Template Ageing and Quality Analysis in Time-Span separated Fingerprint Data

Simon Kirchgasser
Department of Computer Sciences
University of Salzburg
Jakob-Haringer-Str. 2
5020 Salzburg, AUSTRIA
skirch@cosy.sbg.ac.at

Andreas Uhl
Department of Computer Sciences
University of Salzburg
Jakob-Haringer-Str. 2
5020 Salzburg, AUSTRIA
uhl@cosy.sbg.ac.at

Abstract

We confirm earlier findings on the existence of significant fingerprint template ageing on a dataset acquired with commerical off-the-shelf optical fingerprint sensors exhibiting a time-span of 4 years using two different minutiae-based recognition schemes. A subsequent analysis of the quality of imprints involved in false matches (type 1 and type 2 errors, respectively) does not give clear evidence that reduced quality of time-separated data can be made responsible for the observed template ageing effects. Thus, the cause for the observed template ageing remains unclear and is subject to further investigation.that fingerprint ageing is a possible explanation for the biometric menagerie in time separated data

1. Introduction

Ageing phenomena potentially affect recognition accuracy of biometric systems. The most general term dealing with this issue, *template ageing*, as being defined by ISO/IEC 19795-1:2006 ("Information technology - Biometric performance testing and reporting -...", Section 6.4.6), relates to the fact that longer time intervals generally make it more difficult to match samples to templates. This effect refers to the increase in error rates caused by time-related changes in the biometric pattern, its presentation and the used sensor.

The state-of-the-art in dealing with template ageing varies strongly among biometric modalities [10]. While template ageing effects are quite accepted to be present e.g. in speaker recognition and strategies have been developed in this field to develop algorithms maintaining recognition accuracy using speaker data separated by significant time intervals, for other modalities like iris recognition even the presence of template ageing effects remains controversial. As soon as the presence of template ageing has been demonstrated for a modlaity, the question for its cause arises naturally to be able to eventually better cope with corresponding higher error rates.

For fingerprints (FPs), until now, it is generally assumed that a fingerprint (FP) pattern is fully formed at the gestational age of 24 weeks and is a rather stable modality [6]. Galton's study on the permanence of human FP's ridges and furrows [4] is probably the first study related to FP ageing. Nevertheless, skin ageing can be measured by high-frequency skin ultrasonography, prophilometry, skin microrelief descriptors [5], or skin topography changes from capacity images (analysing 3D profile suffering from ageing, introducing wrinkles and getting deeper enlarging cells) [2]. FP ageing results in loss of collagen [13] causing skin being dryer and looser in elderly people. Furthermore, several acquired skin diseases affect FP recognition accuracy as well [3].

With respect to behavioural effects, less accurate presentation of fingers due to arthritis has been identified [13]. Age groups have been considered in [16] revealing that kids' FP verification performance suffers when compared to adults, in terms of both equal error rate (EER) as well as Receiver Operating Characteristics (ROC).

FP template ageing has been investigated and documented on a low number of datasets so far. Forensic FP data from the German federal criminal police office (BKA) is analysed in [1]. A significant recognition performance degradation for FP in intervals of 10 to 30 years has been found. Another study [18], considering small time-intervals, revealed slight degradation of recognition performance even for a 16 week time-span, based on finger 3D range data. Using a handprint data base (5 year time-span), acquired by a flatbed scanner, the effects of FP template ageing become manifest in a 2-4 times lowered EER performance caused by roughly 33% decreased genuine comparison scores [17]. A recent extensive study [19] confirms the decrease of genuine scores for longer time-spans (maximum 7 years) on forensic data. The only study so far being conducted on time

