A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Influence of Test Protocols on Biometric Recognition Performance Estimation

Teodors Eglitis[1] , Emanuele Maiorana[1] , Patrizio Campisi[1]

Abstract: The performance of a biometric system is commonly evaluated by the obtained recognition rates and comparing the results against the ones reported in the literature on the same database. An aspect that has not received the deserved attention in the literature concerns the influence, on the achieved rates, of the test protocol employed to select the enrol and probe data. We provide a detailed analysis of the impact of the experimental choices on the estimated performance, considering the recommendations provided by ISO/IEC 19795 standard. We use the UTFVP finger vein database, reproducing results presented in the literature using multiple protocols. Our experiments highlight the possibility of obtaining equal error rates reduced by half simply by changing the test protocol.

Keywords: Database protocols, reproducible research, vascular biometrics.

1 Introduction

The recognition capabilities of a biometric system are evaluated by running tests on datasets of biometric samples captured from a set of subjects. To this aim, in-house databases are often collected, especially for innovative modalities at an early stage of development. The availability of public datasets enables researchers to perform in-depth research and reproduce others' work on more established modalities according to a test protocol, which determines how the considered data are used. However, the details regarding such employed protocols are often not provided with due care, making it hard to compare the new results against those previously achieved in literature, even when performing tests on the same data.

In this paper, we conduct an extensive analysis of the impact of the used experimental protocols on estimating a biometric recognition system performance. We want to highlight the importance of adequately describing the procedure followed when conducting tests on a given database by investigating the extent to which recognition rates may vary depending solely on how the considered data is exploited. We consider the recommendations of the ISO/IEC 19795 standard "Biometric performance testing and reporting" [In06] when defining the testing procedures. Vascular biometrics is used as the reference scenario. This modality has recently attracted considerable attention from industry and academia due to its advantages over more traditional biometric traits, with an ever-increasing number of papers published recently. At least eleven public databases containing finger-vein samples,

[1] Department of Industrial, Electronic, and Mechanical Engineering, Roma Tre University, Via Vito Volterra 62, 00146 Rome, Italy
{teodors.eglitis, emanuele.maiorana, patrizio.campisi}@uniroma3.it

and eight databases with palm-vein images, are currently publicly available [Uh20]. The University of Twente Finger Vascular Pattern (UTFVP) database [TV13], one of the first publicly available finger-vein datasets and among the most cited ones, is employed in this paper to assess the influence of the used test protocol on the recognition rates.

The remainder of this paper is structured as follows: Section 2 summarizes the testing recommendations provided by the ISO/IEC 19795 standard. Section 3 describes the database, recognition approach, the test protocols, and the performance metrics used. Finally, a discussion on the obtained results and conclusions are given in Section 4.

2 Test Protocol Recommendations

Guidelines for designing and testing biometric recognition systems are provided by the ISO/IEC joint technical committee (JTC) 1/SC 37. Standards for test protocols are defined in the ISO/IEC 19795, “Information technology – Biometric performance testing and reporting” documents, currently consisting of ten parts. Those relevant for our study are “Part 1: Principles and framework” [In06], first published in 2005, and “Part 2: Testing methodologies for technology and scenario evaluation" [In07], initially released in 2006, which describe the recommended scientific practices for technical performance testing. Recommendations from ISO/IEC 19795-1 [In06] for the definition of test protocols can be summarized as follows:

