Protocol Based Similarity Evaluation of Publicly Available Synthetic and Real Fingerprint Datasets

Dominik Söllinger, Simon Kirchgasser, Andreas Uhl
Multimedia Signal Processing and Security Lab
University of Salzburg
5020 Salzburg, Austria
{dsoellinger, skirch, uhl}@cs.sbg.ac.at

Andrey Makushina, Jana Dittmann
Advanced Multimedia and Security Lab
Otto von Guericke University Magdeburg
39106 Magdeburg, Germany
andrey.makrushin@ovgu.de
jana.dittmann@iti.cs.uni-magdeburg.de

Abstract

Several attempts have been made recently to generate synthetic fingerprint data. This has become necessary after legal changes in Europe and some US states in order to allow and continue long-term developments in the field of fingerprint biometrics. Apart from utilizing traditional methods (often based on Gabor filters), deep convolutional neural networks are widely used to generate synthetic fingerprint samples. The current study aims at comparing several publicly available synthetic fingerprint datasets with several datasets that consist of imprints taken from real people. To enable a comparison, first a detailed description of these datasets is carried out. Secondly, an available 4-level protocol is used, which is supposed to show similarities and/or differences between real and synthetic fingerprint samples in terms of quality assessment and non-mated as well as mated comparison scores' behavior. Creators of synthetic datasets should feel encouraged to report the resemblance of their synthetic samples to real FPs by using the proposed protocol.

1. Introduction

In recent years, many efforts have been made by various governments to increase the privacy protection of individuals. The European Union has issued the General Data Protection Regulation (GDPR), the California introduced the California Consumer Privacy Act (CCPA) and Illinois updated the Personal Information Protection Act (PIPA) (all three now exhibit comparable regulations). These regulations were absolutely necessary in order to protect people's data privacy, confidentiality, and integrity, given that the technical possibilities for unauthorized use of personal data are constantly improving, especially due to the rapid advances made with deep neural networks.

Biometric data, whether it is fingerprint (FP) data, facial data, or iris images, is used in many ways in everyday life. However, a lot of data (exhibiting a large variety of natural variations) is needed in order to improve or redesign biometric applications. Unfortunately, the acquisition of biometric data from real people is very costly, not only in terms of the time required to collect the data, but also in terms of the availability of human resources since participants need to be willing to agree to GPDR or comparable regularities. A potential solution to overcome the high efforts and costs could be the usage of synthetically generated biometric data. Furthermore, the replacement of real biometric samples with synthetic ones, would be a step towards open research by sharing biometric data which could otherwise not be made public [23].

In the best case, synthetically generated biometric samples are capable of representing all known natural variation, e.g. all types of basic FP patterns (loop, arch and whorl) or minutiae point types (ridge ending, bifurcation, island or crossover), which are potentially contained in biometric samples of real people. The chances that these sample variations are included in synthetically created FPs can either be controlled by quasi-deterministic strategies (traditional generation approaches) or by learning the likelihood that the named variations are detected in the training data (more recent deep learning based methods). Moreover, ensuring equal distribution of sample attributes (e.g. race, gender, age) can be done during a synthetic FP generation process as well [19].

The most influencing factors, which need to be detected within the synthetic data, are (a) less discriminative differences among generated non-mated subjects and (b) identity leakage, in case too much similarity between training samples (which are likely to be from real subjects in the first place) and generated ones is given. In the first case, the synthetic data from different subjects would be too similar

and thus, behave differently from biometric data from real subjects if evaluated with an automated authentication system. In the second case, privacy/identity leakage would be an obvious threat since personal information from the training data will be contained in the synthetic data. This must only be to such a small extent that no conclusions can be drawn about the persons (identities) used during training. In both described cases it is hardly possible to detect such issues by a visual inspection, especially as the number of generated samples is likely to be large. As a consequence, an automated quality and performance based evaluation is mandatory.
Currently only one study [17] is known, which tries to define general requirements that need to be met in order to create synthetic FP samples that behave like real FPs and further, only [12] defines a protocol to evaluate the resemblance of synthetic mated FPs to real FP samples. However, when creating synthetic biometric samples, typically the first step is to generate only non-mated samples, describing different subjects. Thus, the suggested protocol [12] is limited to mated samples and therefore cannot be used to evaluate the resemblance to pristine of non-mated FP samples which are created in many recent works, e.g. [2, 18].

Contribution of Work: The contribution of this work is to address some neglected aspects regarding the evaluation of synthetic datasets. First of all, the protocol proposed in [12], which was originally designed to assess the resemblance of mated FP samples to real FPs, is extended to assess non-mated samples. Secondly, the work studies which of several publicly available FP datasets resemble real FP data using the proposed evaluation protocol. At this point, it is the only study available that compares public synthetic datasets with various real ones in terms of biometric quality and recognition performance. Furthermore, it is proposed that the extended protocol should be used in the future for the evaluation of synthetic FP databases in order to achieve a general basis for comparison. The only aspect that is excluded in the course of this work, but must not be neglected, relates to the aspect of identity leakage already mentioned above. The obtained results show that in order to fully assess the authenticity of a synthetic FP dataset, someone needs to have access to the synthetic as well as the corresponding real data. However, real data can often not be made public due to company specific or legal restrictions. Creators of synthetic datasets should therefore feel encouraged to use the proposed protocol to report the resemblance of the generated synthetic data to the real data, especially when the real data utilized cannot be provided to the public.

