Handwritten Signature On-Card Matching Performance Testing

Olaf Henniger[1] and Sascha Müller[2]

[1] Fraunhofer Institute for Secure Information Technology, Darmstadt, Germany
olaf.henniger@sit.fraunhofer.de
[2] Technische Universität Darmstadt, Darmstadt, Germany
mueller@sec.informatik.tu-darmstadt.de

Abstract. This paper presents equipment and procedures for on-card (in-situ) performance testing of biometric on-card comparison implementations using pre-existing databases of biometric samples. A DTW-based on-line signature on-card comparison implementation serves as an example test object. The test results presented are false match rates and false non-match rates over a range of decision thresholds on a per-test-subject basis. The results reveal considerable differences in the comparison-score frequency distribution among test subjects, which necessitates the setting of user-dependent decision thresholds or comparison-score normalization.

Keywords: On-card comparison, biometric performance testing, handwritten signature.

1 Introduction

If sufficiently resistant against direct and indirect attacks, biometric methods can be deployed for user authentication purposes in smart cards that provide security-relevant functions (like the creation of electronic signatures or the authorization of transactions) or carry data worthy of protection (like medical data). An important component of evaluating the security of a biometric system is testing its performance in terms of error rates.

This paper presents means of testing the performance of on-card comparison implementations using databases of biometric samples. In contrast to [1], on-card comparison is not emulated in a software library installed on a PC; the testing is rather conducted on physical cards. The advantage of this is that, by moving the point of control and observation closer to the implementation under test (IUT), the confidence in its correct functioning may increase.

The paper also reports about performance test results for an on-line signature on-card comparison prototype. Moreover, it closely investigates the issue why do behavioral biometric characteristics require either the setting of user-dependent decision thresholds [2] or the normalization of comparison scores [3].

The rest of the paper is organized as follows: Section 2 introduces the on-card comparison IUT. Section 3 reports which aspects of performance are measured. Section 4 introduces the test database used. Section 5 deals with the means of

J. Fierrez et al. (Eds.): BioID_MultiComm2009, LNCS 5707, pp. 276–283, 2009.

testing, i.e. equipment and procedures used. Section 6 details the test results for the IUT. Section 7 compares strengths and weaknesses of actual on-card testing and off-card emulation. Section 8 summarizes the results and gives an outlook.

2 Implementation under Test

The IUT is a prototype of an on-card comparison algorithm for handwritten signatures [4]. It uses a dynamic time warping (DTW) algorithm with Sakoe/Chiba band [5]. DTW can be considered among the best approaches for on-line signature verification [6,7]. The data sent to the card are time series for x and y coordinates in the compact format of [8]. The sample rate is 100 per sec.

The implementation platform are Java cards, i.e. smart cards with a Java Card Virtual Machine. Java cards are good for prototyping smart card applications, but their computing speed is limited because Java byte code is interpreted at run-time. The same comparison algorithm has been installed on different Java cards with different computing speeds.

At any one time, a single signature is stored on the card as reference signature. The storage space for the reference signature is restricted and cannot hold signatures that take longer than 5 sec to sign. For testing purposes, no retry counter is used, i.e. the biometric verification method is not blocked after a number of failed verification attempts. For faster testing, the IUT also does not prompt for user authentication prior to changing the reference data.

For testing purposes, the IUT returns the comparison score of each comparison of a probe signature with a reference signature as two-octet unsigned integer. The comparison scores are distance (or dissimilarity) scores, i.e. decrease with similarity. The IUT tries (in vain) to normalize the distance scores by dividing each DTW distance by the number of sample points of the reference signature.

3 Performance Metrics

The false non-match rate (FNMR) of a biometric verification system is the proportion of genuine attempts falsely declared not to match the biometric reference. The false match rate (FMR) of a biometric verification system is the proportion of impostor attempts falsely declared to match the biometric reference [9]. Error rates involving impostor attempts can be measured using either

- *zero-effort attempts* where impostors present their own biometric characteristics as if attempting successful verification against their own reference or
- *active impostor attempts* where impostors intentionally imitate the biometric characteristics of other persons.

In cases where impostors may easily imitate aspects of the required biometric characteristics, such as handwritten signatures, an FMR measurement based on active-impostor attempts is more predictive of the performance in practice. Therefore, here the active-impostor FMR is considered.