- the test phase should be conducted on data unavailable during algorithm development [In06, § 5.5.3.a];
- collection of enrolment and probe data should be separated at least by days [In06, § 6.5.5];
- when reporting error rates, the “rule of 3” and “rule of 30” [In06, § 6.6.1], which relate the number of probes with the achievable error confidence intervals, should be taken into account. It is remarked that handling ten probes for ten subjects is not equivalent to having a hundred subjects each with only a single probe, although, for certain protocols, this produces an equal number of comparisons;
- data from the same subject and the same modality, yet different instances (e.g., distinct eyes, fingerprints, finger-veins) can be used to represent distinct users [In06, § 6.6.3.b];
- collected samples should be excluded from the database only if a predetermined criterion is violated [In06, § 7.1.6];
- each test subject should be enrolled only once [In06, § 7.3.1.1];
- impostor comparisons involving data captured from the same subject (e.g., vascular data from different fingers of the same person, representing different *virtual* users) should not be performed because intra-individual data are likely to contain more similarities than data from different individuals [In06, § 7.6.1.3];
- zero-effort impostors can be selected by randomly choosing biometric templates or by doing a full cross-comparison [In06, § 7.6.3.1.1];
- enrolment templates can be used as impostor data in case different feature extractors are applied to enrolment and probe samples [In06, § 7.6.3.3.b].

Several of the ISO/IEC 19795 recommendations mentioned above, e.g., enrolment and probe data being captured at different days, or computing a minimum number of comparisons to validate error rates, are often not respected in the employed test protocols, thus affecting the reliability of the reported performance.

Additional suggestions on the test protocols to be used have been proposed in the literature. For instance, when evaluating biometric systems performing verification, in [JKR15] it has been suggested to use training, validation, and testing sets derived from different subjects, to avoid positive bias in the estimation of performance such as false match rate (FMR), false non-match rate (FNMR), and equal error rate (EER). [Ma15] (published in 2015) recommends using a Receiver Operating Characteristic (ROC) and Detection Error Tradeoff (DET) curves; providing False Acceptance Rate (FAR) in the range $\{10^{-4}, 10^{-2}\}$; compare algorithms using verification (1:1) experimental setup instead of 1:N, arguing that 1:1 comparisons more clearly indicates the algorithm effectiveness (if the specific research does not concern identification), and disclose detailed information about software, database, algorithms, and computational efficiency. Paper [MZB16], published in 2016, analyze different methods of data division as enrol and probe data, namely hold-out (selecting percentage of the data as probe samples), cross-validation (using n folds and repeating calculations n times, every time using the different fold as probe data) and leave-one-out methodology (cross-validation where the number of folds equals the number of samples in the dataset). Authors summarize published results and offer their own, using different data division scenarios on three face databases.

Our investigation is similar to the [MZB16], the main differences are that we focus on more exotic data division in protocols often used in vascular biometrics experiments, we summarize and follow the recommendations provided by the ISO standards, hoping that our research will be beneficial to the novices in the field.

3 Method

The UTFVP database, upon which the performed tests are conducted, comprises data recorded from 60 subjects. Two samples from three fingers (index, ring, and middle finger) of both hands have been captured during two sessions separated by 15 days for each of the involved individuals. For each finger, images 1-2 are obtained in the first recording session, and images 3-4 in the second one, for a total of 4 biometric samples. The database, therefore, consists of 360 different finger-vein classes for a total of 1440 vascular pattern images. Samples from UTFVP are processed using the maximum curvature (MC) feature extractor [MNM07]. The similarity score between two templates is evaluated using the Miura match (MM) algorithm [MNM07]. Such comparison is not symmetrical and can generate different scores if the two templates are switched in places.

Biometric systems working in verification modality have been considered, with the EER used to characterize their recognition performance. A summary of EERs reported in papers using MC for feature extraction and MM for template comparison is given in Table 1. These results already show the significant variability exhibited in literature for tests conducted on the same database with the same processing pipeline, yet resorting to different test protocols. Nonetheless, since a different number of classes has been considered in the referenced papers, it is impossible to properly evaluate the influence of the employed test protocols on the obtained performance analyzing these data. Conversely, our analysis investigates the test protocols used in the papers mentioned in Table 1, whose details are given in the following, while keeping as unaltered as possible any other aspect in the performed comparisons. Furthermore, additional testing strategies highlight the variability of the achieved performance depending on the employed protocol.

Tab. 1: Reported recognition results, UTFVP database, MC extractor, MM matcher.

