Fingerprint Template Ageing Revisited - It's the Quality, Stupid !

Simon Kirchgasser and Andreas Uhl
Department of Computer Sciences
University of Salzburg, AUSTRIA
{skirch, uhl}@cs.sbg.ac.at

Katy Castillo-Rosado, David Estévez-Bresó,
Emilio Rodríguez-Hernández and José Hernández-Palancar
Advanced Technologies Application Center
(CENATAV), La Habana, CUBA
{krosado, destevez, erodriguez, jpalancar}@cenatav.co.cu

Abstract

Several studies conducted on time-span separated fingerprint datasets revealed recognition performance degradation as compared to commonly achieved accuracies. The reported accuracy reduction is mainly caused by a shift of the genuine score distributions towards the impostor ones. This fact raised discussions about the reasons behind eventual template ageing effects in fingerprint recognition. Analysing the only publicly available time-separated fingerprint dataset (CASIA Fingerprint Subject Ageing Version 1.0), it is demonstrated in the current study that (i) advances in recognition technology mitigate the earlier observed decrease in recognition accuracy and (ii) advances in fingerprint quality assessment now lead to the conclusion that it is indeed fingerprint quality which causes the earlier observed recognition impact. Current results on this dataset do not suggest the existence of a fingerprint template ageing effect.

1. Introduction

Over the last years, several studies have been conducted on time-span separated fingerprint (FP) datasets. Most of these investigations revealed a recognition performance degradation introduced by a reduction of the genuine comparison scores. Obviously, there might exist several factors causing the observed recognition accuracy reduction, and it is definitely not clear that FP template ageing causes these effects. However, only few studies so far tried to investigate the reasons behind the exhibited recognition performance degradations in time-span separated FP databases.

Proper quality assessment is a mandatory precondition to ensure a stable and comparable application of FP recognition systems. Thus, it is natural to look at differences in FP quality eventually causing the observed phenomena. If the imprints' quality is indeed responsible for reduced genuine comparison scores, then it is incorrect to qualify the observed effects as "template ageing" as these are eventually identified as not being time-dependent. Instead, such effects would be properly termed as "template changes" introduced by variations during data acquisition. In fact, controversial results have been obtained, whether the observed reduced genuine comparison scores can be attributed to FP quality differences: The authors of [21] concluded that indeed quality differences can be made better responsible for the reduced genuine comparison scores as compared to time-separation data (using non-public FP forensic data), while the authors of [7, 8] do not identify quality as the reason for the observed effects, and thus do not rule out FP template ageing as a potential reason (analysing the public available "CASIA Fingerprint Subject Ageing Version 1.0" database).

The authors of [7] considered so called ghost fingerprints as another potential reason for the described genuine score decrease (again on the "CASIA Fingerprint Subject Ageing Version 1.0"). This investigation was motivated by the high amount of ghost fingerprints present in the respective data. These problematic structures in FP images are introduced by a non sufficient sensor plate cleaning during FP acquisition and are typically not detected by a FP based quality assessment. In case the acquisition protocol is not constant over time, one might observe a different amount of ghost fingerprints in time-separated data eventually causing the decreased genuine comparison score. Of course, such an effect is not at all time-related as this effect could also occur between acquisition sessions, which are not separated

978-1-5386-7180-1/18/$31.00

by time intervals at all. Thus, recognition degradation in time separated data caused by a different number of ghost fingerprints could not be qualified as template ageing. However, the authors showed that decreased genuine comparison scores still prevail even after ghost fingerprints have been removed, thus, ghost fingerprints do not contribute to the observed effects.

Since [7, 8] have been published, there have been new developments regarding (i) FP quality assessment (NIST FP Image Quality 2.0 (NFIQ 2.0) as discussed in Section 5 has been introduced) and (ii) algorithmic accuracy in commercial FP recognition systems (VeriFinger and Innovarics SDK as described in Section 3). Also, we intend to apply more sophisticated (statistical) analysis concerning eventual quality differences. Extending previous investigations on the CASIA Fingerprint Subject Ageing Version 1.0 data we aim to answer the following research questions with this work: (a) Do we observe similar recognition performance degradations using more recent FP recognition systems? (b) Does the application of NFIQ 2.0 on the imprints describe the data's quality better or more thoroughly compared to NFIQ 1.0 regarding quality differences in the datasets ? (c) Does the application of proper statistical tests clarify the question if quality differences among the data can be reported and made be responsible for the detected recognition accuracy degradations? (d) Does the consideration of quality difference of imprints causing false matches lead to clearer results with respect to eventual quality differences as compared to the corresponding imprints' average quality (as used in [8])?

The rest of this paper is organised as follows: In Section 2 a thorough description of related work is given. The used FP databases and applied recognition SDKs will be described in Section 3. A subsequent detailed discussion of recognition experiments is done in Section 4. The considered FP quality assessment methodologies, applied statistical methods, performed experiments and corresponding results are analysed in Section 5, before concluding this study in Section 6.

2. FP Recognition on Time-separated Data

Almost all investigations on time-separated FP data report a recognition performance decrease, caused by a reduction of the genuine match scores which can be described by a shift of the genuine score distributions towards the impostor ones.

