Assessment of Synthetically Generated Mated Samples from Single Fingerprint Samples Instances

1st Simon Kirchgasser, 2nd Christof Kauba and 3rd Andreas Uhl
Department of Computer Sciences, University of Salzburg, Salzburg, Austria
{skirch, ckauba, uhl}@cs.sbg.ac.at

***Abstract*—The availability of biometric data (here fingerprint samples) is a crucial requirement in all areas of biometrics. Due to recent changes in cross-border regulations (GDPR) sharing and accessing biometric sample data has become more difficult. An alternative way to facilitate a sufficient amount of test data is to synthetically generate biometric samples, which has its limitations. One of them is the generated data being not realistic enough and a more common one is that most free solutions are not able to generate mated samples, especially for fingerprints. In this work we propose a multi-level methodology to assess synthetically generated fingerprint data in terms of their similarity to real fingerprint samples. Furthermore, we present a generic approach to extend an existing synthetic fingerprint generator to be able to produce mated samples on the basis of single instances of non-mated ones which is then evaluated using the aforementioned multi-level methodology.**

***Index Terms*—fingerprint, synthetic sample generation, mated samples, performance evaluation**

I. Introduction

Fingerprint (FP) recognition can be considered as mature biometric solution with many different application areas in our modern world. Nevertheless, there is always a need to improve existing algorithms and to come up with novel, better performing (nowadays often CNN-based) ones, especially for particular use cases like contact-less acquisition and latent/low quality FPs. Hence, one major requirement for the development of new algorithms or the adaption of existing ones is the availability of training/test data, in this case FP samples.
The main sources of test data are either collected real FP samples from various subjects or synthetically generated FP samples. Real captured samples are preferable, as they are as close as possible to the data the system is going to process after deployment. However, acquiring a sufficient amount of FP samples, especially for training a CNN-based solution, is a time and labour intense work. Another problem that has been recently arisen are cross-border regulations on security of private data (GDPR), imposing severe restrictions on preserving, sharing and processing of person related data. Since biometric data like face, iris and FP images are considered to be a special category of private data, sharing databases of biometric samples has become very difficult, leading to an insufficient number of available datasets and thus, preventing further scientific research and industrial development of reliable biometric access control systems and forensic investigation tools.

This research was funded by the Austrian Research Promotion Agency (FFG) grant number 873462, project name "BioCapture".

If synthetically generated FP samples are used instead, the aforementioned issues can be resolved as the sample generation does not involve any data subjects. These synthetic samples do not contain any sensible private information. Hence, sharing synthetically generated datasets is not affected by cross-border privacy regulations. However, synthetically generated samples are no real samples and can only resemble the properties of real samples to a certain extent. The higher this extent, the better the "utility" of the synthetically generated samples. The goal is to produce synthetic FP samples that are indistinguishable from their real counterparts in terms of their main biometric characteristics.
An important question is how to assess the "utility" of the synthetic data, i.e. to assess the similarity of the synthetic and real samples (are the synthetic samples indistinguishable from real ones). In this work, we introduce a systematic, stepwise methodology to assess the synthetic data's utility by measuring the comparison scores' similarity behaviour instead of performing investigations regarding quality aspects or FP sample appearance variations. Hence, we suggest four levels of similarity:

- *Level 1:* The most coarse one is a similarity in terms of recognition performance only (e.g. by evaluating the equal error rate or other performance measures).
- *Level 2:* The next level is a visual evaluation based on the comparison score histograms and their corresponding score distributions.
- *Level 3:* This third analysis part is a refinement of the second level employing different histogram comparison metrics.
- *Level 4:* The last level is the most fine grained one which is the application of a statistical test on the comparison scores.

If samples pass the last one, the are highly likely to pass the other three as well and thus, to achieve a high similarity to real fingerprint samples. However, generating samples exhibiting this high level of similarity is difficult and might not be needed depending on the use case. For example if the use case is a performance evaluation only, the similarity based on a recognition performance or histogram comparison level is sufficient. On the other hand if a deep learning network for feature extraction should be trained, the maximum possible level of similarity has to be achieved, otherwise if the network is trained on synthetic samples it might fail for real ones.