separated data (4 years) collected with off-the-shelf commerical FP scanners [9] confirms the presence of FP template ageing by analysing user-group specifc effects which are introduced by the so called "Doddington Zoo" concept (not confirming all Doddington Zoo-related results found in [17], i.e. the statement "Short-term goats extend to long-term goats" [17] could not be verified in [9]).
While the presence of FP template ageing effects seems to be pretty well confirmed considering the overall trend in all these results, the reason(s) for these observed effecthat fingerprint ageing is a possible explanation for the biometric menagerie in time separated datats are hardly investigated so far. At least, a recent study [8] excludes sensor ageing as a possible reason for observed FP template ageing. Apart from this result, only [19] looks into factors causing FP template ageing: Covariate-fit analysis reveals that FP image quality explains observed genuine score variation better than subject's age and time-span among FP samples and templates included in the corresponding forensic dataset used in this study.
Based on those recent results we pursue two major objectives with this work analysing the same dataset as in [9] (4 years separated commercial FP scanner data): First, we aim at verifying the presence of FP template ageing using traditional EER and ROC analysis (as in [9], only Doddington Zoo-related effects are documented). Second, we conduct an analysis relating FP image quality to observed incorrect FP matches with the aim to examine the relation of time-span and eventually observed quality differences comparing our results to [19] (using a different, forensic dataset). Based on these results, a discussion on the eventual cause(s) for observed FP template ageing effects will be conducted.
The rest of the paper is organised as follows: In Section 2, the datasets and FP recognition systems employed in experiments are introduced. Section 3 describes the experimental setup wrt. examining template ageing effects and discusses our analysis of the quality of FP images that are involved in type 1 and type 2 matching errors, respectively. Section 4 concludes this paper with a discussion on the eventual cause(s) for observed FP template ageing effects.

2. Datasets and Recognition Systems

When performing an ageing oriented FP analysis, several factors affect the corresponding FP acquisition process: The extent of non-ageing based variability within the dataset should be as low as possible (e.g. concerning illumination differences etc.). Ideally, the sensor used for the first acquisition should be exactly the same as for the other acquisition processes in the following year(s). In this case, cross sensor effects can be excluded and only sensor ageing could influence FP ageing analysis (see [8] for negating this). FP quality, eventually to be considered as a non-ageing-based variability, should therefore be kept constant across time-separated acquisitions.
Basically two main data bases, provided by the biometrics research team at the Center for Biometrics and Security Research (CBSR) at the Chinese Academy of Sciences, Institute of Automation (CASIA) were taken into account in our study: The first dataset, "CASIA 2009", is a subset of the CASIA-FPV5[1] database and contains 980 FP images of 49 volunteers. For both hands, it includes FP scans of forefinger and second finger, 5 prints per finger. All images have been captured by a U.are.U 4000 scanner, produced by DigitalPersona. This is an optical scanner with a resolution of 512 dots per inch (dpi). All FP images are 8-bit/pixel grayscale images and have a resolution of 328x356 pixel. The second dataset, "CASIA 2013", consists of five different datasets of FP images. Each subset contains 980 FP images of the same 49 volunteers as in CASIA 2009, including 20 imprints for each user using the same fingers. Those five single datasets have been acquired by 3 different sensor types. Two instances of U.are.U 4000 scanners and two instances of U.are.U 4500 scanners were used to acquire the imprints of two datasets per scanner. For the remaining 5th dataset a TCRU1C sensor was selected for the acquisition process. The U.are.U 4500 is very similar to the U.are.U 4000 and therefore the specifications with respect to resolution, image dimensions and bit depth are identical. The TCRU1C sensor is a capacitive FP scanner with a resolution of 508 dots per inch (dpi). The imprints have a resolution of 256x360 pixel. FP acquisition by different sensor types enables an analysis of cross-sensor effects during the experiments. In this study, matching score information and corresponding recognition results were obtained by applying two different minutiae based FP recognition systems:
NIST Biometric Image Software (NBIS): Implemented by the National Institute of Standards and Technology (NIST)[2]; in this work release 5.0.0 was used.
VeriFinger (NEURO) The *VeriFinger SDK*[3], developed by Neurotechnology, is minutiae based as well. Release 7.1 includes algorithmic solutions enhancing the performance on rolled and flat FP matching, tolerance to FP translation, rotation and deformation as well as adaptive image filtration.

3. Experiments and Results

3.1. Experimental Setup

In the following experimental analysis 11 different datasets were used. dataset A refers to CASIA 2009. Every time a dataset is named with B as first letter, one of the single datasets of CASIA 2013 is considered. B1 is the dataset acquired with the TCRU1C sensor. B2 and B3 are the datasets for which a U.are.U 4000 sensor was used.