In [20], a time-interval of 16 weeks has been sufficient to reveal a slight degradation in recognition accuracy conducting experiments on 3D finger range data. Further, [14] reported an EER increase using the Korea Fingerprint Recognition Interoperability Alliance (KFRIA) database, which was acquired using three different commercial sensors (2 optical and 1 capacitive sensor type) and exhibits a time span of 1 year. Another study, using hand-print data collected with a common flat-bed scanner, covering a 5 year time-span, reports recognition degradation effects caused by roughly 33% decreased genuine match scores [19]. Comparable results on the same database were obtained by [18]. Furthermore, [21] confirms this genuine score decrease on a forensic database (covering time-spans of up to 7 years) as well. In [8], the same effects have been confirmed once more performing the experiments on the CASIA Fingerprint Subject Ageing Version 1.0 data which will be considered in this study too. The authors concluded that template ageing could be an eventual explanation for the observed recognition accuracy reduction.

Further investigations focused either on longer time intervals [1], non-minutiae based comparison schemes [9], or on the description of observed reduced comparison scores by analysing user-group specific effects [6]. In [1], a forensic dataset provided by the German federal criminal police office (BKA), exhibiting time intervals of 10 to 30 years, was considered. As result it was concluded that the recognition accuracy is lower when the time interval is increased. The "Doddington Zoo" concept [2] was used in [6] to reveal the presence of similar recognition accuracy degradation effects as reported before (again, the respective CASIA data was analysed). Additional investigations on the same data [9] described very similar effects with respect to decreased comparison accuracy in case of applying non-minutiae based recognition systems. However, only few studies done on time-separated FP data so far performed experiments to highlight the reasons which could cause the recognition accuracy's degradation. The sole description of increased error rates is not sufficient enough to be compliant with the definition of template ageing effects, because for template ageing, it must be proven that these effects are introduced by time-related changes. The underlying definition is given in the ISO/IEC biometric testing standard ISO/IEC 19795-1, which states that "Longer time intervals generally make it more difficult to match samples to templates due to the phenomenon known as template ageing" [11]. Further, it is defined that "template-ageing" might be introduced by an "increase in error rates caused by time-related changes in the biometric pattern, its presentation, and the sensor". Nevertheless, various non-time related reasons are known, which might cause the observed increased error rates. In [21] the described genuine comparison score reduction was explained by quality based influences. This conclusion was drawn after the application of a covariate-fit analysis model which revealed that image quality explains the observed increased errors better as time-related changes. In [7, 8], it was possible to detect: (i) a genuine score distribution shift towards the impostor scores distribution, (ii) almost no influence of the imprints' quality upon the detected effect and (iii) ghost FPs are not the reason for the observed

effects. This led the authors of [7, 8] to conclude that neither quality nor ghost fingerprints can be made responsible for the observed recognition accuracy degradations and thus, template ageing can not be ruled out as eventually causing these effects on the considered database.

Apart from investigations focusing on time-span related recognition performance, there have been several studies focusing on the impact of age on recognition performance relying on age groups datasets. The aspect of considering different human age groups is very interesting and important for the analysis of time-separated FP recognition performance: If it is possible to report a reduced performance for more elderly age groups compared to younger ones, the potential chances of detecting time-related effects in databases containing time-separated imprints, at least for long time-spans, would be higher of course. In fact, in [15] the authors could show that elderly age groups exhibit a worse behaviour in terms of FP quality and comparison accuracy applying a statistical one-way analysis of variance (ANOVA) and the calculation of Pearson correlation coefficients. The same database, acquired by an optical and a capacitive sensor, was reused in [12] and [13], where the authors focused on a minutiae count, a biometric quality and performance figure based analysis. The age group containing the imprints of the oldest volunteers did receive the worst results among all considered groups. In [17] a similar study was performed, but the core aspect considered very young people as well. For most performed experiments it was reported that kids' FP performance suffers compared to the adults' comparison accuracy [17]. Finally, in [4] the main goal was to cope with different age groups and observed quality degradation effects by performing isotropic rescaling methods on the children's data.

3. Datasets and Recognition Methodology

A dataset intended towards analysing FP ageing effects can still be influenced by various non-ageing related factors. As a consequence, potential reasons for such non time-related template differences, e.g. illumination variations or other differences in acquisition protocol, should be avoided. Furthermore, it is important that the same sensor(s) is/are used at all data collection events, because otherwise cross sensor effects might be introduced (which again might easily be misinterpreted as template ageing effect).

The authors decided to use the only publicly available time-separated FP data, which was already considered in other related studies. This decision leads to the possibility of comparing to or even extending results presented in [7, 8, 9]. The corresponding data was acquired at the Center for Biometrics and Security Research (CBSR) at the Chinese Academy of Sciences, Institute of Automation (CASIA). The entire database "CASIA Fingerprint Subject Ageing Version 1.0" is publicly available[1]. On the one hand it contains a subset of the also publicly available CASIA-FPV5 database[2], and on the other hand a dataset which was collected 4 years later in 2013, respectively. The first set "CASIA2009" contains 980 FP images of 49 volunteers, whose both fore- and second fingers have been acquired (5 imprints per finger). During the capturing process an U.are.U 4000 scanner, produced by DigitalPersona, was used. This optical device has a resolution of 512 dots per inch (dpi), which results in 8-bit/pixel grayscale images with a resolution of 328x356 pixel. The same set-up was considered in 2013 once again. This resulted also in 980 FP images using the same volunteer's fingers as in 2009 and the same number of imprints per finger. The main difference in the "CASIA2013" database is that it contains 5 independent subsets (980 images each), which have been collected using 3 different sensor types. Two instances of an U.are.U 4000 scanner were used as well as two instances of an U.are.U 4500 scanner (datasets uru4000_1, uru4000_2, uru4500_1, uru4500_2). In terms of image resolution and bit depth the U.are.U 4500 device has the same specifications as reported for the U.are.U 4000. The fifth subset was acquired by a capacitive TCRU1C sensor characterised by a resolution of 508 dots per inch (dpi) and an image resolution of 256x360 pixel.