There are several free of charge [1], [2] as well as commercial approaches [3] to generate synthetic FP samples. The free approaches are limited in that they are either not able to produce mated samples at all or that the produced mated samples are far from resembling the properties of real mated samples. Only the commercial solutions [3] are available to produce realistic mated as well as non-mated samples. For further details on other synthetic fingerprint generators, the interested reader is referred to [4].

This work contains two main contributions: First, an approach to generate mated samples based on a synthetic FP generator which is originally not capable of producing mated samples in combination with several image manipulation techniques originally used in image watermark benchmarking [5], [6]. A generated sample is distorted/manipulated in various ways to derive the mated samples from the output by the synthetic FP generator.

Second, a multi-level methodology to assess synthetically generated fingerprint samples, based on the above mentioned four levels of similarity, recognition performance based, comparison score distribution based, histogram comparison based and the statistical test. This methodology is then utilised to verify the effectiveness of the above mentioned mated sample generation approach, i.e. to assess the "utility" of the generated mated samples. Starting from a real dataset, one sample per finger is used and several synthetic, mated samples are generated. These synthetic samples are then compared against the remaining real ones in terms of different strategies/levels.

The paper is organized as follows: A detailed description of the proposed approach is given in Section II. Section III explains the experimental set-up to assess the similarity of the generated mated samples and real fingerprint samples. The experimental results are presented and discussed in Section IV. Finally, Section V concludes this paper and gives an outlook on future work.

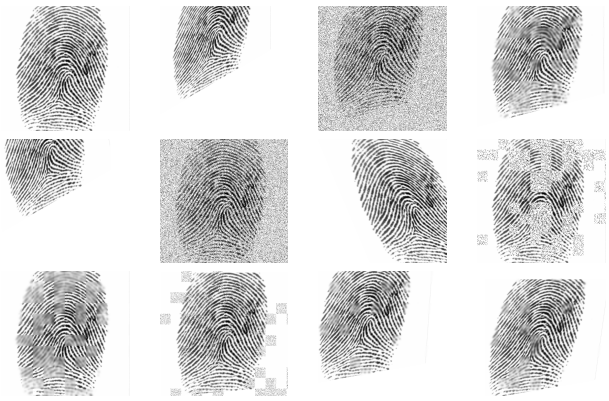

Fig. 1. Generated synthetic mated FP samples from top left beginning: original, aff, affglobal, afflocal, afftrans, globalnoise, rotation, single, smooth, mix1, mix2, mix3

II. Synthetic Mated FP sample generation

To be able to generate synthetic mated FP samples it is necessary to apply methods which are capable of manipulating the content of an imprint in a realistic manner. This methodology is provided by a model-based FP manipulation software (using similar methods as StirMark [5] and StirTrace [6]). The proposed method is a general purpose data generation one (not only for augmentation) using basic image operations to model intra-class variability. It can be extended/adapted to better reflect fingerprint distortions caused by unfavourable acquisition conditions. The following methods are available:

translation: allows to shift the FP image along x- and/or y-axis, with the shifting parameters - pixel range: [0, 150].

affine transformations: including FP rotation (range: [-15°, 15°]), x- and y-axis stretching as well as compression (parameter range for both types: [-25°, 25°]). Translation and this method simulates different positional finger placement variations relatively to the capturing device.

smoothing: The imprint is partitioned into non-overlapping blocks (block size range in pixel: [1, 32]) and an average, median or Gaussian Filter is applied to a randomly selected number of blocks. The latter one is controlled by a sigma ranging from 0 till 1.5 while the other two can be adjusted by the filter kernel size. The ratio of blocks that are smoothed can be set from 0 to 100%. Thus, it is also possible to apply the filtering to the entire imprint. Global Gaussian-based smoothing was applied only in two cases while average and median smoothing did not result in any realistic FP samples (wet fingertips or the use of hand lotions should be simulated).

noising: It is possible to apply global as well as local noising. Two different noising methods have been implemented. Gaussian noising and Speckle noising. The first method exhibits the same parameters as described in the smoothing case, while the latter one adds multiplicative noise to the FP image which can be controlled by different variances (range: [0.00, 0.25]), allowing to simulate dirt or damage on the sensor plate.