[1] http://biometrics.idealtest.org/dbDetailForUser.do?id=7
[2] http://www.nist.gov/itl/iad/ig/nbis.cfm
[3] http://www.neurotechnology.com/verifinger.html

Table 1: Characteristic individual performance values of NBIS matching for all datasets using all matches.

dataset	**EER**	**AGS**	**AIS**	**FAR$_{100}$**	**ZeroFAR**
single					
A	7.42	64.03	6.78	0.13	0.34
*B*1	8.95	64.87	6.64	0.15	0.39
*B*2	8.17	64.63	6.53	0.13	0.35
*B*3	9.07	53.69	6.83	0.18	0.81
*B*4	**5.96**	70.56	6.37	0.10	0.91
*B*5	7.30	67.30	6.34	0.14	0.97
crossed					
*C*1	**12.63**	47.61	6.58	0.26	0.57
*C*2	14.76	44.71	6.51	0.29	0.58
*C*3	14.37	43.81	6.71	0.29	0.87
*C*4	13.18	49.06	6.52	0.25	0.97
*C*5	13.46	48.65	6.50	0.25	0.99

Table 2: Characteristic individual performance values of NEURO matching for all datasets using all matches.

dataset	**EER**	**AGS**	**AIS**	**FAR$_{100}$**	**ZeroFAR**
single					
A	2.07	508.56	0.005	0.04	0.08
*B*1	3.17	499.66	0.001	0.06	0.08
*B*2	**1.96**	562.11	0.005	0.04	0.06
*B*3	4.00	464.01	0.029	0.08	0.81
*B*4	2.04	553.72	0.013	0.04	0.73
*B*5	3.69	484.50	0.021	0.07	0.98
crossed					
*C*1	**5.32**	356.57	0.002	0.10	0.22
*C*2	5.97	359.21	0.005	0.12	0.25
*C*3	6.16	350.87	0.015	0.12	0.90
*C*4	5.81	368.43	0.006	0.11	0.90
*C*5	6.73	352.25	0.008	0.13	0.99

Table 3: Characteristic individual performance values of NBIS matching for the crossed datasets.

dataset	**EER**	**AGS**	**AIS**	**FAR$_{100}$**	**ZeroFAR**
	crossed - excluding time-span impostor scores				
*C*1	**13.03**	47.58	6.71	0.26	0.54
*C*2	15.22	44.68	6.64	0.29	0.57
*C*3	13.86	43.76	6.80	0.29	0.80
*C*4	13.26	49.14	6.57	0.25	0.71
*C*5	13.43	48.71	6.55	0.26	0.82
	crossed - randomly selected scores				
*C*1	**12.82**	47.56	6.58	0.26	0.53
*C*2	14.79	44.69	6.51	0.29	0.57
*C*3	14.10	43.84	6.71	0.29	0.75
*C*4	13.15	49.07	6.52	0.25	0.60
*C*5	13.38	48.74	6.50	0.25	0.71

B4 and B5 are the datasets acquired with the U.are.U 4500 sensor. Apart from those 6 "single" datasets, 5 so-called "crossed" sets were constructed. Those contain both the imprints of CASIA 2009 (A) and one of the 5 datasets of CASIA 2013 - 1960 images in total. dataset C1 includes the imprints of CASIA 2009 (A) and the TCRU1C sensor recordings of CASIA 2013 (B1). C2 and C3 denote the combination of A and U.are.U 4000 sensor FP images of CASIA 2013 (B2 and B3). The remaining data sets C4 and C5 result from combining A and the U.are.U 4500 sensor CASIA 2013 imprints (B4 and B5). The evaluation of the recognition accuracy is based on the procedure used in all four FP Verification Contests (FVC), for example see [11]. Due to the described specifications of the datasets, a different amount of genuine and impostor scores is computed. For the single sets (A, B1 - B5), 1960 genuine and 95550 impostor matches were computed, respectively. The crossed datasets' (C1 - C5) number of genuine scores and impostor scores is 4.5 times and 4 times the size of the single sets, respectively. In order to be able to directly compare performance figures without introducing bias due to different dataset size, the identical number of matching scores should be employed for all datasets during the performance evaluation. For this purpose, a randomized selection strategy of the scores was conducted for the crossed datasets to select 1960 genuine and 95550 impostor matches. To ensure a balanced evaluation of the performance figures this selection was repeated $\binom{10}{5}$ times and the obtained performance results were averaged.