The remainder of the current work is structured as follows: Various reference literature is discussed in the following Section 2, excluding studies that will discussed separately in Section 3. In Section 4 the generation of additional mated samples and the experimental methodology are presented, while the description and discussion of the evaluation results is done in Section 5. Section 7 concludes this study including an outlook on planned future work.

2. Related Work

Regardless of which biometric modality is considered, according to [19] methods for generating synthetic biometric samples can be categorized into three classes: (a) adaptation, (b) synthesis and (c) reconstruction from a biometric template. If an adaptation based approach is selected, the main goal should be to modify an existing (real or synthetic) biometric sample to mimic specific acquisition conditions. In the context of the current study, adaptation techniques are utilized to create mated from non-mated samples if the investigated synthetic FP creation method is not capable of producing mated sample (see Section 4.1 for details). The second class, synthesis, contains methods that aim to create synthetic samples with predefined conditions from scratch. A prominent method that falls into this category is SFinGe [6]. The third class, reconstruction, describes various methods, that make use of existing biometric samples and create synthetic ones from them by learning a model. Furthermore, there is the possibility to combine synthesis and reconstruction as done in [4].
Apart from the classification just discussed, the various methods for generating synthetic data can also be categorized in another way: (i) methods that make use of traditional modeling methods and (ii) methods that are based on a data-driven approach. Model-based methods have been applied in rather early works on FP synthesis, while in recent years data-driven approaches have gained importance. Hence, at first a summary of model-based studies is given, followed by a discussion of investigations applying data-driven concepts.
In [26] a model, describing the FP ridge pattern orientation by utilizing core and delta information, was presented. This concepts was applied in an extended manner in [6] allowing to model so called master FPs using an iterative application of Gabor filters and a method to simulate realistic distortions. A drawback of [6] is that the generated distortions are, from a statistically point of view, not always representative compared to real FPs. This was partly solved in [11] as possible realistic distortions are derived from real FP samples. As skin diseases are known to influence FP authentication processes it would be beneficial to be able to include such complicating factors to synthetic FP data as well [8].
Data-driven synthesis approaches (which includes the synthesis of FP data) exhibit a specific advantage over model-based application as no knowledge about image semantics

is necessary. Nevertheless, this advantage can only be exploited if enough discriminative (real or synthetic) training data is available.
A Wasserstein generative adversarial network (GAN) was applied to generate master imprints being capable of matching several real FPs in [3]. A different approach, the so called FingerGAN [21], is able to simulate FP samples that are originally from the FVC2006 [5] and PolyU [14] dataset. A more recent method, the SynFi approach [24], allows the generation of high-resolution FPs generation by combining a GAN and a super-resolution network. A combination of a convolutional autoencoder and an improved Wasserstein GAN was used in [22] to generate synthetic rolled and plain FPs. In [25], progressive growing GAN, StyleGAN and StyleGAN2 were applied to create realistic partial FP patterns, while in [27] a CycleGAN was used for texture transfer from real FPs to Anguli [1] generated FPs before a super-resolution network increased the image size of the generated samples.

3. Fingerprint Datasets and Generators

This section introduces the FP datasets that are investigated in this work. Each dataset provides either real or synthetic FP samples, except the FVC dataset series, which provides both types of samples. A brief summary of the utilized datasets (generators) can be found in Table 1. Note that all datasets used in this work are publicly available.

AMSL Synthetic FP Datasets: The AMSL Synthetic FP dataset collection[1] provides three synthetic FP datasets. The first dataset, named *AMSL SynFP SGR v1*, provides 50k non-mated FP samples. These synthetic samples were created using a StyleGAN2-ADA model. The second and third dataset, named *AMSL SynFP P2P v1* and *AMSL Synth P2P v2*, respectively, provide 40k mated FP samples each. These samples were created in a two step process. In the first step, minutiae were extracted from randomly generated Anguli samples. Each minutiae pattern then underwent a series of transformations, which included random rotations, shifts and cuts, to create mated minutiae patterns. In the second step, natural-looking FPs were generated by passing each minutiae pattern through a Pix2Pix model pretrained to reconstruct a FP from its minutiae. Note that a combined set of three datasets was utilized to train each model. This set is composed of the Neurotechnology Cross Match, FVC2002 DB1 A+B and FVC2004 DB1 A+B dataset.

Anguli Generator: Anguli [1] is a freely available synthetic fingerprint generator which builds upon the algorithms utilized by SFinGe [7]. Samples generated by Anguli mimick a plain FP with a resolution of 500 DPI. Anguli can also generate mated samples, however, the quality of generated mated impression leaves a lot to be desired. Therefore, no mated samples created by Anguli are examined in this work. Instead, mated samples are created using the technique outlined in Sec. 4.1.