FNMR and FMR depend on an adjustable decision threshold determining the required degree of similarity of probe sample and reference. The lower the FMR at a certain threshold value, i.e. the fewer forgeries are accepted, the higher is the FNMR, i.e. the fewer genuine samples are accepted as well, and vice versa. Depending on the concrete requirements, an appropriate threshold value has to be chosen, reconciling security (low FMR) with usability (low FNMR).

A receiver operating characteristic (ROC) curve shows the rate of impostor attempts accepted (FMR) on the x axis against the corresponding rate of genuine attempts accepted ($1-$FNMR) on the y axis, plotted parametrically as a function of the decision threshold [9].

To find out the worst-case performance for individuals, the ROC curves are plotted on a per-test-subject basis. This requires multiple genuine and impostor samples for each test subject.

A significant performance metric of a biometric system is the equal error rate (EER), even though it does not summarize all characteristics of an ROC curve. The EER is the error rate at that threshold value where FNMR = FMR. When the threshold is moved away from that value, then either FNMR or FMR deteriorate beyond EER.

Because FMR and FNMR of a biometric system could be improved by abstaining from processing low quality samples, also the failure-to-enrol (FTE) rate is reported. The FTE rate of a biometric system is the proportion of enrolment attempts for which the system fails to complete the enrolment process.

4 Test Corpus

The error rates can be estimated experimentally with some statistical significance using large test databases. The corpus of samples used is a publicly available subset of the database [10] consisting of signature samples of $n = 100$ test subjects. For each test subject, there are $m_g = 25$ genuine samples and $m_f = 25$ skilled forgeries. For the forgery attempts, the impersonators had the original signatures available on paper and were allowed to imitate them to the best of their ability. They were allowed to practice the signatures to be forged, to look at the original while forging, and even to retrace the original.

5 Test System

The tests are controlled by a software called KARMASYS (card-manipulation system). KARMASYS allows, via an off-the-shelf card terminal, sending commands to a smart card and recording the responses received from the card. Fig. 1 shows an overview of the test equipment.

A KARMASYS test script comprises a sequence of application protocol data units (APDUs) to be sent to the card under test. Change Reference Data APDUs are used for changing the reference data stored on the smart card; Verify APDUs are used for transmitting probe samples to the card [11]. For compliance with [11], signature time series data blocks that are too large for a single regular-length APDU are split into chains of APDUs.

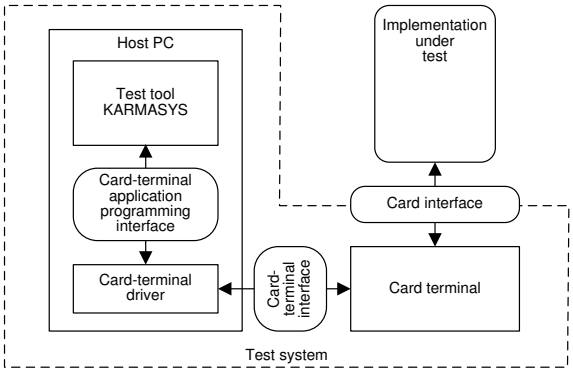


Fig. 1. Architecture of the test system

For each of the $n = 100$ test subjects, a test script was created using `awk` scripts and batch files. The output is a sequence of APDUs for testing pairs of signature samples of a test subject. The test script for a test subject includes

- $m_g = 25$ CHANGE REFERENCE DATA APDU chains,
- $\sum_{i=1}^{m_g}(i-1) = 300$ VERIFY APDU chains containing genuine signatures (as DTW is commutative, there is no need to compare the genuine signatures (g_k, g_j) after comparing (g_j, g_k)), and
- $m_g \cdot m_f = 625$ VERIFY APDU chains containing forged signatures.

6 Performance Test Results

6.1 Failure-to-Enrol Rate

For 30% of all test subjects a failure to enrol occurred because at least one of their genuine signature time series data blocks was too long to be stored on the smart card as reference data. Provided that long signatures are more difficult to forge than shorter ones, these failures are detrimental to the overall FMR.