3.2. Recognition Accuracy Analysis

For determining recognition accuracy, 5 different characteristic performance figures were considered for all the subsets described in Section 2: Equal Error Rate (EER %), Average Genuine Score (AGS), Average Impostor Score (AIS), the lowest FRR for FAR less or equal to 0.1% (FAR$_{100}$), and Zero False Acceptance Rate (ZeroFAR). The best achieved EER values are indicated in bold in the following tables. For both recognition systems the performance figures are displayed in Tables 1 – 4. In the first experiments, no randomised selection of the matching scores was done. Instead the results include all possible genuine and impostor matches (see Tables 1 and 2). Randomised matching selection was applied in two variants and results in a balanced number of matches for all datasets (see Section 3.1). The first variant only considers impostor scores without including any time-span matches in the randomised selection, thus only matches among imprints of the same year were taken into account. The second variant uses a randomised selection of all genuine and impostor scores. The results are listed in Tables 3 and 4. There is hardly any difference among the various experimental results for the crossed datasets. Neither the random selection strategy to avoid bias due to unbalanced dataset sizes nor the inclusion of time-separated imposter scores have any impact on the overall trends observable in the results. The most important result which can be seen in Tables 1 – 4 is the significant increase of the EER, FAR$_{100}$, and ZeroFAR figures comparing the single and crossed datasets. Looking at

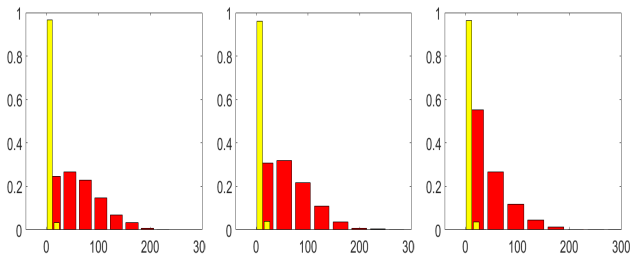
Table 4: Characteristic individual performance values of NEURO matching for the crossed datasets.

dataset	**EER**	**AGS**	**AIS**	**FAR$_{100}$**	**ZeroFAR**
	crossed - excluding time-span impostor scores				
$C1$	**5.29**	356.93	0.004	0.10	0.18
$C2$	5.98	359.12	0.006	0.12	0.21
$C3$	6.21	350.21	0.017	0.12	0.82
$C4$	5.75	368.77	0.009	0.11	0.55
$C5$	6.73	352.36	0.013	0.13	0.77
	crossed - randomly selected scores				
$C1$	**5.33**	356.95	0.002	0.10	0.17
$C2$	6.00	359.12	0.005	0.12	0.18
$C3$	6.16	350.82	0.015	0.12	0.71
$C4$	5.80	368.33	0.006	0.11	0.44
$C5$	6.71	352.39	0.009	0.13	0.64

the AGS and AIS values, a clearly observable reduction of the genuine scores and a more or less stable behavior of the impostor scores can be observed. These effects are present for both NBIS and NEURO recognition schemes. In fact, the stability of the AIS values and the high decrease in the AGS leads to the assumption that ageing-related degradations could be responsible for these observations and confirms earlier finding in ageing-related analysis (e.g [17, 19]). However, based on the results, it can be stated that not the security aspect but the user convenience is impacted by the decreased genuine scores. Another interesting effect can be observed focusing on cross-sensor effects. As opposed to the expectations, datasets C2 and C3, which contain time-separated data acquired with the same sensor type, do *not* exhibit the best accuracy results. For NBIS, even the opposite is observed for EER and FAR$_{100}$. Thus, it seems that an identical sensor type is not important to get good results but, as we shall see in the subsequent section, the quality of the acquired FP images determines the resulting accuracy no matter which sensor type is used.
The tendency of the genuine scores becoming more similar to the impostor ones for the crossed datasets is displayed graphically in Figures 1 and 2, where the x-axis denotes the matching scores and y-axis the percentage scaled from 0 to 1. For all crossed datasets the situation is similar. A clear shift of the genuine score distribution to the left approaching the impostor distribution can be observed for C2 data for both recognition schemes. Thus, we have a very clear confirmation of template ageing effects in terms of quantitative performance measures and qualitative genuine score distribution shape. In the following Section 3.3 we will investigate the role and contribution of FP quality in/to the observed effects.