Clarkson FP Dataset: The Clarkson [2] FP Dataset is a publicly available dataset composed of 50k non-mated synthetic FP samples which were created using StyleGAN. The model was trained on a set of 72k FPs captured with a CrossMatch Guardian Sensor.

PLUS Synthetic FP Datasets: The PLUS Synthetic FP datasets collection[2] currently provides one synthetic dataset named *PLUS SynFP RealScan v1*. The dataset is composed of 50k non-mated samples generated using StyleGAN2-ADA. It was trained on all samples from the RealScanG1 sensor provided in the PLUS-MSL-FP dataset.

PLUS-MSL-FP Dataset: The PLUS-MSL-FP dataset [13] is a publicly available multi-sensor FP dataset composed of FP samples from all ten fingers of 63 different subjects (5 impressions per finger). Each fingers was captured with ten different acquisition devices, and recaptured four times over a time-span of two years. The datasets consists of $\sim$128k images.

PrintsGAN Dataset: The PrintsGAN [9] dataset is a mated synthetic FP datasets composed of 525k rolled FP impressions from 35k unique identities. To create a series of mated FP samples, first, a binary masterprint is created using BigGAN. Next, mated binary samples are created by warping the masterprint using a Thin Plate Spline (TPS) and segmenting out random parts of the finger. This step again utilizes a GAN to predict the parameters of the TPS as well as the segmentation mask. Last, binary FPs are convert into realistic FPs using a generative autoencoder.

FVC Datasets: The FVC datasets were created to verify the FP recognition performance of the algorithms submitted to the Fingerprint Verification Content (FVC) 2002 [15], 2004 [16] and 2006 [5], respectively. Each datasets is composed of four subsets, three real subsets (DB1-3) and one synthetic subset (DB4). To allow competitors to tune their algorithms, each subset was split into two sets (referred to "a" and "b"). Further details about the number of images in each subset can be found in Table 1. Note that only the competition sets (DB1a-DB4a) are studied in this work.

[1] https://gitti.cs.uni-magdeburg.de/Andrey/gensynth

[2] https://wavelab.at/sources/Kirchgasser23b/

	Dataset	Subset	Ref.	Impression Type	Incl. mated samples	DPI	Subjects × Impressions	Total
Real	FVC2002	DB1a	[15]	Plain, Optical	✓	500	100 × 8	800
	FVC2002	DB2a	[15]	Plain, Optical	✓	569	100 × 8	800
	FVC2002	DB3a	[15]	Plain, Capacitive	✓	500	100 × 8	800
	FVC2004	DB1a	[16]	Plain, Optical	✓	500	110 × 8	800
	FVC2004	DB2a	[16]	Plain, Optical	✓	500	110 × 8	800
	FVC2004	DB3a	[16]	Swipe, Thermal	✓	512	110 × 8	800
	FVC2006	DB1a	[5]	Electric Field	✓	250	140 × 12	1680
	FVC2006	DB2a	[5]	Optical	✓	569	140 × 12	1680
	FVC2006	DB3a	[5]	Thermal sweeping	✓	500	140 × 12	1680
	PLUS-MSL-FP	Columbo	[13]	Plain, Capacitive	✓	500	630 × 20	≈12.6k
	PLUS-MSL-FP	NB-3010-U	[13]	Thermal	✓	500	630 × 20	≈12.6k
	PLUS-MSL-FP	RealScanG1	[13]	Plain, Optical	✓	500	630 × 20	≈12.6k
Synthetic	AMSL SynFP	P2P v1	[18]	Plain, Optical	✓	500	4000 × 10	40k
	AMSL SynFP	P2P v2	[18]	Plain, Optical	✓	500	4000 × 10	40k
	AMSL SynFP	SGR v1	[18]	Plain, Optical		500	-	50k
	Anguli Generator	-	[1]	Plain	✓	500	-	-
	Clarkson	-	[2]	Plain, Optical		500	-	50k
	PrintsGAN	-	[9]	Rolled	✓	500	35k × 15	525k
	PLUS SynFP	RealScanG1	-	Plain, Optical		500	-	50k
	FVC2002	DB4a	[6]	Plain	✓	500	100 × 8	800
	FVC2004	DB4a	[7]	Plain	✓	500	100 × 8	800
	FVC2006	DB4a	[5]	Plain	✓	500	140 × 12	1680

Table 1. Summary of all datasets and generators utilized in this work. The table shows the main characteristic of each dataset such as the impression of each subset, total number of impressions or subject, DPI, etc. Note that the term "subject" refers to a finger (not a person). A detailed description of each dataset can be found in Sec. 3.

4. Data Preprocessing and Evaluation

The statistical evaluation conducted in this work requires equally sized sets of mated FP samples. As the number of mated samples and impressions is different in each dataset, and some datasets do not provide mated samples, each datasets had to undergo some preprocessing. These preprocessing steps are described in the following two subsections, while the third subsection presents the utilized evaluation methodology.