6.2 Per-Test-Subject Performance

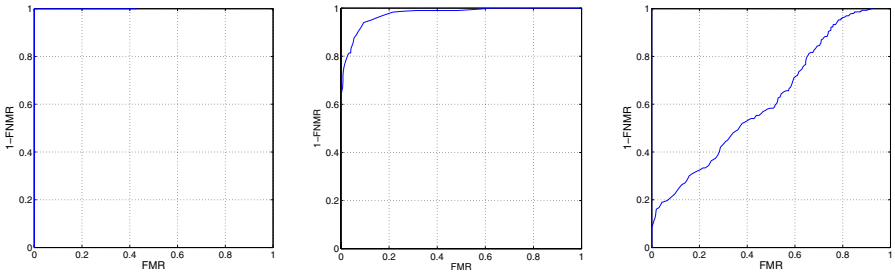
Fig. 2(a) through 2(c) show examples of per-test-subject performance test results in ROC curves. Considerable individual differences show up. The per-test-subject EERs of all enrolled test subjects range from 0% to 44.7%:

- For four test subjects (5.7% of all 70 enrolled test subjects) whose signatures are long and of high complexity as in Fig. 3(a), perfect discrimination between genuine and forged signatures appears possible, i.e. all forgery attempts resulted in a higher distance score than any genuine signature verification attempt. The EER is 0% in this case, and the ROC curve passes through the upper left corner of the diagram, cf. Fig. 2(a).

– For two test subjects whose signatures are short and of low complexity as in Fig. 3(b), the EER is above 40%, cf. Fig. 2(c). For these test subjects the outcome of signature verification is nearly as random as tossing a coin.

These differences may be due to inter-individual differences in the stability (intra-individual invariance) and the forgeability of signatures. They may also be due to differences in the skills applied when attempting to forge the signatures. Quality scores [12] can be used for deciding whether a handwritten signature is long and complex enough to be hard to forge.

Fig. 4 shows the relative frequency distribution of per-test-subject EERs. For a third of all enrolled test subjects, the EER amounts to less than 5%. The median value of the per-test-subject EERs of all enrolled test subjects (i.e. the



(a) Test subject 42 (b) Test subject 91 (c) Test subject 15

Fig. 2. ROC curves

(a) Test subject 42 (long, complex) (b) Test subject 15 (short, simple)

Fig. 3. Signature examples

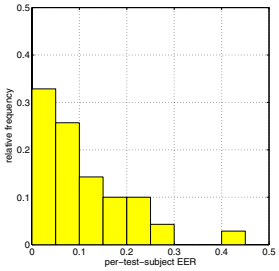


Fig. 4. Histogram of per-test-subject EERs

value separating the higher half of the values from the lower half) is 8.3%. The mean value of the per-test-subject EERs of all enrolled test subjects is 10.7%; their standard deviation is 9.6%.

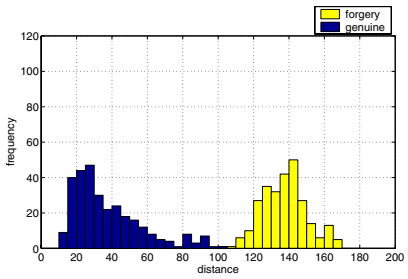
6.3 Overall Performance

The decision threshold can be set individually for each user or chosen to be identical for all users. A common decision threshold for all users has the advantage that it needs to be set only once for all. However, a common threshold can achieve an acceptable overall performance only if the frequency distribution of the comparison scores is nearly the same for all users. Otherwise, the optimal thresholds for distinguishing between genuine and forgery samples lie at different threshold values for different users.

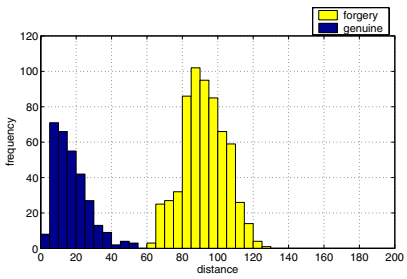
In order to achieve an acceptable overall performance in case of different comparison-score frequency distributions among users,

- the threshold should be set individually for each user based on the similarity of the samples presented during enrolment [2] or
- the comparison scores should be normalized in such a way that their frequency distribution becomes similar for all users [3].

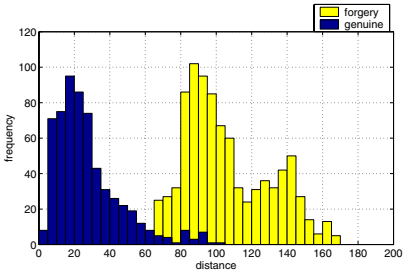
For illustration, see the distance score distributions for the test subjects 42 and 62 in Fig. 5(a) and (b). For each of them, all genuine distance scores are smaller than any forgery distance score and an optimal threshold value can be found that



(a) Test subject 42



(b) Test subject 62



(c) Test subjects 42 and 62 together; genuine and forgery score distributions overlap

Fig. 5. Distance score histograms

allows perfect discrimination between genuine signatures and forgeries. The two optimal threshold values, however, are different. If a common threshold value is to be used for both test subjects, then, no matter which value is chosen, false matches and false non-matches are unavoidable, see Fig. 5(c). This effect is aggravated when all test subjects are taken into consideration. If the tested prototype operated with a common decision threshold for all test subjects, its overall EER were 19.1%, i.e. significantly worse than the median and mean EER values reported in Section 6.2.