To simplify the results' description in the following sections, some abbreviations need to be introduced. These are illustrated in Table 1. The first abbreviations (A, B_i) were chosen to describe the so-called "single" 2009 and 2013 datasets (containing only the data acquired in these years) independently from each other. The recognition accuracy on each single database will represent a baseline for the detailed analysis concerning the systems' performance and eventual ageing effects (see Section 4). All databases, which have been abbreviated using C_i or D_i, refer to "crossed" datasets where two different sets from 2009 or 2013 have been combined into a new one. C_i datasets contain data leading to potential ageing related results because imprint from 2009 and one set of 2013 were combined. To eliminate possible cross-sensor effects we constructed only one dataset (D_{23}) which contains the sole images acquired with an U.are.U 4000 scanner and are stored in B_2 and B_3. This D_{23} dataset was built to showcase results that can be used during the "ageing" datasets' (C_i) comparison. In case the performance figures are equal in C_i and D_{23} and exhibit the same degradation as reported in previous studies considering time separated FP data, it is clear that template ageing can be excluded as potential reason for this observation as time-differences were excluded by the construction of D_{23}.

FP Recognition Systems: The state-of-art algorithms in

[1] http://biometrics.idealtest.org/dbDetailForUser.do?id=15

[2] http://biometrics.idealtest.org/dbDetailForUser.do?id=7

Abbrev.	Orig. Name	Abbrev.	Orig. Name
single		*crossed*	
A	'CAISA2009'	C_1	A and B_1
B_1	TCRU1C	C_2	A and B_2
B_2	uru4000$_1$	C_3	A and B_3
B_3	uru4000$_2$	C_4	A and B_4
B_4	uru4500$_1$	C_5	A and B_5
B_5	uru4500$_2$	D_{23}	B_2 and B_3

Table 1. Abbreviations used for the processed datasets.

FP recognition obtain the comparison score information by an application of minutiae based feature extraction systems. In this study we applied 4 different recognition systems:

NIST Biometric Image Software (NBIS): Implemented by the National Institute of Standards and Technology (NIST)[3]; in this work Release 5.0.0 was used.
VeriFinger (NEURO): The *VeriFinger SDK*[4], developed by Neurotechnology, is minutiae based as well. The current Release 10.0 was applied. It turns out that this release contains algorithmic enhancements improving the overall performance of the system compared to results reported in [8] where Release 7.1 was applied on the imprints.
Innovatrics (ANSI, ISO and IDKiT): The fingerprint recognition SDKs *ANSI, ISO* and *IDKiT*[5] were developed by the Slovakian company Innovatrics and have been selected to use an additional common off-the-shelf (COTS) recognition system.
CFIM: This algorithm is based on the Delaunay triangulation using a given minutiae set [5]. The comparison algorithm performs the comparison between two models which describe the spatial and directional minutiae relationships by using a triangle-based representation.

The comparison scores for each dataset were calculated using the protocol as used in all FP Verification Contests (FVC), e.g. [10]. This methodology excludes all symmetric matches, thus no correlation among the received scores is possible.

4. FP Recognition on CASIA Fingerprint Subject Ageing Version 1.0

In the following we discuss the results obtained by the FP recognition systems applied to our dataset considering three popular performance figures: Equal error rate (EER %), the lowest false non match rate (FNMR) for a false match rate (FMR) less or equal to 0.1% (F_{100}) and the zero false match rate (zFMR). In case performance figures on C_i and D_{23} are worse compared to both figures obtained on A and B_i they are highlighted in yellow. If a C_i value is worse,

[3]http://www.nist.gov/itl/iad/ig/nbis.cfm
[4]http://www.neurotechnology.com/verifinger.html
[5]https://www.innovatrics.com/tools-sdks/

data	NBIS			VeriFinger 7.1		
	EER	F_{100}	*zFMR*	**EER**	F_{100}	*zFMR*
A	7.42	0.13	0.34	2.07	0.04	0.08
B_1	8.95	0.15	0.39	2.07	0.04	0.08
B_2	8.17	0.13	0.35	1.96	0.04	0.06
B_3	9.07	0.18	0.81	4.00	0.08	0.81
B_4	5.96	0.10	0.91	2.04	0.04	0.73
B_5	7.30	0.14	0.97	3.69	0.07	0.98
C_1	12.63	0.26	0.57	5.32	0.10	0.22
C_2	14.76	0.29	0.58	5.97	0.12	0.25
C_3	14.37	0.29	0.87	6.16	0.12	0.90
C_4	13.18	0.25	0.97	5.81	0.11	0.90
C_5	13.46	0.25	0.99	6.73	0.13	0.99

Table 2. Recognition performance of NBIS and Verifinger 7.1 comparison according to [8].