4.1. Generation of mated samples

Four synthetic FP datasets, *i.e.*, *PLUS SynFP RealScanG1*, *Clarkson*, *AMSL SynFP SGR v1* and *Anguli*, do originally not provide mated FP samples. Since various experiments conducted in this work rely on the presence of mated samples, a protocol to create these mated ones had to be developed first. This protocol is based on methods similar those provided by StirMark [10] and StirTrace [20]. The selected transformation methods manipulate the FPs' content in a realistic manner by compressing/stretching the imprints according to the x- and y-axis, rotating and translating them. The parameters controlling the transformations were sampled from a differently parameterized normal distribution. For each generated mated sample, the transformation method and the parameter values were selected on a random basis each time. Hence, similarity between mated samples is ensured, but they exhibit a rather different shape after performing the manipulations, resulting in a fairly complicated dataset. As a consequence, it can be assume that the recognition performance of those datasets containing mated samples generated by this process will be worse compared to the other synthetic datasets. The code and further details for the mated sample generation are publicly available[3].

4.2. Preparation of equally sized folds

For a correct statistical comparison of the genuine and imposter distributions of each dataset, all datasets are required to have the same number of genuine and imposter pairs. This is not the case since each dataset provides a different number of subjects as well as mated impressions. To overcome this problem, multiple equally sized data folds were randomly sampled from each dataset as follows: 100 subjects were randomly chosen from each dataset. For each of these subjects, eight impressions were randomly selected to create a fold. As a result, each fold is composed of 800 FP samples from 100 different subjects and abbreviated using 'F' for fold and a number from 1 to 5 indicating the exact

[3] https://wavelab.at/sources/Kirchgasser23b/

fold, e.g. *RealScanG1-F1*. In case of the Anguli dataset the five folds are represented by five subsets that are independent from each other.

4.3. Evaluation Methodology

The synthetic datasets summarized in Table 1 were evaluated based on the following aspects: (a) sample quality, (b) non-mated sample score distribution behavior, (c) mated sample score distribution behavior and (d) recognition performance. The real datasets mentioned in Table 1 act as reference datasets for the distribution of real data.

The FP sample quality is measured using NFIQ2[4] as quality assessment metric. Mated and non-mated score are using the VeriFinger SDK 12.1 [5] as biometric recognition system.

The evaluation of the synthetic datasets follows the 4-level evaluation protocol detailed in [12]. The 4-levels are as follows:

1. **Level:** Comparison of the recognition performance (e.g., equal error rate) of each dataset. Question: Does a synthetic dataset behave similarly to a real dataset? If yes, to which dataset is it similar?

2. **Level:** Visual comparison of quality, mated and non-mated score distributions. Question: Does a synthetic dataset have a distribution similar to a real dataset? If yes, to which one is it similar?

3. **Level:** Metric-based score distribution similarity assessment. The metric-based evaluation aims to compare the variability among the score distributions, where distances between distribution pairs are used as a measure for variability. The distance between two score distributions is measured either using Chi-Squared distance (CHI), Histogram Intersection (HI) and Jensen-Shannon divergence (JS). Intuitively, if a synthetic dataset behaves similarly to a real dataset, its distance(s) to real datasets should be similar to distances among real data.

   Question: Is the variability between a synthetic dataset and real data clearly dissimilar to the variability among real data?

4. **Level** Direct statistical comparison of each dataset's quality, mated and non-mated score distribution using a Mann-Whitney-U test. Question: Which synthetic dataset is dissimilar to each individual real dataset from a statistical point of view? For which dataset can dissimilarity not be excluded?

[4] https://github.com/usnistgov/NFIQ2

[5] https://www.neurotechnology.com/verifinger.html

Level-3 evaluation protocol: Given a synthetic dataset D_{synth}, calculate the distance d_synth between the synthetic dataset and the first real dataset. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Level-4 evaluation protocol: To account for variations in the score distribution among different folds of a real dataset (a MWU-Test indicated dissimilar fold distributions), it is impossible for a fold of a synthetic dataset to be similar to all folds of a real dataset. To take this into account, the dissimilarity between a synthetic dataset and a real dataset is assessed as follows: Each fold of a synthetic dataset is compared to each fold of the real dataset with a MWU-Test. The count is then calculated for how many synthetic folds have at least one real fold that is not dissimilar.

5. Results

This result section is structured into several paragraphs. The first paragraph discusses the results of the recognition performance evaluation which corresponds to the first level of the evaluation methodology explained in Sec. 4.3. The second, third, and fourth paragraphs discuss the results of the quality, non-mated and mated score evaluations, respectively. The evaluation of quality, non-mated and mated scores is done using the second, third and fourth level of the evaluation methodology.

Recognition Performance: Tables 2 and 3 summarize the average biometric recognition performance (described by the average equal error rate (**avEER**) and the corresponding mean comparison scores for mated ($\overline{\mathbf{mated}}$) and non-mated ($\overline{\mathbf{nmated}}$) FP samples. Furthermore, the standard deviations of the comparison scores are given as well ($\mathbf{s_{mated}}$ and $\mathbf{s_{nmated}}$). To simplify the comparison of real datasets with synthetic ones, the results are shown in two distinct tables. While Table 2 shows the results for real datasets, Table 3 presents the results for synthetic datasets.