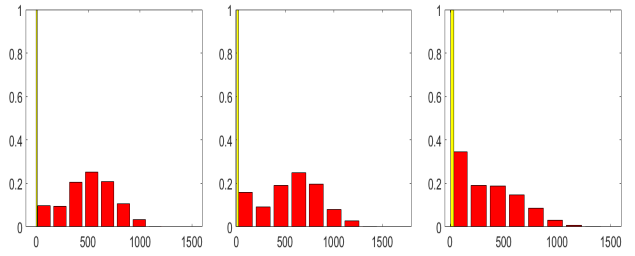
3.3. Fingerprint Quality Analysis

For this analysis, we use the following rationale: We determine the quality of FP images involved in type 1 and type 2 erroneous matches when considering time separated data.



a: Set A b: Set B2 c: Set C2

Figure 1: Genuine (colored red) and Impostor (colored yellow) score distribution of NBIS A, B2 and C2 dataset.



a: Set A b: Set B2 c: Set C2

Figure 2: Genuine (colored red) and Impostor (colored yellow) score distribution of NEURO A, B2 and C2 dataset.

If the quality of these FPs is not lower than the overall average quality of FPs, it is not lower quality that causes the observed template ageing effects and a thorough investigation is required for identifying the actual reasons (e.g. different acquisition conditions not impacting measured quality or subject ageing).
There are different approaches to determine quality of FP images. The first FP specific approach is the NIST FP Image Quality (NFIQ)[4], also included in the NBIS software, which uses various information of a FP image, like minutiae feature information and local orientations to calculate a quality value from 1 (best) to 5 (worst) [15]. The second FP specific approach, the Image Quality of FP (IQF)[5] [14], selects certain parameters of the power spectrum based on the Fast Fourier Transform (FFT) to calculate a score between 0 (worst) and 100 (best). An entirely different approach is to determine generic image quality without exploiting the fact that the input images are containing FPs [7]. There are various generic quality metrics which ca be used for that purpose. We used the non-reference metric Blind Referenceless Image Spatial Quality Evaluator (BRISQUE)[6] [12]. As result of this measurement a score value between 100 (worst) and 0 (best) is obtained.

[4]http://www.nist.gov/itl/iad/ig/nigos.cfm#Releases
[5]http://www2.mitre.org/tech/mtf/
[6]http://live.ece.utexas.edu/research/Quality/index.htm

The FP image examples displayed in Figure 3 have the

a: A: userID 13 #1 b: userID #3 A: 13 c: B2: userID 13#8

Figure 3: Volunteer 13 representing NFIQ/IQF value: 2/97 same NFIQ/IQF value but very different similarity scores (eventually caused by ageing effects). All of the imprints of user 13 have an NFIQ/IQF value of 2/97. The NBIS/NEURO matching scores between imprint 13 #1 and 13 #3 are 60 and 606, respectively. The matching scores between 13 #1 and 13 #8 are 16 and 84, respectively. Although all imprints exhibit identical biometric quality, the matching scores between imprints separated by a time span are considerably lower. In the figures, minutiae not present in the corresponding imprint separated by 4 years time span are depicted by green crosses (see colour print). Obviously, a considerable amount of minutiae is changed which may be a consequence of the type of matcher and imprints used. In order to augment this qualitative finding with quantitative results, we first compute average biometric quality values of all datasets.

As NFIQ values must not be directly averaged, a weighted-sum approach was used instead, as introduced in [15]. We used the identical weights as suggested in the original work. After computing the NFIQ average, a [0,100] range is obtained for all figures (NFIQ, IQF and BRISQUE), where a value of 0 indicates the lowest quality (NFIQ, IQF) and 100 denotes the best (NFIQ, IQF) or the other way round (BRISQUE). In the leftmost column of Tables 5 – 9, the results of computing these average quality values are displayed for reference.

In order to determine the quality of the imprints involved in type 1 errors (false accepts) and type 2 errors (false rejects), we defined a set of 5 thresholds experimentally, for which the false accepted and false rejected matches (and the corresponding imprints involved) have been determined. We choose the thresholds to cover a wide range of operating conditions of the recognition systems, depending either on the NBIS or the NEURO software.

In Table 5 the quality results for the false accepted matches employing NBIS are displayed. Comparing the results with the average quality values of the entire datasets in the leftmost column, it is obvious that except for a few threshold dependent fluctuations no highly significant difference

Table 5: Average image quality of imprints involved in false accepted matches depending on different decision thresholds using NBIS.