In the following different combinations of the before mentioned methods are applied to generate realistic synthetic mated FP samples. The best performing settings are described in Table I. The proposed software and the corresponding setting files can be downloaded from the webpage: www.http://wavelab.at/sources/Kirchgasser21b/. Figure 1 shows several mated examples using some of the named methods.

III. Evaluation protocol

The evaluation is done on the basis of four assessment levels which are applied in a top-down manner to exploit several levels of similarity in terms of granularity based on EER, mated score distributions, histogram comparison metrics and statistical analysis. The first level based on the EER represents the most general part of the analysis. The real datasets' EERs are compared to the ones calculated for datasets including synthetically generated mated samples. Subsequently, the knowledge gained from the first analysis step is refined by using the mated score distributions. The histograms based on the mated comparison scores are generated and score density distributions are fitted using a uni-variate kernel density estimate based fitting. This allows to illustrate similarities and differences from a visual point of view. The third level utilises three standard histogram metrics: Histogram

TABLE I
SETTINGS USED TO GENERATE REAL-SYNTHETIC/SYNTHETIC FP SAMPLES

setting name	*setting description*
aff	several affine transformations (transf.) are applied (excluding rotation)
affglobal	affine transf. either combined with 1 of 4 global smoothing (Gauss) options or 1 of 3 global noising (Speckle) options (each combination generating one mated sample)
afflocal	affine transf. combined with either 1 of 4 local smoothing (Gauss) options or 1 of 3 local noising (Speckle) options (each combination generating one mated sample)
afftrans	affine and translation transformations
globalnoise	global noising based on different levels of Speckle noise
rotation	image rotation in the range of -10 till 25°
single	one setting each which applies either affine transformation, global noising, local noising (Speckle), local noising (Gauss), rotation, local smoothing (Gauss) or translation
smooth	local smoothing (Gauss) with varying smoothing ratio and Gauss Sigma
mix1	affine transformations, translation and local Speckle based noising combinations
mix2	affine transformations, translation and local Gauss based smoothing combinations
mix3	affine transformations, translation and global Speckle based noising combinations

TABLE II
EER RESULTS (IN %) FOR DIFFERENT DATASETS USING INNOVATRICS ANSI: ORIGINAL EER (SECOND COLUMN) AND EER OBTAINED ON DATASETS CONTAINING SYNTHETIC MATED SAMPLES (THIRD TILL LAST).

dataset	**original**	*aff*	*affglobal*	*afflocal*	*afftrans*	*globalnoise*	*rot*	*smooth*	*single*	*mix1*	*mix2*	*mix3*
FVC2002 Db1a	0.336	0.683	0.269	0.000	0.327	0.413	0.369	0.071	0.273	0.000	0.066	0.000
FVC2004 DB1A	1.969	5.278	2.899	1.151	8.044	1.657	0.655	0.562	1.479	1.900	4.506	0.172
CASIA T2	1.712	1.876	8.898	1.071	1.624	11.179	2.061	1.0231	6.122	1.463	0.963	2.742
CASIA uru4500-1	1.943	3.272	2.143	1.773	3.408	1.911	2.065	1.527	2.933	1.106	1.705	1.143
PLUS IBColumbo	0.097	45.908	46.013	45.824	46.128	43.498	46.426	44.448	46.083	47.182	45.504	45.398
PLUS NB3010	3.615	43.656	44.537	43.361	45.023	44.532	42.789	43.865	43.363	43.033	43.750	43.329
PLUS V311	0.836	45.326	47.228	45.996	45.137	44.932	44.957	45.869	46.683	47.107	45.385	46.893
FVC2002 Db4a	0.683	4.132	5.580	2.008	2.194	6.415	1.562	3.074	3.871	0.9235	1.669	0.627
FVC2004 DB4A	1.140	2.619	3.873	1.043	1.893	4.303	1.718	1.529	2.734	0.492	0.477	0.657
CNN	-	0.850	0.000	0.000	0.336	0.000	1.205	0.000	0.002	0.000	0.000	0.000
Anguli Basic	-	0.851	0.000	0.000	0.336	0.000	1.205	0.000	0.002	0.000	0.000	0.000
Anguli Moisture	-	0.957	2.873	0.000	0.480	2.239	1.599	0.036	0.769	0.035	0.009	0.000
Anguli Frag	-	1.029	0.000	0.000	0.301	0.000	1.266	0.103	0.000	0.000	0.000	0.000
Anguli Intens	-	0.851	0.000	0.000	0.336	0.000	1.205	0.000	0.037	0.000	0.000	0.000
Anguli Noise	-	0.765	0.000	0.000	0.361	0.000	1.305	0.000	0.003	0.000	0.000	0.000