As can be seen in Table 2, for real FP datasets, the **avEER** is lower than 1% (except FVC06-DB1A) which indicates an overall good recognition performance. This observation is consistent with the relatively high $\overline{\mathbf{mated}}$ scores and a low $\mathbf{s_{mated}}$. Note that the FVC06-DB1A is well known to be a fairly challenging dataset with characteristics (images resolution of 96×96 pixels and 250DPI) that are entirely different from all other datasets.

Dataset	avEER	$\overline{\mathbf{mated}}$	$\mathbf{s_{mated}}$	$\overline{\mathbf{nmated}}$	$\mathbf{s_{nmated}}$
FVC02-DB1A	0.254	281.70	7.78	76.01	31.43
FVC02-DB2A	0.251	287.70	6.27	77.57	33.86
FVC02-DB3A	0.446	246.53	10.19	80.78	33.73
FVC04-DB1A	0.479	218.04	7.75	78.90	33.23
FVC04-DB2A	0.452	226.55	9.49	73.98	37.22
FVC04-DB3A	0.253	212.77	7.79	67.14	39.33
FVC06-DB1A	3.522	142.91	5.81	58.34	23.19
FVC06-DB2A	0.251	342.54	7.49	72.54	35.39
FVC06-DB3A	0.323	263.04	8.87	95.67	37.88
Columbo	0.423	286.95	8.09	89.55	32.57
NB-3010-U	0.915	221.76	11.98	86.35	31.67
RealScanG1	0.251	335.05	9.47	93.34	34.66

Table 2. The averaged EER (in percent) over all five folds, the average mated comparison scores ($\overline{\mathbf{mated}}$), the standard deviation of the mated comparison scores ($\mathbf{s_{mated}}$) and the average non-mated comparison scores ($\overline{\mathbf{mmated}}$) as well as the standard deviation of the non-mated comparison scores ($\mathbf{s_{nmated}}$) for the real FP datasets are presented.

Dataset	avEER	$\overline{\mathbf{mated}}$	$\mathbf{s_{mated}}$	$\overline{\mathbf{nmated}}$	$\mathbf{s_{nmated}}$
AMSL P2P-v1	6.973	69.85	21.46	23.03	43.16
AMSL P2P-v2	0.251	192.42	24.05	44.23	41.89
AMSL SGR-v1	3.802	247.72	11.24	145.63	38.98
Anguli	5.070	227.18	21.21	117.83	33.29
Clarkson	2.512	254.47	9.93	153.52	37.21
Printsgan	0.270	196.06	6.59	57.33	38.76
SynFP RealScanG1	2.610	240.32	9.64	148.25	34.78
FVC02-DB4A	0.252	249.93	19.83	58.47	32.60
FVC04-DB4A	0.409	221.57	13.51	63.88	32.18
FVC06-DB4A	0.269	219.33	15.53	55.82	30.74

Table 3. The recognition performance evaluation results for the synthetic FP datasets.

Comparing Table 3 with Table 2, it can be seen that the recognition performance of the synthetically generated FP datasets is worse compared to the real ones. In particular, it is noticeable that the average EER is the worst (highest) for those synthetic datasets whose mated FP samples were generated using the method described in Section 4.1. This observation is expected due to two reasons. First, the generation process results in mated FP samples that exhibit more distortions caused by rotation and other affine transformation based variations, than the other synthetically generated mated samples. Secondly, the applied VeriFinger SDK 12.1 seems to be particularly sensible to these introduced distortions. Moreover, it is noticeable that for the dataset with the highest EER (AMSL P2P-v1) the $\overline{\mathbf{mated}}$ is much lower than for the others and the $\mathbf{s_{nmated}}$ is the highest among all evaluated real and synthetic datasets.
Based on the comparison of Table 2 with Table 3, it is further possible to state that the EER of a synthetic FP dataset should be below 1% if this dataset behaves similarly as a real one. This is currently only the case for AMSL P2P-v2, Printsgan and the SFinge datasets.

Quality assessment: Figure 1 presents the Level-3 evaluation results for NFIQ2 scores, comparing the variability between a synthetic dataset and real data to the variability among real data. The numbers inside each cell show the number of folds that are found real by the Level-3 evaluation protocol. Hence, a higher number in the cell indicates a higher chance that samples contained in a synthetic dataset are similar to real data.

As can be seen in Figure 1, most synthetic datasets are considered real by the employed evaluation protocol, independent of the utilized distribution comparison matrix. Only the AMSL-SGR-v1, AMSL-P2P-v1, and AMSL-P2P-v2 are clearly dissimilar from real data.

