingerprint evaluations and developments. *Proceedings of the IEEE*, 2006.