Intersection (HI), which measures similarity in a range of [0, 1], where 0 corresponds to no similarity and 1 to a perfect one. Chi-Squared distance (Chi) and Kullback-Leibler divergence (KL) [7] measure divergence in a range of [0, unbounded], where 0 corresponds to perfect similarity and the higher the values the less similarity.

The metric values are calculated for three different scenarios: real/real, where all pairs of the real FP sample datasets are compared to each other in order to establish a baseline to compare the following two cases to. The second case is the real/real-synthetic one, where real-synthetic refers to mated samples generated on the basis of real samples (FVC2002 Db1a, FVC2004 DB1A, CASIA T2, CASIA uru4500-1, PLUS IBColumbo, PLUS NB3010 and PLUS V311), whereas in the third case, real/synthetic, mated samples have been generated on the basis of synthetically generated samples (FVC2002 Db4a, FVC2004 DB4A, CNN, Anguli Basic, Anguli Moisture, Anguli Frag, Anguli Intens, Anguli Noise). If the metric values for the second and third case report less similarity than for case one, the synthetic samples can be clearly distinguished from the real ones. The real datasets have been captured utilising different capturing devices (different characteristics limiting their metric based similarity), while all the involved samples in the second and third case originate from the same capturing device, exhibiting similar characteristics. The results are presented as averaged values over the included datasets in Table VI as well as for the second case per dataset in Table V.

Finally, the most fine-grained part (assessment level 4) of the top-down evaluation is a statistical analysis based on the Mann-Whitney-U test [8] which is applied to prove if the score distributions of the synthetically generated mated samples exhibit a similar statistical behaviour as the real mated samples do. This test is a non-parametric test for two independent sample sets and allows a t-test identical interpretation of the results. However, the Mann-Whitney-U test is computed based on rank sums rather than means as it is done using a t-test.

Several real and synthetic FP datasets are evaluated: The real FP datasets were collected by the use of several capturing devices including thermal, capacitive and (multispectral) optical ones, while the synthetic samples were generated by traditional [3] or deep-learning methods [2].

FVC 2002/2004 Db1/DB1 are subsets of the databases established for the second/third Fingerprint Verification Contest [9] and have been collected by optical capturing devices. All samples exhibit a resolution of at least 500dpi and each subset contains 800 imprints from 100 fingers.

The *CASIA Fingerprint Subject Ageing Version 1.0* [10] was collected using one capacitive and two optical scanners which results in a total of six subsets (2009, 2013). Five imprints

of both index and middle fingers have been acquired from 49 capturing subjects. Thus, each subset contains 980 images with a resolution of 500 dpi. In the current study only the subset acquired by the capacitive sensor (T2) and one acquired by the optical capturing device (uru4500-1) are considered.
The third real FP database is the *PLUS Multi-Sensor and Longitudinal Fingerprint Dataset (PLUS MSL)*, containing 108106 FP samples collected by 10 different capturing devices. In this work only the samples acquired by the thermal device (NB3010), one capacitive device (IBColumbo) and one multispectral one (V311) are utilised.