Paper	# classes	# gen. comp.	# imp. comp.	EER (%)
[TV13]	325	1950	842 400	0.4
[Va14]	325	3900	1 684 800	0.49
[Va14]	108	216	46 224	1.39
[Id21]	360	5760	2 067 840	0.6 [1]
[Id21]	325	3900	1 684 800	0.7 [1]
[Id21]	192	768	146 688	1.1 [1]
[KRU14]	35	210	9 520	6.443
[KU20]	360	3600	64 620	0.37
[KU15]	360	2160	10 620	0.6 [2]

3.1 Considered Test Protocols

A test protocol is defined specifying which similarity scores are computed to estimate the achievable recognition capabilities. The possible score types are depicted in Figure 1, showing a confusion matrix obtained comparing samples belonging to two classes from the UTFVP database. The following groups of scores, identified with the symbols reported hereafter:

╲ : scores generated by genuine comparisons, with the enrolment image compared against itself, resulting in a perfect similarity. Such scores correspond to the diagonal line of the confusion matrix, with 4 scores for each class in UTFVP;

◹ : genuine scores obtained comparing a probe sample with an enrolment sample having a lower index number (e.g., image 1 of finger 1 serves as enrolment template and image 2 of finger 1 as probe). Such group comprises genuine scores located above the diagonal line (╲) of the confusion matrix;

◺ : genuine scores located below the diagonal line (╲) of the confusion matrix. For every class, there are 6 ◹ and 6 ◺ scores;

◥ : impostor scores, located above the diagonal line ╲. Such scores are calculated comparing probe samples with enrolment samples having a lower class index (e.g., enrolment image 2 of finger 1 compared with the probe image 1 of finger 2);

◣ : impostor scores, located below the diagonal line ╲.

If scores from only one side of the diagonal (◹ or ◺; ◥ or ◣) are used, then all the samples in a database are compared only once. For consistency, if genuine scores from only *one side* (e.g., ◹) are considered, the same *side* for impostor scores (e.g., ◥) is also taken into account in our analysis. We evaluate testing strategies implemented in two of the most commonly used open-source, reproducible research frameworks available to test vein-based biometric recognition systems: PLUS OpenVein Toolkit (PLUS) [KU20], written using Matlab, and BOB [An17], a comprehensive signal processing framework, writ-

[1] Reproduced in this work

[2] Employs histogram equalization in preprocessing

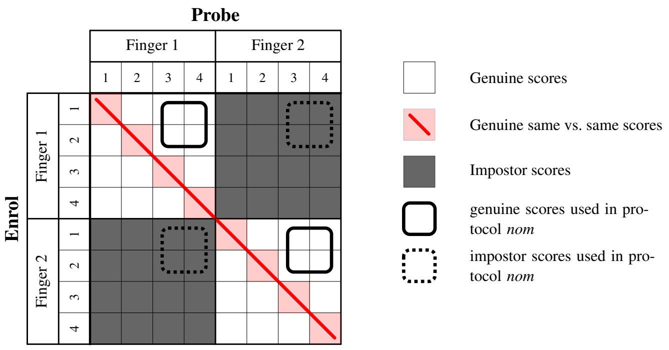

Probe
Finger 1
Finger 2
1
2
3
4
1
2
3
4
Enrol
Finger 1
Finger 2
Genuine scores
Genuine same vs. same scores
Impostor scores
genuine scores used in protocol *nom*
impostor scores used in protocol *nom*

Fig. 1: Graphical representation of the considered comparison scores.

ten in Python, designed for biometric experiments. The specific library dedicated to vein recognition is *bob.bio.vein*. In more detail, the test protocols executable in the aforementioned open-source frameworks, and employed in the performed tests on UTFVP, are:

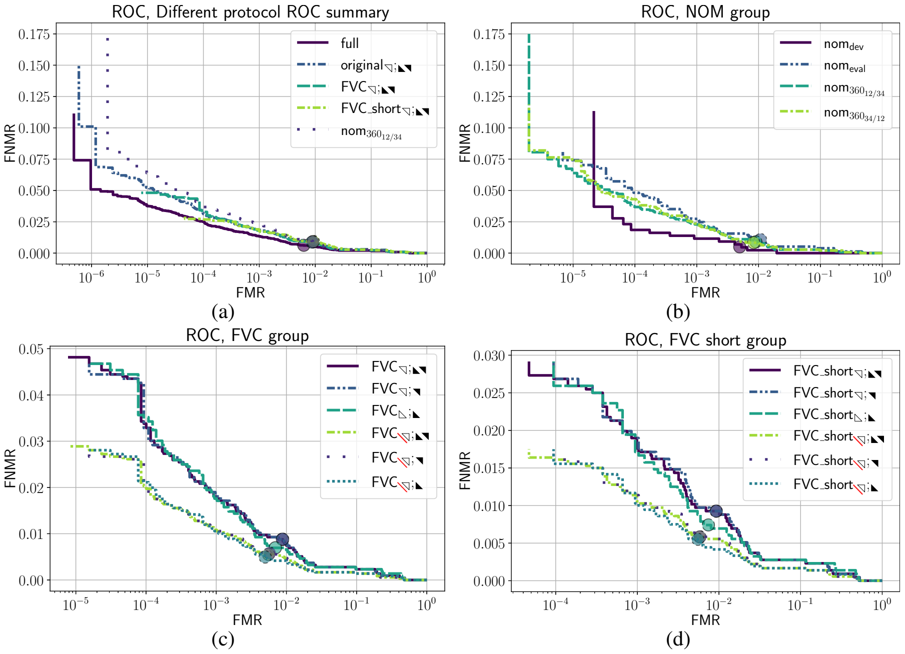
- **original** [TV13]: considering the UTFVP database, it reserves 35 class data for parameter tuning, and uses the remaining 325 classes for performance evaluation. Unsymmetrical genuine comparisons (◹ or ◺) are employed to estimate recognition capabilities, resulting in $325 \cdot 6 = 1\,950$ genuine scores;
- **FVC** (PLUS): derived from the FVC2004 fingerprint verification contest, uses all 360 classes for testing. Since there are discrepancies between the description and formulas in [KU20] and the source code [Ka21], it is unclear whether the protocol uses ╲ genuine scores. Thus, different genuine score combinations are here used;
- **FVC_short** (PLUS): unlike FVC, a reduced number of impostor scores is considered, with only the first image from the same fingers as the enrolled sample used as impostor probe (e.g., left hand, middle fingers).
- **full** (BOB): considers all possible comparisons from 360 classes, including those in ╲, therefore consisting of $(360 \cdot 4)^2$ computations, with $360 \cdot 4 \cdot 4 = 5\,760$ genuine and $360 \cdot 4 \ \cdot 359 \cdot 4 = 2\,067\,840$ impostor scores;
- **1vsall** (BOB): analogous to *full*, but using only 325 class data, excluding those used for parameter estimation in the *original* protocol;
- **nom** (BOB): designed according to ISO/IEC 19795-1 suggestions, with Session-1 data used as enrolment templates, and Session-2 data as probes [Id21]. Furthermore, as recommended in [JKR15], the 60 available subjects are split into three disjoint subsets:
  - the train subset comprises samples from 10 subjects (60 fingers), used for setting the feature extractor parameters;
  - the development subset comprises samples from 18 subjects (108 fingers), used for parameter determination, including system threshold;
  - the evaluation subset comprises data from 32 subject (192 fingers), employed to estimate the achievable performance.

  We also report results derived from comparisons carried out on all the available 360 finger-vein classes, denoting with nom360_{S1vsS2} the use of Session-1 data for enrolment and Session-2 data for probes, and vice versa for nom360_{S2vsS1}

Since it often happens that not enough information about impostor scores is provided, for protocols *original*, *FVC* and *FVC_short* we explore { ◥; ◣; ◣◥ }. All the scores needed in the considered test protocols are computed exploiting the BOB framework with *full* protocol. We then select the specific scores needed for each protocol and compute the associated results[3]. Such an experiment ensures that the only aspect varying between different tests is the used protocol, with no other implementation detail impacting the results reported in the next section.