time-related template changes (i.e. template ageing) or non time-related changes might be the reason for it. However, if we observe performance figure degradation in set D_{23} time-related template changes can be ruled out as a possible reason. In case $C_{2,3}$ datasets' performance values are worse than those of D_{23}, this is an indication that eventually, time-related changes might cause the observed effects. Table 2 reproduces results of [8] (also using the older VeriFinger version) while Table 3 displays results of the recognition schemes additionally considered in this paper. It is clearly visible that ANSI, ISO, IDKit, and the more recent VeriFinger Release 10 perform much better than results reported in [8, 9] and those of CFIM. On the one hand we can be observe that the absolute recognition performance of these recognition systems is better on datasets A and B_i. On the other hand, and more importantly for the scope of this study, it is interesting to detect that formerly reported performance degradations on time-separated data (i.e. datasets C_i) are not present in most of the considered cases. For all Innovatrics' implementations a performance decrease can only be observed in terms of zFMR, while EER as well as F_{100} are typically in-between figures obtained on A of B_i. For A or B_i data's zFMR the number of corresponding false non matches must be higher compared to the number of false non matches detected at the EER threshold. This is obviously identical for C_i datasets. However, the sole detection of a zFMR decrease in C_i can be explained by the introduction of a few additional cross-time related false non matches. Thus, the observation of a performance decrease in terms of zFMR occurs not unexpected and seems reasonable.
For VeriFinger 10 we still observe an EER increase on C_1 and C_5. Hence, we can answer research question (a) clearly: (More) recent commercial COTS SDKs do exhibit previously seen recognition performance degradations on time-separated data to a much lesser extent if they do at all.The performance on dataset D_{23} is typically seen between the performance of dataset C_2 and C_3, respectively, thus, no

additional time-related changes can be assumed for the C_i data based on these results. Furthermore, when eventual cross-sensor comparison effects are considered, the recognition results of C_2 and C_3 (involving only U.are.U 4000 scanners) are expected to be superior compared to $C_{1,4,5}$ which are based on cross-sensor comparison. Results do not at all support this assumption. We also observe, that for some recognition techniques, A, B_3 and B_5 performance figures are inferior to those of other datasets. It will be interesting to compare this observation to the quality determined on these datasets in the next section.

5. Quality Analysis of suspicious Imprints

In [7] and [8], no specific statistical analysis was conducted to verify the obtained observations. The results indicated no quality difference in the datasets separated in time (the conclusion was drawn by considering average quality values and box plots only). Further, the methodology applied to analyse the quality of imprints involved in false matches seems questionable. In [8], the mean quality of the two imprints involved in an incorrect match (false positive or false negative) was determined and compared to the mean quality of the entire datasets. However, in case of two imprints with very different quality the mean can be simply identical to that of two imprints sharing the same quality, thus, this strategy is simply not sensible. Instead of considering the quality mean of two imprints causing false matches we consider the absolute quality difference between the imprints in the current analysis.

FP Quality Assessment: We focus on two well-known and standardised assessment methodologies:

The first approach is the *NIST FP Image Quality 1.0 (NFIQ 1.0)*[6]. This method was chosen to compare the results of the current study to those of the previously conducted one [8]. The method, included in the NBIS software, uses various FP related information, like minutiae position and local orientations to calculate a quality value from 1 (best) to 5 (worst) [16]. According to the fact that NFIQ 2.0 values range from 0 till 100 (where a value of 0 indicates the lowest quality) we adjusted the NFIQ 1.0's values accordingly to facilitate better comparability. A weighted-sum approach, as proposed in [16] was considered, by applying the same weights as suggested in the original work.

The second FP specific approach is the recent *NIST FP Image Quality 2.0 (NFIQ 2.0)*[7]. This algorithm exhibits increased reliability and accuracy in terms of determining which FP sample is going to fail in the recognition stage, with respect to its previous version (NFIQ 1.0). The method is based on fourteen features with high predictive power selected from 155 quality features reported in literature. The applied classifier (binary classification) was trained using a random forest with all the features describing the feature vector. The final quality score expresses the probability that a FP sample belongs to the FP class of highest utility multiplied by 100 and rounded to its closest integer.

Statistical Methodology: The statistical methodology used to determine the existence of significant quality differences is explained subsequently.

Overall quality analysis: The main objective is to determine if there is a significant difference in terms of quality values with respect to the factors "year" $\{2009, 2013\}$ or "database" $\{A, B_1, B_2, B_3, B_4, B_5\}$. In order to determine if there are significant differences between the groups, non-parametric tests were used, due to the fact that the assumptions of normality and homoscedasticity in all the groups were not met. The Mann-Whitney U-test was used to analyse the "year" factor and the Kruskal-Wallis with Scheffé *post hoc* test is employed for the "database" factor, respectively.

Analysis of the false matches: The variable of interest (Q-diff) is the absolute value of the quality difference of the FP's qualities involved in each match. The objective is to determine if there are significant differences in terms of Q-diff between the detected false matches (FM and FNM) and all performed matches. In order to achieve a balanced design as well as the independence between the samples, a random sampling was made on the set of all matches for each comparison. The analysis was carried out for different decision thresholds and both error distributions (FM matches vs. randSample (all matches) and FNM matches vs. randSample (all matches)). The Mann-Whitney-U test was used in this analysis as well.