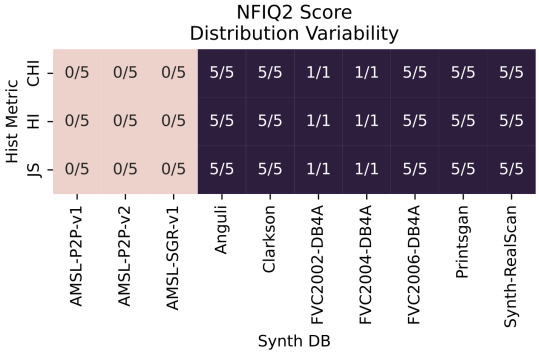

Figure 1. Result of the Level-3 evaluation performed on NFIQ2 quality scores. CHI denotes the Chi-Squared distance, HI the Histogram Intersection and JS the Jenson-Shannon divergence.

Figure 2 shows the results of the Level-4 evaluation which directly compares individual NFIQ2 score distributions of synthetic and real datasets. The numbers inside each cell show how many folds of a synthetic dataset exhibit at least one fold with a similar distribution among the real datasets. Thus, the presented results show a potential similarity between the FVC2006-DB4 and NB-3010-U dataset (5/5 folds), PrintsGan and FVC2004-DB1A (5/5 folds), PrintsGan and IBColumbo (5/5 folds), Anguli and NB-3010-U dataset (5/5 folds) as well as AMSL-P2P-v1 and IBColumbo (5/5 folds). Furthermore, in 4 out of 5 folds, Synth-RealScanG1 has at least one non-dissimilar fold in the RealScan dataset. This was expected since Synth-RealScanG1 was directly created from RealScan samples.

Non-mated comparison score evaluation: Figure 3 presents the Level-3 evaluation results for non-mated scores, comparing the variability between a synthetic dataset and real data to the variability among real data. As can be seen, only the non-mated score distributions of AMSL-SGR-v1, Clarkson and partially Printsgan are similar to the distributions of real data.

Figure 4 shows the results for the Level-4 evaluation, where a fold of each synthetic dataset is directly compared to all folds of a real dataset. Based on the figure, it can be

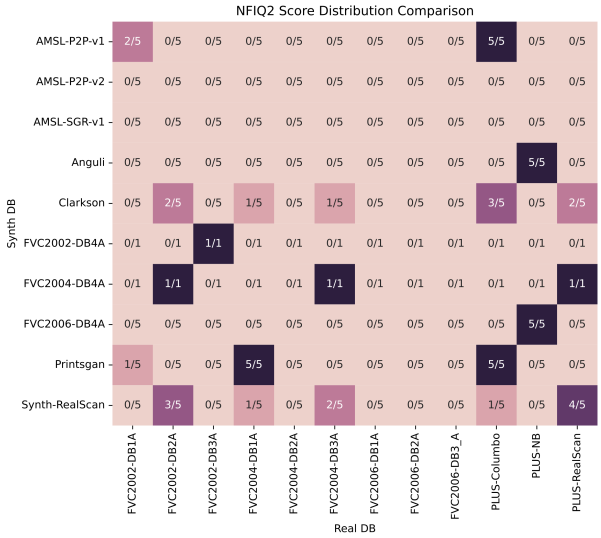

Figure 2. Results of the Level-4 evaluation on NFIQ2 quality scores. Each cell in the table represents the number of folds in the synthetic dataset where the MWH test accepted at least one fold from the real dataset. The significance value for the MWU-Test is 0.01.

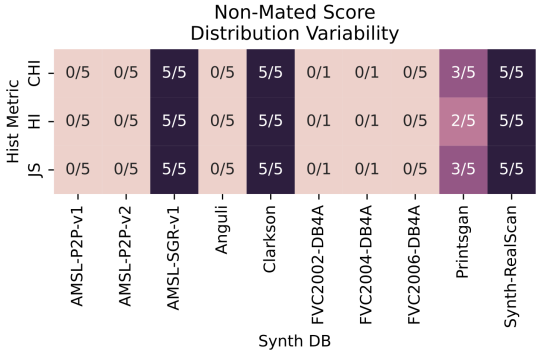

Figure 3. Result of the Level-3 evaluation performed on non-mated scores. CHI denotes the Chi-Squared distance, HI the Histogram Intersection and JS the Jensen-Shannon divergence.

claimed that there is more or less no similarity between the non-mated score distributions of synthetic and real data.

Mated comparison score evaluation: Figure 5 shows the results of the Level-3 evaluation for mated scores. As can be seen, most synthetic datasets are similar to real datasets. The only synthetic datasets which is dissimilar according to the Level-3 evaluation protocol is AMSL-P2P-v1.