4 Discussion and Conclusions

The obtained EERs are summarized in Table 2, while Figure 2 depicts the associated ROC curves in terms of FNMR vs FMR. For protocols involving all the 360 available subjects, the most notable difference is between FVC⟍◹;◣ and FVC_short◹;◣◥, with the latter EER being 86% worse than the former.



(a) (b) (c) (d)

Fig. 2: ROC curves for the performed tests (FMR and FNMR reported in absolute values, not as percentage). Circles indicate EER values.

It has to be noted that ⟍ scores should not be considered to estimate recognition rates since they can severely distort the obtained results, as shown in Table 2 and by the ROC curves in Figures 2c and 2d. A considerable impact on recognition performance is produced from choices regarding the employed impostor scores, using a single probe for each impostor resulting in a misleading worsening of the performance obtained in our tests. Moreover, Figures 2c and 2d show that, when adopting a non-symmetrical comparison approach such as the MM, selecting only ◥ or ◣ impostor scores may have a significant influence on the obtained results, and therefore both groups should be considered to represent an average behaviour.

[3] Code available at `https://gitlab.com/biomedia4n6-public/biosig2021-influence-of-test-protocols/`

The ISO standard suggestion of using data from different acquisition sessions as enrolment and probe samples should be followed whenever possible. The results obtained using the full (*All vs All*) protocol are notably better than those referred to a $\text{nom}_{360_{S1vsS2}}$ approach. Nonetheless, only this latter resembles how a biometric system works in real-life applications. Such observation is also in line with the ISO standard suggestions arguing that generating more scores from fewer subjects is not equivalent to having more subjects with the same number of comparison scores. It is also to be remarked that the obtained results could depend on which session is employed to provide enrolment data, as observed in Figure 2a, where protocol $\text{nom}_{360_{S1vsS2}}$, using Session-1 data for enrolment and Session-2 as probes, turns out to be more challenging than $\text{nom}_{360_{S2vsS1}}$. The observed behaviour confirms the need for collecting multi-session databases to test biometric systems properly.

Tab. 2: Recognition results obtained exploiting scores generated according to the *full* protocol. Protocols with less than 360 classes are shaded. Corresponding ROC curves are shown in Figure 2.

Protocol	# classes	# gen. comp.	# imp. comp.	EER (%)
full ◺╲◹;◣◥	360	5 760	2 067 840	0.6238
1vsall ◺╲◹;◣◥	325	5 200	1 684 800	0.6731
parameter tuning	35	210	9 520	0.1103
original◹;◣◥	325	1 950	1 684 800	0.9149
original◹;◥	325	1 950	842 400	0.9231
original◺;◣	325	1 950	842 400	0.7692
FVC◹;◣◥	360	2 160	129 240	0.8797
FVC◹;◥	360	2 160	64 620	0.8793
FVC◺;◣	360	2 160	64 620	0.6946
FVC╲◹;◣◥	360	3 600	129 240	0.5556
FVC╲◹;◥	360	3 600	64 620	0.5780
FVC╲◹;◣	360	3 600	64 620	0.4968
FVC_short◹;◣◥	360	2 160	21 240	0.9267
FVC_short◹;◥	360	2 160	10 620	0.9244
FVC_short◺;◣	360	2 160	10 620	0.7423
FVC_short╲◹;◣◥	360	3 600	21 240	0.5836
FVC_short╲◹;◥	360	3 600	10 620	0.5836
FVC_short╲◹;◣	360	3 600	10 620	0.5508
nom_{dev}	108	432	46 224	0.4954
$\text{nom}_{\text{eval@dev}}$	192	768	146 688	1.0781
$\text{nom}_{360_{S1vsS2}}$	360	1 440	516 960	0.9032
$\text{nom}_{360_{S2vsS1}}$	360	1 440	516 960	0.8333

It can be observed that the "rule of 3" and "rule of 30" mentioned in the ISO specifications, although often overlooked, should be instead considered when reporting low error rates, e.g., in the order of 10^{-5}. Even if the EERs reported in Table 2 are not so low as to require special care, the FNMR and FMR rates reported in Figure 2 should be carefully evaluated under this perspective. As a general recommendation, if a rough performance estimate is needed, as for grid-type parameter search, it could be reasonable to not take the rules mentioned above into account, whereas they should be considered when reporting the outcomes of the performed research.