Mann-Whitney U-Test [3]: This test is a non-parametric alternative to the t-test for two independent samples. The interpretation of the test is essentially identical to the interpretation of the result of a t-test, except that the U-test is computed based on rank sums rather than means.

Kruskal-Wallis Test [3]: Is an extension of the U-test for more than two independent samples being a non parametric alternative for one-way ANOVA. The test determines whether the medians of two or more groups are different. If the result is considered statistically significant it can be affirmed that at least there are two different groups. Then, which groups significantly differ from each other has to be determined using a multiple comparison (post hoc) test (Scheffé Test), correcting the bias. Multiple comparison tests correct the estimation bias that occurs if all pairs of means (central tendency values) are tested independently with Mann-Whitney or t-student type tests, which causes the type I error to increase, to ensure that type I error does not exceed a pre-established level.

Expectations: In case the imprints' overall quality compared between the data separated in time is statistically dif-

[6] http://www.nist.gov/itl/iad/ig/nigos.cfm#Releases

[7] https://www.nist.gov/services-resources/software/development-nfiq-

data	**VeriFinger** 10			**ANSI**			**ISO**			**IDKiT**			**CFIM**		
	EER	F_{100}	*zFMR*	**EER**	F_{100}	*zFMR*	**EER**	F_{100}	*zFMR*	**EER**	F_{100}	*zFMR*	**EER**	F_{100}	*zFMR*
A	1.57	0.019	0.09	3.48	0.047	0.26	3.47	0.047	0.25	2.96	0.000	0.23	5.07	0.064	0.15
B_1	1.54	0.017	0.03	1.75	0.021	0.08	1.73	0.020	0.07	1.88	0.000	0.04	5.90	0.080	0.15
B_2	0.60	0.006	0.01	1.74	0.019	0.03	1.75	0.019	0.07	1.20	0.000	0.03	5.19	0.068	0.16
B_3	2.66	0.032	0.76	4.09	0.057	0.07	4.15	0.057	0.83	3.57	0.000	0.83	7.77	0.109	0.84
B_4	0.79	0.008	0.85	1.76	0.019	0.75	1.77	0.021	0.74	1.48	0.000	0.75	5.08	0.063	0.96
B_5	1.66	0.018	0.95	2.68	0.038	0.71	2.77	0.038	0.72	2.38	0.000	0.69	7.90	0.109	0.94
C_1	1.70	0.019	0.10	2.02	0.026	0.36	2.04	0.025	0.34	2.67	0.000	0.28	11.92	0.180	0.34
C_2	1.36	0.014	0.01	3.17	0.042	0.33	3.14	0.04	0.27	2.50	0.000	0.29	14.05	0.207	0.71
C_3	2.27	0.026	0.74	3.49	0.049	0.80	3.48	0.049	0.79	3.09	0.000	0.80	13.15	0.207	0.94
C_4	1.35	0.015	0.89	3.05	0.041	0.84	2.99	0.040	0.83	2.57	0.000	0.83	13.50	0.220	0.73
C_5	1.80	0.019	0.96	3.24	0.044	0.77	3.19	0.044	0.78	2.90	0.000	0.74	13.88	0.216	0.98
D_{23}	1.72	0.019	0.67	2.22	0.028	0.73	2.25	0.028	0.72	1.85	0.000	0.73	9.95	0.158	0.92

Table 3. Performance figures for the VeriFinger, Innovatrics-ANSI/ISO/IDKiT and CFIM FP recognition system.

ferent, the (earlier) observed recognition accuracy decrease can be attributed to this quality difference. However, in case the quality is higher for more recently acquired fingerprint data (i.e. ageing has progressed in subjects), time-related effects can be ruled out for being the reason for the described recognition degradation (as ageing cannot be expected to increase quality). Of course, if the quality difference between imprints involved in false matches over time is larger than the difference between imprints where the matches do not involve cross-time matches, we have again a clear indication of quality being responsible for the effects (earlier) observed on time-separated data.

variable	Mann-Whitney-U Test (NFIQ 1.0), $\alpha = 0.01$			
	Rank Sum		**Z**	**p-value**
	2009	**2013**		
values	3134461	14155679	5.5867	0.000

Table 4. Mann-Whitney-U Test (NFIQ 1.0) results based on year information.

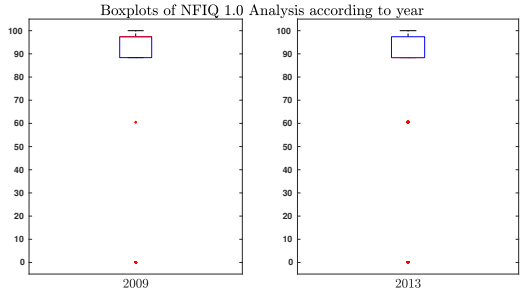

Figure 1. Boxplots of NFIQ 1.0 Analysis done according to year information.

Figure 1 shows the NFIQ 1.0 quality in box plots comparing 2009 to 2013. Although looking pretty similar, please note that the median is clearly lower in 2013, and overall, the quality is very high. The results of the corresponding statistical analysis are shown in Table 4, clearly indicating that the imprints of 2009 and 2013 are not from the same continuous distribution with equal medians. So, contrasting to results on this dataset published earlier [8], we are able to identify statistically significant quality differences between time-separated data using NFIQ 1.0. Corresponding results, obtained when analysing NFIQ 2.0 assessment, are displayed in Table 5 and Figure 2. Statistical analysis of NFIQ 2.0 values confirms the results of NFIQ 1.0 stating a significant quality difference between data separated in time.

variable	Mann-Whitney-U Test (NFIQ 2.0), $\alpha < 0.01$			
	Rank Sum		**Z**	**p-value**
	2009	**2013**		
values	35400185	137501215	13.5742	0.000

Table 5. Mann-Whitney-U Test (NFIQ 2.0) results based on year information.