The utilised synthetic datasets include FVC 2002 Db4a and FVC 2004 DB4A [3] and generated synthetic samples using the ANGULI generator [1] as well as a CNN method [2], each one containing samples of 100 different fingers, exhibiting the same resolution of 500 dpi. The ANGULI ones have been generated by varying parameters controlling e.g. core type, moisture type, ridge fragmentation, ridge intensity, and ridge noise. To assess the recognition performance of the FP samples two commercial state-of-the-art minutiae-based approaches have been applied: ANSI SDK developed by Innovatrics (https://www.innovatrics.com) and VeriFinger SDK 11.0 developed by Neurotechnology (https://www.neurotechnology.com).

IV. Experimental Results and Discussion

Initially, the different settings for generating the synthetic mated samples were optimised based on the FVC2002 Db1a using Innovatrics ANSI until the samples passed all four assessment levels (cf. first row of Table III). Subsequently, these settings were applied to generate synthetic mated samples for all remaining experiments.
The first assessment level, depicted in Table II shows the EER result for Innovatrics ANSI. The results for VeriFinger highly differ from these in most cases. They are in general similar to the worst EER results obtained on the PLUS MSL dataset. Thus, the synthetic mated sample generation was not successful for VeriFinger and therefore no detailed results are included. This indicates that the synthetic generation process can have a high impact on the general performance and this impact does not need to be the same across different recognition systems. However, in most cases the EER obtained on datasets using real FP samples as original input is mostly comparable to the EER observed on the original datasets without synthetic mated samples. The PLUS MSL datasets are an exception as the EER in these cases is inferior to all others. Furthermore, some of the settings seem to generate highly similar mated FP samples leading to a perfect EER of 0% in several cases, which is not a desirable result. For the synthetic datasets generated by [1] and [2] a direct comparison between original EER and the EER values of the newly generated subsets/settings is not possible as these methods are not able to produce mated FP samples. Thus, these datasets are not included in the subsequent evaluation.

TABLE V
Histogram metrics averaged over all generated subsets/settings using Innovatrics ANSI and VeriFinger (including standard deviation in case of HI).

	dataset	*CHI*	*HI*	*KL*
Innovatrics ANSI	FVC2002 Db1a	299909	0.6742 ± 0.0891	1372057
	FVC2004 DB1A	173330	0.7816 ± 0.0531	893702
	CASIA T2	152240	0.8627 ± 0.1042	767728
	CASIA uru4500-1	339924	0.6494 ± 0.1237	1484145
	PLUS IBColumbo	579940	0.6615 ± 0.2136	2383041
	PLUS NB3010	258919	0.6545 ± 0.2059	1025624
	PLUS V311	578892	0.6484 ± 0.2144	1644704
	FVC2002 Db4a	199635	0.6408 ± 0.1545	857437
	FVC2004 DB4A	172528	0.7823 ± 0.0539	890059
VeriFinger	FVC2002 Db1a	115812	0.7812 ± 0.0919	661224
	FVC2004 DB1A	145732	0.7467 ± 0.1323	806668
	CASIA T2	180800	0.6594 ± 0.3044	1139679
	CASIA uru4500-1	204364	0.7509 ± 0.0892	1106211
	FVC2002 Db4a	117647	0.7222 ± 0.0905	612534
	FVC2004 DB4A	115897	0.7284 ± 0.0788	606479

TABLE VI
Average values (and standard deviation in case of HI) for histogram metrics applied on distributions of real/real, real/real-synthetic and real/synthetic mated samples.