For score normalization the user should present several signatures at enrolment time. The one with the smallest average distance to all other presented signatures should be chosen as reference signature. At verification time, each distance score should be divided by the average distance of the reference signature to that user's other signatures presented at enrolment [6].

6.4 Time Needed

The time needed for transmitting probe samples to the card and for on-card comparison using the DTW algorithm grows linearly with the size of the probe samples. That means, if it takes 10 sec on the fastest tested Java card to verify a signature which it took 1 sec to write, then it takes 50 sec to verify a signature which it took 5 sec to write. As different Java cards with different speed have been used during the test campaign, here the time needed is not evaluated statistically.

7 On-Card Testing vs. Off-Card Emulation

While off-card emulation needs less time than actual on-card (i.e. in-situ) testing, in-situ testing allows measuring additional performance and robustness indicators, such as the actual time needed, and detecting errors that remain hidden in case of off-card emulation. For instance, the following error in the IUT has been revealed by in-situ testing: A frequently used loop-control variable got allocated EEPROM space instead of RAM space. This slowed down operations and repeatedly lead, after several thousand signature comparisons, to the destruction of the misused EEPROM cells and thus of the cards. After allocating RAM space to the loop-control variable, the life time of the cards greatly increased (no outage since then).

8 Summary and Outlook

The performance of biometric systems depends on the algorithms, the environment, and the population. Therefore, caution must be exercised when comparing these results with those of other systems tested using different test data.

There is ample potential for improvement of the prototype IUT:

- Processing time: 10 sec per 100 sample points is too long. A speed-up is expected from faster smart cards and optimizations of the algorithm.
- FTE rate: To avoid failures to enrol, more memory space is needed for reference signatures.

- Frequency distribution of distance scores: This differs from person to person for the IUT. The distance scores need to be normalized to allow using a uniform decision threshold.
- Quality control: Too short and too simple signatures should be rejected at enrolment time to make forgeries more difficult.

The performance of future versions of the handwritten signature on-card comparison implementation can be tested, largely automatically, using the presented test equipment and procedures. The test equipment and procedures can also be used, with different corpora of biometric samples, to test the performance of on-card comparison implementations for other biometric characteristics.

References

1. Grother, P., Salamon, W., Watson, C., Indovina, M., Flanagan, P.: MINEX II – Performance of fingerprint match-on-card algorithms – Phase II report. NIST Interagency Report NISTIR 7477, NIST, Gaithersburg, MD, USA (2008)
2. Jain, A., Griess, F., Connell, S.: On-line signature verification. Pattern Recognition 35, 2963–2972 (2002)
3. Fiérrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Target dependent score normalization techniques and their application to signature verification. In: [13], pp. 498–504
4. Henniger, O., Franke, K.: Biometric user authentication on smart cards by means of handwritten signatures. In: [13], pp. 547–554
5. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-26(1), 43–49 (1978)
6. Kholmatov, A., Yanikoglu, B.A.: Identity authentication using improved online signature verification method. Pattern Recogn. Letters 26(15), 2400–2408 (2005)
7. Yeung, D.Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., Rigoll, G.: SVC2004: First international signature verification competition. In: [13], pp. 16–22
8. Information technology – Biometric data interchange formats – Part 7: Signature/sign time series data. International Standard ISO/IEC 19794-7 (2007)
9. Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. International Standard ISO/IEC 19795-1 (2006)
10. Ortega-Garcia, J., Fiérrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.J., Vivaracho, C., Escudero, D., Moro, Q.I.: MCYT baseline corpus: a bimodal biometric database. IEE Proceedings Visual Image Processing 150(6), 395–401 (2003)
11. Information technology – Identification cards – Part 4: Organization, security and commands for interchange. International Standard ISO/IEC 7816-4 (2004)
12. Müller, S., Henniger, O.: Evaluating the biometric sample quality of handwritten signatures. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 407–414. Springer, Heidelberg (2007)
13. Zhang, D., Jain, A.K. (eds.): ICBA 2004. LNCS, vol. 3072. Springer, Heidelberg (2004)