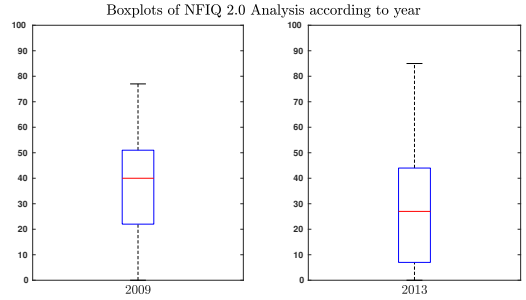

Figure 2. Boxplots of NFIQ 2.0 Analysis done according to year information.

However, there is also a clear difference between NFIQ 1.0 and 2.0 quality, respectively. As detectable in Figure 2, the quality values in both years are much worse for NFIQ 2.0 compared to the NFIQ 1.0 figures, which corresponds much better to the visual impression when looking at the imprints. Further, and even more important for this study, is

the fact that in the box plot the quality of 2013 data is clearly lower compared to 2009 (with also higher amount of variation). Thus, trends observed when considering NFIQ 1.0 figures are more clearly confirmed on NFIQ 2.0 data.
Following the general year specific analysis the set-up was refined to a statistical comparison of the different single databases (using a Kruskal-Wallis test). For both NFIQ 1.0 as well as NFIQ 2.0 these tests resulted in a $p - value = 0$ at a significance level of $\alpha = 0.01$ for all performed experiments, resulting in the conclusion that the imprints of the given distinct databases are of different quality when comparing quality figures over time. A graphical representation of the single databases' quality values using box-plots is shown in Figure 3 (NFIQ 1.0) and Figure 4 (NFIQ 2.0).

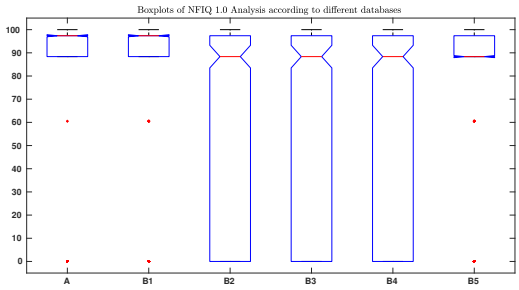

Figure 3. Boxplots of NFIQ 1.0 Analysis done according to the different datasets.

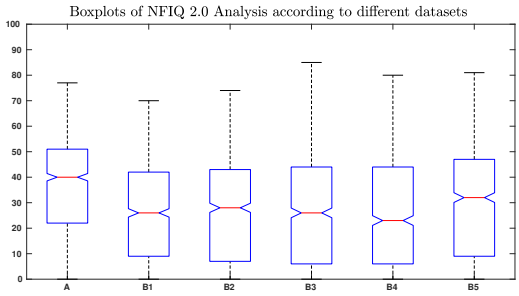

Figure 4. Boxplots of NFIQ 2.0 Analysis done according to the different datasets.

The imprint's NFIQ 1.0 quality in A and B_1 looks identical and also A and B_5 is similar (although the median is clearly lower in B_5). The other 2013 databases contain more low quality images according to NFIQ 1.0. In Figure 4 two observations concerning NFIQ 2.0 results can be made: i) The quality of dataset A is higher as compared to the 2013 databases (as already indicated by the results displayed in Figure 2). ii) The imprints' quality of the single 2013 databases is rather similar and does not exhibit the trend as seen with NFIQ 1.0 (where $B_{2,3,4}$ exhibited clearly worse results). The results of these refined observations considering NFIQ 2.0 using the Kruskal-Wallis Test and a subsequently Multiple Comparisons (Scheffé) Test are shown in in Table 6. In case of rejecting the null hypothesis (quality values are selected from distributions with the same median), the associated values in the table are coloured in red. Thus, it is possible to statistically confirm the results of Figure 4. Each time dataset A is compared to B_1 - B_5 we observe statistically significant difference because of $p = 0$. For comparisons among datasets B_1 - B_5, in most cases it can be assumed that their quality values are selected from distributions with the same median, except for B_5, which exhibits better quality compared to the others.
Based on these results, we are now able to answer research questions (b) and (c) posed in the Introduction: (b) Observed differences are more distinct (and even clearly observed in box plots) for NFIQ 2.0. (c) The application of statistical testing clearly reveals significant differences for both NFIQ 1.0 and NFIQ 2.0 values when comparing quality of 2009 and 2013 data, respectively (which clearly contrasts earlier results in [8]).

Dependt.: Quality	Multiple Comparisons p values (2-tailed); Quality (Quality) Independent (grouping) variable: Sensor Kruskal-Wallis test: H (5, N=5880) = 205.992 4 p = 0.000					
	A	B1	B2	B3	B4	B5
A		0.00	0.00	0.00	0.00	0.00
B1	0.00		1.00	1.00	1.00	0.02
B2	0.00	1.00		1.00	1.00	0.06
B3	0.00	1.00	1.00		1.00	0.002
B4	0.00	1.00	1.00	1.00		0.0006
B5	0.00	0.02	0.06	0.002	0.0006	

Table 6. Kruskal-Wallis Test (NFIQ 2.0) and Multiple Comparisons (Scheff'e Test) results based on database information.