all img. avg.	*dataset*	*threshold values for false accepts*				
		1.0	5.0	10.0	20.0	30.0
NFIQ						
78.40	**A**	78.29	78.29	78.99	79.89	88.52
85.33	**B**1	85.27	85.25	86.06	87.17	97.07
65.84	**B**2	66.18	66.30	68.18	81.04	82.95
64.09	**B**3	64.24	64.24	65.27	78.29	72.73
69.73	**B**4	69.88	69.88	70.43	76.31	58.95
73.09	**B**5	73.23	73.21	73.11	75.10	76.62
81.87	**C**1	81.83	81.80	82.23	83.66	91.54
72.12	**C**2	72.31	72.33	73.96	78.94	86.99
71.25	**C**3	71.34	71.32	72.19	80.06	91.74
74.06	**C**4	74.15	74.19	74.48	78.62	93.43
75.75	**C**5	75.82	75.88	76.06	79.45	80.59
IQF						
95.34	**A**	95.34	95.34	95.46	95.69	97.00
92.85	**B**1	92.83	92.82	92.78	93.32	96.31
95.26	**B**2	95.25	95.25	95.55	96.03	96.62
95.11	**B**3	95.10	95.10	95.48	95.64	91.22
95.25	**B**4	95.24	95.24	95.60	96.59	96.80
95.10	**B**5	95.09	95.09	95.48	96.26	96.40
94.10	**C**1	94.09	94.10	94.28	94.90	96.50
95.30	**C**2	95.30	95.29	95.58	95.71	96.89
95.23	**C**3	95.22	95.22	95.31	96.06	95.91
95.32	**C**4	95.29	95.29	95.34	95.18	92.70
95.22	**C**5	95.22	95.21	95.20	95.81	96.77
BRISQUE						
49.63	**A**	49.56	49.56	49.46	49.52	49.72
31.52	**B**1	31.57	31.62	31.67	31.10	30.89
45.28	**B**2	49.51	49.46	49.16	48.91	48.89
44.51	**B**3	49.51	49.51	49.25	49.58	49.47
46.75	**B**4	49.51	49.51	49.14	49.48	49.48
50.14	**B**5	49.51	49.51	48.90	49.02	49.03
40.58	**C**1	43.59	43.60	43.56	43.40	43.42
47.46	**C**2	49.57	49.55	49.39	49.35	49.44
47.07	**C**3	49.57	49.57	49.42	49.60	49.68
48.19	**C**4	49.57	49.57	49.38	49.45	49.53
49.89	**C**5	49.57	49.57	49.29	49.33	49.43

is present for the NBIS results. For NFIQ and IQF, there is a slight tendency for higher quality (i.e. higher values) observed for the imprints involved in false accepts as compared to the datasets' average (!), while for BRISQUE, quality is lower (i.e. higher vaulues) for imprints involved in false accepts (except for datasets A and B5).

For the NEURO software a similar behaviour is observed (results are not shown and not discussed in detail as the false accepted matches are not responsible for the observed template ageing effects).

In the following we concentrate the attention to the false rejected matches. In Tables 6 and 7 the averaged quality values of the imprints involved in false rejected matches for the single datasets (A, B1 - B5) are displayed. Note that these imprints and their average quality also correspond to the subset of false rejected matches in the crossed sets C1 - C5 in case no matches among imprints acquired in different

years are considered. As IQF exhibits a very low extent of variability overall which can hardly be sensibly interpreted we refrain from further presenting and discussing IQF results.

In Table 6 we observe non-consistent results among different datasets for the NBIS results. For NFIQ, quality of imprints involved in false rejected matches is lower than the average dataset quality (leftmost column) for datasets A and B1 and vice versa for sets B2 - B5. For BRISQUE, B1 and B3 exhibit lower quality for imprints involved in false rejected matches compared to average values, and vice versa for dataset B5 while A, B2 and B4 do not exhibit a clear trend. Overall, we do not observe a clear trend towards lower quality for imprints involved in false rejected matches for not time-sparated (i.e. single) datasets for NBIS.

The NEURO results are displayed in Table 7. Contrasting to NBIS results, we observe quite uniform behaviour for NFIQ: For all datasets, the quality of imprints involved in false rejected matches is significantly lower as compared to the average of the respective datasets. Surprisingly, the quality as determined by BRISQUE does not follow this trend, as for datasets A, B4 and B5 quality of imprints involved in false rejects is slightly higher compared to the average values, while for B1 - B3 the opposite is the case. Thus, overall, we observe a clear trend towards lower quality for lower quality for imprints involved in false rejected matches only for NFIQ, however, still for not time-separated data.