		CHI	HI	KL
Innovatrics ANSI	real/real	298060	0.6447 ± 0.1589	1267500
	real/real-synth.	340450	0.7046 ± 0.1434	1367285
	real/synth.	327880	0.8178 ± 0.1680	919090
VeriFinger	real/real	230940	0.6619 ± 0.1485	1017700
	real/real-synth.	161677	0.7345 ± 0.1544	928446
	real/synth.	311970	0.8259 ± 0.1670	872020

All in all, the EER is too vague to reliably judge if synthetic samples show a similar behaviour as real FP samples. Not only the recognition system is influencing the evaluation but also the input data and the selected settings. Hence, a more detailed analysis is beneficial.
Continuing with the second assessment level, the comparison score histograms depicted in Figure 2 underpin that on the FVC2002 Db1a, which was used for optimisation, a high similarity between the original and the newly generated synthetic samples is present. For all other combinations of selected settings and datasets the similarity is acceptable, as shown in the second example using the synthetic FVC2002 Db4a. On the other hand, as depicted in the third and fourth histogram, the differences are higher compared to the first two.
To further refine the top-down analysis, especially for the cases where the EER and the score distribution based evaluation resulted in contrary findings, the next step is to apply assessment level three, the histogram metrics to the score distributions. Their results are listed in Tables V and VI. The results for the PLUS MSL using VeriFinger are omitted as they follow the same trend as using Innovatrics ANSI. The results given in Table V reveal two distinct cases: most of the metric results confirm the preceding evaluation but some are contrary. In case of PLUS NB3010, CHI and KL indicate a higher similarity compared to the other two PLUS MSL subsets, while HI indicates the same extent of similarity. Contrasting to this the EER values of all three PLUS MSL datasets indicating a low similarity of the synthetic vs the real samples. Hence, the evaluation using the histogram metrics is still ambiguous.

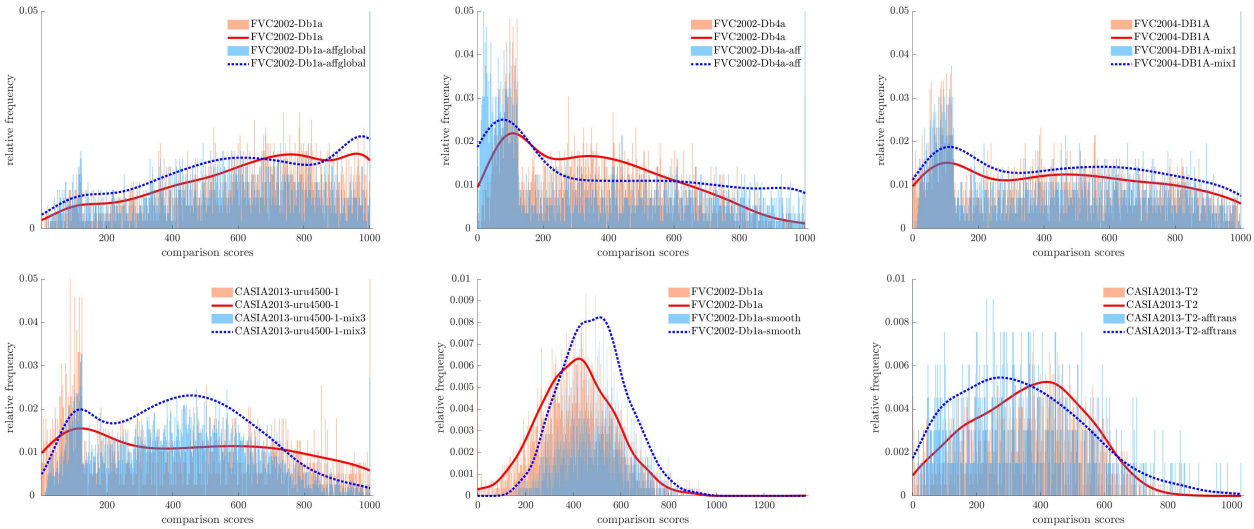

Fig. 2. Mated score distributions plots (Innovatrics ANSI first four examples and VeriFinger last two ones), showing for several datasets and settings the differences between original mated sample scores and synthetic ones.