Comparing Figures 3 (NFIQ 1.0) and 4 (NFIQ 2.0) also shows that quality assessment contrasts to the different B_i datasets: While $B_{2,3,4}$ are rated as being of lower quality and B_1 as rather identical to A by NFIQ 1.0, only B_5 is better rated as the other B_i by NFIQ 2.0 and B_1 is very different in terms of quality compared to A according to NFIQ 1.0. Note also, that interestingly the quality values do not at all correspond to the trend in comparison accuracy for some recognition schemes, where worse accuracy is observed for A, B_3 and B_5, where especially A and B_5 are particularly well rated in terms of quality.
After focusing on the overall quality of the entire datasets, we refine the quality based analysis by looking at the quality of imprint-pairs leading to false matches in the recognition experiments (as done earlier in [8]). For each match the absolute difference between the two imprints' quality values is calculated. The determination of the particular false accept and reject matches was done at certain decision thresholds, which have been experimentally chosen to represent the

system's performance best. For NBIS, $5, 10, 20, 30, 50$ have been chosen, for VeriFinger $5, 20, 50, 70, 100$, for CFIM $20, 30, 50, 75, 100$, and for all other recognition systems the selected thresholds are $50, 150, 250, 350, 450$. After computing the quality differences, these values are statistically tested against randomly selected sets of all quality difference values using the Mann-Whitney U-Test once again. The significance level $\alpha = 0.01$ is set like in the previous tests.

For ANSI, ISO, IDKit and VeriFinger it was not possible to observe any false (non) matches at the selected thresholds, which would have been introduced by 2009 vs. 2013 image pairs. Thus, only the false matches from 2009 vs. 2009 and 2013 vs. 2013 imprints can be made responsible for performance figures deviating from perfect accuracy. Consequently, time-related reasons for any observed recognition accuracy decrease can be completely ruled out for these recognition schemes.

The statistical analysis of quality difference of false-match causing imprint pairs considering NFIQ and CFIM displays a similar trend even though the performance of these systems is much worse compared to the commercial SDKs. As example, we show results of the statistical analysis for CFIM in tables 7 and 8. These tables' entries report the $p - values$ for each experiment instance. The cases where significant differences are found with a $p < 0.01$ are highlighted in red colour. The analysis was conducted by selecting *only* the cross year matches for the C_i datasets (2009 vs. 2013 imprints) and all matches for the remaining datasets.

dataset	$p - values$ for false matches				
	20	30	50	75	100
A	0,001	0,006	0.479	0.547	0.431
B_1	0,000	0,000	0.171	0.679	0.311
B_2	0,000	0,002	0.827	0.464	0.785
B_3	0,000	0,000	0,001	0.323	0.618
B_4	0,000	0,000	0.048	0.016	0.467
B_5	0,000	0,000	0.061	0.338	0.931
C_1	0.057	0.903	0.642	0.927	0.981
C_2	0,000	0,002	0.222	0.021	0.044
C_3	0.338	0.599	0.015	0,000	0,001
C_4	0,000	0,000	0.105	0.658	0.562
C_5	0,002	0.268	0.199	0.029	0.028
D_{23}	0,000	0,000	0,004	0.046	0.219

Table 7. Mann-Whitney U-test $\rho - values$ for the quality differences of CFIM false matches against a random sample of all matches using NFIQ 2.0.

We note that for both types of false matches time separation does not play an important role. In contrary, for false (non) matches between imprint pairs of the same year we indeed result in differences to all matches more often as compared to false (non matches) between time-separated imprint pairs. This indicates, that the observed quality differences are not associated to any time-related effects. The same trend is observed for NBIS false (non) comparison results as well. Thus, we are now able to answer research question (d): The consideration of quality differences of imprint-pairs causing false (non) matches is highly valuable in that it reveals that (i) for the better performing COTS SDKs, no time-separated imprint pairs cause any comparison errors and (ii) for the SDKs with lower performance, time-separated imprint pairs causing false (non) matches have no stronger deviation from the distribution of imprint-pair differences over the entire dataset than non time-separated imprint pairs have.

dataset	$p - values$ for false non-matches				
	20	30	50	75	100
A	0,000	0,000	0,000	0,000	0,000
B_1	0,000	0,000	0,000	0,000	0,000
B_2	0,000	0,000	0,000	0,000	0,000
B_3	0,000	0,000	0,000	0,000	0,000
B_4	0,000	0,000	0,000	0,000	0,000
B_5	0,000	0,000	0,000	0,000	0,000
C_1	0.015	0.013	0.184	0.078	0,000
C_2	0.793	0.340	0.049	0,000	0,000
C_3	0,001	0,000	0,000	0,000	0,000
C_4	0,000	0,000	0,000	0,000	0,000
C_5	0,006	0.078	0.043	0.083	0,006
D_{23}	0,000	0,000	0,000	0,000	0,000

Table 8. Mann-Whitney U-test $p - values$ for the quality differences of CFIM false non-matches against a random sample of all matches using NFIQ 2.0.