Figure 6 shows the results of the Level-4 comparison for mated comparison scores. As can be seen in the figure, direct comparisons of a synthetic fold with a real fold indicate mostly dissimilar datasets. There are only 3 synthetic

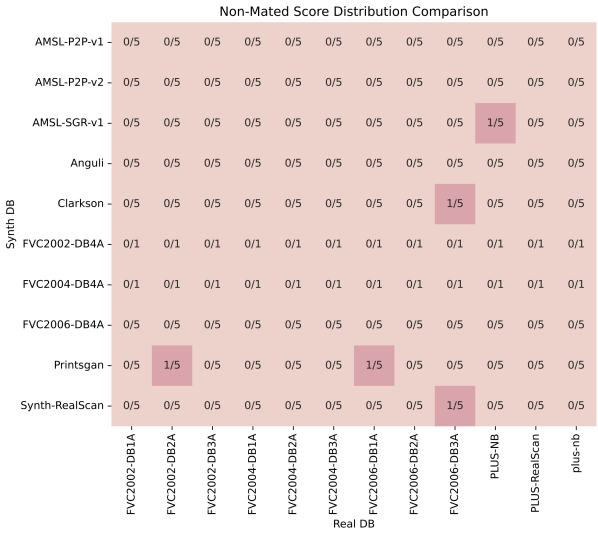

Figure 4. Results of the Level-4 evaluation on non-mated scores. Each cell in the table represents the number of folds in the synthetic dataset where the MWH test accepted at least one fold from the real dataset. The significance value for the MWU-Test is 0.01.

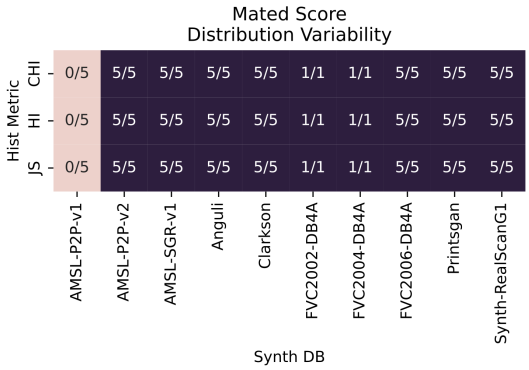

Figure 5. Result of the Level-3 evaluation performed on mated scores. CHI denotes the Chi-Squared distance, HI the Histogram Intersection and JS the Jensen-Shannon divergence.

datasets where all 5 folds have at least one real fold that is not clearly dissimilar according to the MWU-Test. These 3 synthetic datasets are Synth-RealScanG1, FBC2006-DB4A and AMSL-SGR-v1.

6. Discussion

Sec. 5 presented the results of the Level-3 and Level-4 evaluations for quality, non-mated, and mated comparison scores. This section summarizes the findings and discusses their practical implications.

Firstly, it's important to note that when asking how well a synthetic dataset resembles *real data*, the actual question

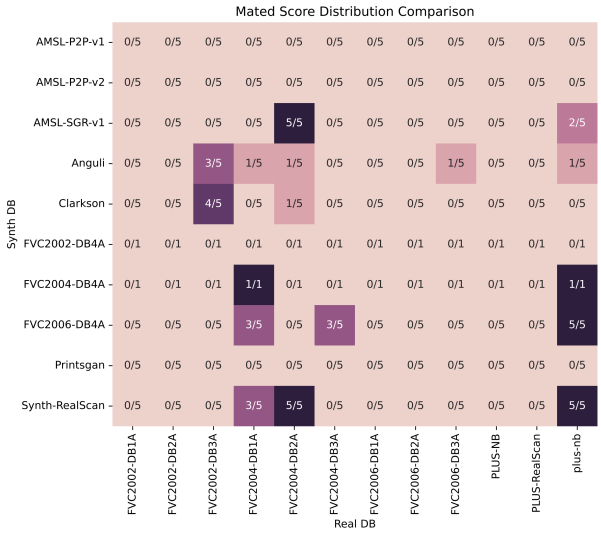

Figure 6. Results of the Level-4 evaluation on mated scores. Each cell in the table represents the number of folds in the synthetic dataset where the MWH test accepted at least one fold from the real dataset. The chosen significance value for the MWU-Test is 0.01.

being asked is how well the synthetic dataset resembles the *set of real databases*. Thus, the statement that a synthetic dataset behaves differently from real data only holds if the chosen set of real databases is representative of all real datasets.
Secondly, it's important to understand that a dissimilarity found in the Level-4 evaluation does not imply that the synthetic dataset cannot be real. In fact, the only scenario where a synthetic dataset should match with a real dataset is when the synthetic dataset was directly created from the real dataset. In this evaluation, this is only the case for Synth-RealScanG1 and RealScanG1. Consequently, Level-3 should be preferred over Level-4 evaluations when evaluating how well a synthetic dataset resembles real data.

Assuming that the chosen set of real datasets is representative of real datasets, only AMSL-P2P-v2, AMSL-SGR-v1, FVC2002-DB4A, and FVC2004-DB4A are obviously not dissimilar in terms of the NFIQ2 score. Considering the mated distribution, only AMSL-P2P-v2 and FVC2004-DB4A might be similar to real data. The analysis of the non-mated distribution does not seem to provide meaningful insights regarding the question of how well a synthetic dataset resembles real data.

7. Conclusion

The current study aims to evaluate how a synthetic dataset resembles real fingerprint datasets. For this purpose, ten publicly available synthetic fingerprint datasets are evaluated using a modified 4-Level protocol, focusing on quality, recognition performance, and mated and non-mated comparison scores.