TABLE III

P-VALUES FOR DIFFERENT DATASETS USING INNOVATRICS ANSI. FOR THE APPLIED MANN-WHITNEY-U TEST AN SIGNIFICANCE NIVEAU OF 0.01 WAS CHOSEN, HENCE ALL P-VALUES EQUAL OR GREATER THAN 0.01 INDICATE THAT THE COMPARISON SCORES OF THE ORIGINAL MATED SAMPLES AND THE SYNTHETICALLY GENERATED ONES HAVE BEEN SELECTED FROM SAME DISTRIBUTIONS.

dataset	*aff*	*affglobal*	*afflocal*	*afftrans*	*globalnoise*	*rot*	*smooth*	*single*	*mix1*	*mix2*	*mix3*
FVC2002 Db1a	0.55	0.79	0.18	0.07	0.01	0.56	0.02	0.46	0.30	0.01	0.06
FVC2004 DB1A	0.36	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.00	0.00
CASIA T2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CASIA uru4500-1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
PLUS IBColumbo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PLUS NB3010	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PLUS V311	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FVC2002 Db4a	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.02
FVC2004 DB4A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE IV

P-VALUES FOR DIFFERENT DATASETS USING VERIFINGER (CF. TABLE III).

dataset	*aff*	*affglobal*	*afflocal*	*afftrans*	*globalnoise*	*rot*	*smooth*	*single*	*mix1*	*mix2*	*mix3*
FVC2002 Db1a	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FVC2004 DB1A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CASIA T2	0.00	0.15	0.00	0.88	0.00	0.00	0.65	0.00	0.00	0.00	0.00
CASIA uru4500-1	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FVC2002 Db4a	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.63	0.00	0.00
FVC2004 DB4A	0.00	0.00	0.00	0.00	0.00	0.78	0.00	0.65	0.00	0.00	0.00

Table VI enables a more high level view by averaging over all datasets for the three cases real/real, real/real-synthetic and real/synthetic. There is obviously a high dissimilarity in the real/real case, indicating that the involved real FP samples are highly distinct. The metrics indicate that for real/synthetic compared to real/real-synthetic the generated samples are more similar, suggesting that the generated mated samples appear to be more realistic if synthetic samples are used as source. In both cases, HI and KL confirm that the synthetically generated mated FP samples are more similar to the real ones than the real ones originating from different datasets (different capturing devices), while CHI shows contrary results.

Based on the first three evaluation levels is not possible to obtain decisive results. Hence, the final part of the analysis is the fourth assessment level, which focusses on the application of the Mann-Whitney-U test (using a significance value of 0.01) and describes the similarities and differences of the computed comparison scores from a statistical point of view. The corresponding results are presented in Table III and Table IV, utilising the Innovatrics ANSI and the VeriFinger recognition system, respectively. With the EER, the visual density distribution and the histogram metrics based analysis being limited in their significance especially in several cases (PLUS MSL datasets), the statistical test application allows a reliable statement. Except for the FVC2002 Db1a and a few other single datasets containing synthetically generated mated samples a clear difference between the synthetic and real mated sample comparison scores is observable. As shown in Table III and Table IV most p-values are 0, which, according to the statistical test, indicates that the synthetic generated

mated samples can be distinguished from the real ones and are thus, not realistic enough.

The results confirmed that the proposed approach is able to generate mated FP samples of sufficient quality to pass levels 1-3 of the assessment methodology for several cases if the single parameters of the involved image manipulation methods are optimised with respect to 1) a particular dataset or 2) a particular FP recognition system. On the other hand, the generated mated samples failed level 4, the statistical test, indicating that they do not exhibit the necessary utility on a fine grained level. Currently, the approach needs to be optimised for each of the involved datasets and recognition systems, i.e. the settings do not generalise to arbitrary datasets and recognition systems. However, if the generated FP samples only have to satisfy an EER based performance requirement (similar recognition performance to real FP samples) or exhibit a similar mated score distribution, the proposed methodology is sufficient (cf. Table II, Figure 2).