6. Conclusion

The detailed analysis of the CASIA Fingerprint Subject Ageing Version 1.0 dataset reveals interesting results. Employing more recent FP recognition technologies, earlier observed accuracy reduction when considering comparison across time is significantly reduced (to a slight increase of zFMR only, while EER and F_{100} are no longer affected). Contrasting to earlier results, the authors find statistically significant quality differences between imprints acquired in 2009 and 2013, considering both NFIQ 1.0 and NFIQ 2.0 quality, respectively. The overall quality is consistently rated as being better in 2009 imprint and still suggests time-related reasons for observed effects in recognition performance decrease. The more detailed analysis on imprint pairs causing false (non) matches entirely rules out time-related effects to cause quality differences or reduced recognition accuracy on time-separated data, at least for recent COTS SDKs. But also for less accurate recognition schemes time-related effects do not seem to contribute to observed recognition effects on the investigated dataset. Thus, earlier observed effects are not caused by time-related changes resulting in template ageing, so: It's the quality, stupid !

References

[1] M. Arnold, C. Busch, and H. Ihmor. Investigating performance and impacts on fingerprint recognition systems. In *Information Assurance Workshop, 2005. IAW '05. Proceedings from the Sixth Annual IEEE SMC*, pages 1 – 7, june 2005.

[2] G. R. Doddington, W. Liggett, A. F. Martin, M. A. Przybocki, and D. A. Reynolds. SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Int'l Conf. Spoken Language Processing (ICSLP)*, 1998.

[3] J. D. Gibbons and S. Chakraborti. Nonparametric statistical inference. In *International encyclopedia of statistical science*, pages 977–979. Springer, 2011.

[4] C. Gottschlich, T. Hotz, R. Lorenz, S. Bernhardt, M. Hantschel, and A. Munk. Modeling the growth of fingerprints improves matching for adolescents. *Information Forensics and Security, IEEE Transactions on*, 6(3):1165 – 1169, sept. 2011.

[5] J. Hernández-Palancar, A. Muñoz-Briseño, and A. Gago-Alonso. Using a triangular matching approach for latent fingerprint and palmprint identification. *IJPRAI*, 28(7), 2014.

[6] S. Kirchgasser and A. Uhl. Biometric menagerie in time-span separated fingerprint data. In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'16)*, pages 1–12, Darmstadt, Germany, 2016.

[7] S. Kirchgasser and A. Uhl. Fingerprint template ageing vs. template changes revisited. In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'17)*, pages 1–12, Darmstadt, Germany, 2017.

[8] S. Kirchgasser and A. Uhl. Template ageing and quality analysis in time-span separated fingerprint data. In *Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA '17)*, pages 1–8, New Delhi, Indien, 2017.

[9] S. Kirchgasser and A. Uhl. Template ageing in non-minutiae fingerprint recognition. In *Proceedings of the International Workshop on Biometrics and Forensics (IWBF '17)*, pages 1–6, Coventry, United Kingdom, 2017.

[10] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2002: Second fingerprint verification competition. In *Pattern recognition, 2002. Proceedings. 16th international conference on*, volume 3, pages 811–814. IEEE, 2002.

[11] A. Mansfield. Iso/iec 19795-1 biometric performance testing and reporting: Principles and framework, fdis ed., jtc1/sc37/working group 5, aug. 2005, 2005.

[12] S. Modi and S. Elliott. Impact of image quality on performance: Comparison of young and elderly fingerprints. In *Proceedings of the 6th International Conference on Recent Advances in Soft Computing (RASC'06)*, pages 449–454, 2006.

[13] S. Modi, S. Elliott, J. Whetsone, and H. Kim. Impact of age groups on fingerprint recognition performance. In *IEEE Workshop on Automatic Identification Advanced Technologies*, pages 19–23, 2007.

[14] J. Ryu, J. Jang, and H. Kim. Analysis of effect of fingerprint sample quality in template aging. In *NIST Biometric Quality Workshop II*, nov 2007.

[15] N. Sickler and S. Elliott. An evaluation of fingerprint image quality across an elderly population vis-a-vis an 18-25 year old population. In *Security Technology, 2005. CCST '05. 39th Annual 2005 International Carnahan Conference on*, pages 68 – 73, oct. 2005.

[16] E. Tabassi and P. Grother. Quality summarization. Technical report, 2007. NISTIR7422.

[17] A. Uhl and P. Wild. Comparing verification performance of kids and adults for fingerprint, palmprint, hand-geometry and digitprint biometrics. In *Proceedings of the 3rd IEEE International Conference on Biometrics: Theory, Application, and Systems 2009 (IEEE BTAS'09)*, pages 1–6. IEEE Press, Oct. 2009.

[18] A. Uhl and P. Wild. Ageing effects in fingerprint biometrics. In M. Fairhurst, editor, *Age Factors in Biometric Processing*, chapter 2.4, pages 153–170. IET, London, UK, 2013.

[19] A. Uhl and P. Wild. Experimental evidence of ageing in hand biometrics. In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'13)*, Darmstadt, Germany, Sept. 2013.

[20] D. Woodard and P. Flynn. Finger surface as a biometric identifier. *Computer Vision and Image Understanding*, 100:357–384, 2005.

[21] S. Yoon and A. Jain. Longitudinal study of fingerprint recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 112(28):8555–8560, 2015.