In conclusion, the performed tests and the obtained results demonstrate the need to accurately describe comprehensive test protocols when evaluating the recognition performance on a given biometric database.

References

[An17] Anjos, A.; Günther, M.; de Freitas Pereira, T.; Korshunov, P.; Mohammadi, A.; Marcel, S.: Continuously Reproducing Toolchains in Pattern Recognition and Machine Learning Experiments. In: International Conference on Machine Learning (ICML). August 2017.

[Id21] UTFVP Fingervein Database User's Guide, `https://www.idiap.ch/software/bob/docs/bob/bob.db.utfvp/v3.0.5/guide.html` (accessed February 17, 2021).

[In06] International Organization for Standardization: Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. ISO/IEC, pp. 19795–1, 2006.

[In07] International Organization for Standardization: Information technology – Biometric performance testing and reporting – Part 2: Testing methodologies for technology and scenario evaluation. Standard, ISO/IEC, Geneva, CH, 2007.

[JKR15] Jain, A.; Klare, B.; Ross, A.: Guidelines for best practices in biometrics research. In: 2015 International Conference on Biometrics (ICB). pp. 541–545, 2015.

[Ka21] OpenVein-toolkit, Matcher.m, `https://gitlab.cosy.sbg.ac.at/ckauba/openvein-toolkit/-/blob/master/Matcher.m#L1008` (accessed February 17, 2021).

[KRU14] Kauba, C.; Reissig, J.; Uhl, A.: Pre-processing cascades and fusion in finger vein recognition. Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI), pp. 99–110, 01 2014.

[KU15] Kauba, C.; Uhl, A.: Sensor ageing impact on finger-vein recognition. In: 2015 International Conference on Biometrics (ICB). pp. 113–120, 2015.

[KU20] Kauba, C.; Uhl, A.: An Available Open-Source Vein Recognition Framework. In (Uhl, Andreas; Busch, Christoph; Marcel, Sébastien; Veldhuis, Raymond, eds): Handbook of Vascular Biometrics. Springer International Publishing, Cham, pp. 113–142, 2020.

[Ma15] Matey, J. R.; Quinn, G. W.; Grother, P.; Tabassi, E.; Watson, C.; Wayman, J. L.: Modest proposals for improving biometric recognition papers. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–7, 2015.

[MNM07] Miura, N.; Nagasaka, A.; Miyatake, T.: Extraction of Finger-Vein Patterns Using Maximum Curvature Points in Image Profiles. IEICE - Trans. Inf. Syst., E90-D(8):1185–1194, August 2007.

[MZB16] Mery, D.; Zhao, Y.; Bowyer, K.: On accuracy estimation and comparison of results in biometric research. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–8, 2016.

[TV13] Ton, B. T.; Veldhuis, R. N. J.: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: Proceedings of the 2013 International Conference on Biometrics (ICB), Madrid, Spain. pp. 1–5, 2013.

[Uh20] Uhl, A.: State of the Art in Vascular Biometrics. In (Uhl, Andreas; Busch, Christoph; Marcel, Sébastien; Veldhuis, Raymond, eds): Handbook of Vascular Biometrics. Springer International Publishing, Cham, pp. 3–61, 2020.

[Va14] Vanoni, M.; Tome, P.; El Shafey, L.; Marcel, S.: Cross-database evaluation using an open finger vein sensor. In: 2014 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings. pp. 30–35, 2014.