Among the ten synthetic fingerprint datasets examined, only four were found not to be dissimilar with any of the twelve real fingerprint datasets, based on their NFIQ2 score. These datasets are AMSL-P2P-v2, AMSL-SGR-v1, FVC2002-DB4A, and FVC2004-DB4A. Considering the mated scores, only the AMSL-P2P-v2 and FVC2004-DB4A datasets were found not to be dissimilar to the examined real datasets.

8. Acknowledgments

The research in this paper is a part of the joint project GENSYNTH (Tools for the Generation of Synthetic Biometric Sample Data), which is funded by the Austrian Science Fund (FWF) under project no. I4272.

References

[1] A. H. Ansari. Generation and storage of large synthetic fingerprint database. *ME Thesis, Jul*, 2011.

[2] K. Bahmani, R. Plesh, P. Johnson, S. Schuckers, and T. Swyka. High fidelity fingerprint generation: Quality, uniqueness, and privacy. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3018–3022. IEEE, 2021.

[3] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross. Deepmasterprints: Generating masterprints for dictionary attacks via latent variable evolution. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2018.

[4] R. Bouzaglo and Y. Keller. Synthesis and reconstruction of fingerprints using generative adversarial networks. *arXiv preprint arXiv:2201.06164*, 2022.

[5] R. Cappelli, M. Ferrara, A. Franco, and D. Maltoni. Fingerprint verification competition 2006. *Biometric Technology Today*, 15(7-8):7–9, 2007.

[6] R. Cappelli, D. Maio, and D. Maltoni. Synthetic fingerprint-database generation. In *2002 International Conference on Pattern Recognition*, volume 3, pages 744–747. IEEE, 2002.

[7] R. Cappelli, D. Maio, D. Maltoni, et al. Sfinge (synthetic fingerprint generator). 2004.

[8] M. Drahanskỳ and O. Kanich. Influence of skin diseases on fingerprints. *Biometrics under Biomedical Considerations*, pages 1–39, 2019.

[9] J. J. Engelsma, S. A. Grosz, and A. K. Jain. Printsgan: Synthetic fingerprint generator. *arXiv preprint arXiv:2201.03674*, 2022.

[10] 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 first ACM workshop on Information hiding and multimedia security*, pages 111–116, 2013.

[11] P. Johnson, F. Hua, and S. Schuckers. Texture modeling for synthetic fingerprint generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 154–159, 2013.

[12] S. Kirchgasser, C. Kauba, and A. Uhl. Assessment of synthetically generated mated samples from single fingerprint samples instances. In *Proceedings of the IEEE Workshop on Information Forensics and Security (WIFS2021)*, pages 1–6, Montpellier, France, 2021.

[13] S. Kirchgasser, C. Kauba, and A. Uhl. The plus multi-sensor and longitudinal fingerprint dataset: An initial quality and performance evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):43–56, 2022.

[14] C. Lin and A. Kumar. Matching contactless and contact-based conventional fingerprint images for biometrics identification. *IEEE Transactions on Image Processing*, 27(4):2008–2021, 2018.

[15] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2002: Second fingerprint verification competition. In *2002 International Conference on Pattern Recognition*, volume 3, pages 811–814. IEEE, 2002.

[16] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2004: Third fingerprint verification competition. In *Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004. Proceedings*, pages 1–7. Springer, 2004.

[17] 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 (IHMMSec'21)*, pages 1–11, Brussels, Belgium (held Online due to Covid), 2021.

[18] A. Makrushin, V. S. Mannam, and J. Dittmann. Data-driven fingerprint reconstruction from minutiae based on real and synthetic training data. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP*, pages 229–237. SCITEPRESS, 2023.

[19] A. Makrushin, A. Uhl, and J. Dittmann. A survey on synthetic biometrics: Fingerprint, face, iris and vascular patterns. *IEEE ACCESS*, 11:33887–33899, 2023.

[20] 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)*, pages 1–6. IEEE, 2015.

[21] S. Minaee and A. Abdolrashidi. Finger-gan: Generating realistic fingerprint images using connectivity imposed gan. *arXiv preprint arXiv:1812.10482*, 2018.

[22] V. Mistry, J. J. Engelsma, and A. K. Jain. Fingerprint synthesis: Search with 100 million prints. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[23] D. S. Quintana. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *Elife*, 9:e53275, 2020.

[24] M. S. Riazi, S. M. Chavoshian, and F. Koushanfar. Synfi: Automatic synthetic fingerprint generation. Cryptology ePrint Archive, Paper 2020/217, 2020. `https://eprint.iacr.org/2020/217`.

[25] S. Seidlitz, K. Jürgens, A. Makrushin, C. Kraetzer, and J. Dittmann. Generation of privacy-friendly datasets of latent fingerprint images using generative adversarial networks. In *VISIGRAPP (4: VISAPP)*, pages 345–352, 2021.

[26] B. G. Sherlock and D. M. Monro. A model for interpreting fingerprint topology. *Pattern recognition*, 26(7):1047–1055, 1993.

[27] A. B. V. Wyzykowski, M. P. Segundo, and R. de Paula Lemes. Multiresolution synthetic fingerprint generation. *IET Biometrics*, 11(4):314–332, 2022.