V. Conclusion and Future Work

The first contribution of this work is a methodology to assess the level of utility (in terms of realistic appearance of the synthetic samples) of synthetic FP samples in comparison to their real counterparts, based on a top-down approach. This evaluation is based on the comparison of synthetically generated mated samples (with real samples as a basis) against real mated samples using the EER, score distribution plots, histogram comparison metrics based on the comparison scores as well as a statistical test. The comparison scores were calculated using state of the art minutiae based fingerprint recognition schemes.

The second contribution is a technique to generate mated samples from (synthetically) generated single instances of non-mated fingerprint samples. It is based on simple image manipulation techniques (rotation, shearing, adding noise, warping) which can be combined and parametrised in order to generate different mated samples. This whole procedure can be used in addition to any existing synthetic fingerprint generator (most free available ones are not able to generate mated samples) in order to obtain mated fingerprint samples. The proposed approach is available free of charge and can be downloaded from our website: http://www.wavelab.at/sources/Kirchgasser21b.

The top-down evaluation results confirmed that the proposed mated sample generation technique is able to satisfy the requirements of realistic mated samples if only the recognition performance or the shape of the comparison score distributions are considered. Furthermore, for some combinations of matchers and datasets the generated mated samples achieve similar properties to the real ones, i.e. they cannot be distinguished from the real ones for the histogram based analysis. On the other hand, the settings for one dataset and matcher combination cannot be directly used for the other datasets and matchers, i.e. the approach currently lacks on generalisability. In the future work, more datasets as well as fingerprint recognition schemes will be included in order to arrive a set of "general best parameters" and subsequently, increase the generalisability of the proposed approach. Moreover, further image manipulation techniques will be included and tested to enhance the properties of the synthetically generated samples and improve their realistic appearance.

References

[1] J. Haritsa, A. Ansari, K. Wadhwani, and S. Jadhav. Anguli: synthetic fingerprint generator. [Online]. Available: http://dsl.cds.iisc.ac.in/projects/Anguli

[2] M. Sadegh Riazi, S. M. Chavoshian, and F. Koushanfar. (2020) Synfi: Automatic synthetic fingerprint generation. [Online]. Available: https://github.com/MohammadChavosh/synthetic-fingerprint-generation

[3] R. Cappelli, A. Erol, D. Maio, and D. Maltoni, "Synthetic fingerprint-image generation," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 3. IEEE, 2000, pp. 471–474.

[4] A. Makrushin, C. Kauba, S. Kirchgasser, S. Seidlitz, C. Kraetzer, A. Uhl, and J. Dittmann, "General requirements on synthetic fingerprint images for biometric authentication and forensic investigations," in *Proceedings of the 9th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'21)*, Brussels, Belgium (held Online due to Covid), 2021, pp. 1–11, accepted.

[5] J. Hämmerle-Uhl, M. Pober, and A. Uhl, "Towards standardised fingerprint matching robustness assessment: The stirmark toolkit – cross-database comparisons with minutiae-based matching," in *Proceedings of the 1st ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'13)*, Montpellier, France, Jun. 2013, pp. 111–116.

[6] R. Merkel, M. Hildebrandt, and J. Dittmann, "Application of stirtrace benchmarking for the evaluation of latent fingerprint age estimation robustness," in *3rd International Workshop on Biometrics and Forensics (IWBF 2015)*. IEEE, 2015, pp. 1–6.

[7] B. T. González, "Analytical comparison of histogram distance measures," in *Proc. of the 23rd Iberoamerican Congress Progress in Pattern Recognition (CIARP'18)*, vol. LNCS 11401. Springer, 2018.

[8] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. CRC press, 2020.

[9] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition (2nd Edition)*, 2009.

[10] Biometrics Ideal Test (BIT. (2018) CASIA Fingerprint Subject Ageing Version 1.0. [Online]. Available: http://biometrics.idealtest.org/dbDetailForUser.do?id=15