Table 6: Average quality of imprints involved in false rejected matches depending on different decision thresholds using NBIS.

all img. avg.	*dataset*	*threshold values*				
		1.0	5.0	10.0	20.0	30.0
NFIQ						
78.40	**A**	51.12	51.12	55.25	59.20	62.50
85.33	**B**1	91.74	71.97	62.98	66.60	71.54
65.84	**B**2	88.37	73.93	77.39	77.61	78.78
64.09	**B**3	88.37	72.73	75.24	73.88	74.24
69.73	**B**4	70.61	70.61	77.63	75.94	76.23
73.09	**B**5	73.03	76.82	79.29	77.48	77.64
BRISQUE						
49.63	**A**	64.58	47.52	45.92	47.00	47.12
31.52	**B**1	49.03	40.72	46.63	44.78	41.24
45.28	**B**2	47.82	45.16	47.13	45.67	45.16
44.51	**B**3	47.84	44.93	47.20	45.53	44.98
46.75	**B**4	47.16	45.58	47.10	45.72	45.56
50.14	**B**5	47.66	47.14	47.81	47.00	47.16

So far, we have not yet extended the quality analysis to "crossed" datasets C1 - C5 which are affected by the template ageing effects. This is done in the following. The detected false rejected matches in datasets Ci were separated into three classes during the refined analysis. The first and third class only contain those matches which have been performed among imprints of the same year (A: 2009 vs 2009 and B1 - B5: 2013 vs 2013), so they exactly correspond to the matches covered in Tables 6 and 7. The second class consists of the remaining cross-year matches: 2009 vs 2013. If the quality values of imprints involved in false rejects of this second class do not exhibit a high extent of degradation compared to the average values of the corresponding entire data sets or if the degradation is less pronounced as compared to the first and second class, then it is quite clear that reduced quality cannot be responsible for the reduced recognition accuracy exhibited for the crossed (i.e. time separated) datasets. The results of the refined analysis of NBIS and NEURO false rejected matches based on their NFIQ/BRISQUE values are shown in Tables 8 and 9.

Table 7: Average quality of imprints involved in false rejected matches depending on different decision thresholds using NEURO.

all img. avg.	*dataset*	*threshold values*				
		5.0	20.0	50.0	70.0	100.0
NFIQ						
78.40	**A**	47.42	47.42	47.42	47.12	45.84
85.33	**B**1	45.53	45.53	46.09	44.99	46.33
65.84	**B**2	24.16	24.16	23.64	24.27	21.61
64.09	**B**3	28.28	28.28	27.93	28.17	26.92
69.73	**B**4	26.53	26.53	25.95	29.37	29.13
73.09	**B**5	34.46	34.46	34.76	31.63	37.51
BRISQUE						
49.63	**A**	43.83	43.83	43.83	44.44	44.57
31.52	**B**1	43.40	43.40	43.40	43.32	43.09
45.28	**B**2	45.80	45.80	45.80	46.00	45.57
44.51	**B**3	44.98	44.98	44.98	44.96	45.05
46.75	**B**4	46.13	46.13	46.14	46.26	46.18
50.14	**B**5	46.74	46.74	46.70	46.44	46.13

Apart from threshold and dataset dependent random fluc-

Table 8: NFIQ/BRISQUE quality refinement analysis using NBIS false rejects information.

all img. avg.	*dataset*	*threshold values*				
		1.0	5.0	10.0	20.0	30.0
	class 1	**NFIQ/BRISQUE -** 2009 **vs.** 2009 see dataset A in Table 6				
	class 2	**NFIQ -** 2009 **vs.** 2013				
81.87	**C**1	62.30	63.45	68.39	71.26	74.04
72.12	**C**2	74.30	69.15	73.51	75.22	76.07
71.25	**C**3	90.17	73.15	69.38	74.24	76.32
74.06	**C**4	67.97	73.82	78.45	67.97	71.12
75.75	**C**5	47.58	66.35	67.30	72.67	74.16
		BRISQUE - 2009 **vs.** 2013				
40.58	**C**1	47.43	49.22	47.08	48.26	48.89
47.46	**C**2	46.72	49.14	48.35	48.93	49.15
47.07	**C**3	46.79	49.53	48.20	48.92	49.36
48.19	**C**4	46.65	49.21	47.50	48.23	48.88
49.89	**C**5	46.71	49.09	47.76	48.74	49.14
	class 3	**NFIQ/BRISQUE -** 2013 **vs.** 2013 see dataset B1-B5 in Table 6				

tuations some interesting observations can be made. The NBIS results for imprints involved in false rejects displayed in Table 8 exhibit slightly worse quality (i.e. higher values) compared to the average C1 - C5 quality as shown in the leftmost column in case of BRISQUE (only for C1, the difference is rather significant). NFIQ values do not exhibit a clear trend: while for C1 and C5 average quality is definitely better, C2 - C4 results depend on the specific decision threshold considered and seem to be quite randomly distributed.

When comparing the false rejected matches' NFIQ quality values to those within datasets A and B1 - B5 (compare Table 6), C1 - C5 values are between A and B1 - B5 values, respectively. Thus, these results do not indicate that the reduced NFIQ quality of time separated matches is responsible for template ageing as observed. On the other hand, for BRISQUE, C1 - C5 values are worse compared to A and B1 - B5 values indicating slightly degraded quality for time separated matches.

Table 9 shows the results obtained when analysing false rejected matches of the NEURO software. For datasets C1, C4, and C5 NFIQ quality values are clearly worse for imprints involved in false rejects as compared to the datasets' average value as shown in the leftmost column. Also for C2 and C3 this is true for most decision threshold values considered. Interestingly, the opposite is the case for BRISQUE results except for C1, where the average result is clearly superior to the false rejected imprints results.

When comparing the false rejected matches' NFIQ quality values to those within datasets A and B1 - B5 (compare Table 7), C1 - C5 values exhibit significantly higher quality compared to A and B1 - B5 values (which shows again, that reduced NFIQ quality of time separated matches is not respondible for template ageing), while the opposite is true for BRISQUE quality values. Thus, while reduced NFIQ quality of imprints involved in false rejected matches definitely cannot be made responsible for template ageing results (interestingly, quality of time-separated false rejected matches is better in several cases), there are some settings where reduced BRISQUE quality can be observed in some of these imprints. Still, results are not consistent enough to make a general statement.

So far, we have only shown averaged results. However, considering the quality mean only may hide important aspects of the distribution of the quality values. Thus, we exemplary present boxplots of the NEURO and NBIS quality values corresponding to thresholds 20.0 and 10.0 from Table 9 and 8. The green squares in the plots depict the average values and the red lines display the median values (see colour print). In Fig. 4 we notice that both mean and median values of the time separated false rejected matches are higher (i.e. better quality) as compared to non time separated matches within datasets C1 - C5. In Figure 5 there

Table 9: NFIQ/BRISQUE quality refinement analysis using NEURO false rejects information.

all img. avg.	*dataset*	*threshold values*				
		5.0	20.0	50.0	70.0	100.0
	class 1	**NFIQ/BRISQUE** - 2009 **vs.** 2009 see dataset A in Table 7				
	class 2	**NFIQ** - 2009 **vs.** 2013				
81.87	**C**1	68.04	68.04	68.71	69.44	70.16
72.12	**C**2	71.91	71.91	72.39	71.87	72.18
71.25	**C**3	68.51	68.51	68.35	70.59	72.21
74.06	**C**4	61.86	61.86	62.07	65.39	67.08
75.75	**C**5	65.89	65.89	65.99	65.98	69.07
		BRISQUE - 2009 **vs.** 2013				
40.58	**C**1	46.15	46.15	46.28	46.83	47.52
47.46	**C**2	46.93	46.93	47.01	47.16	47.91
47.07	**C**3	46.57	46.57	46.62	46.94	47.62
48.19	**C**4	47.15	47.15	47.13	47.04	47.65
49.89	**C**5	46.69	46.69	46.68	47.13	48.12
	class 3	**NFIQ/BRISQUE** - 2013 **vs.** 2013 see dataset B1-B5 in Table 7				

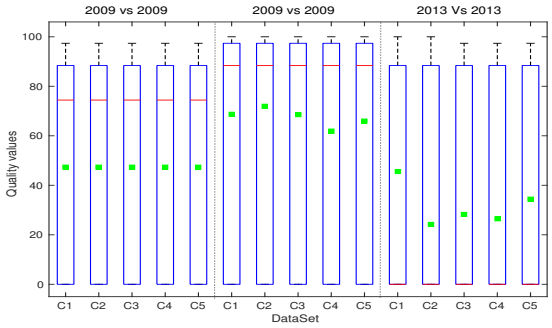


Figure 4: NFIQ quality of NEURO false rejected matches based on decision threshold 20.0.

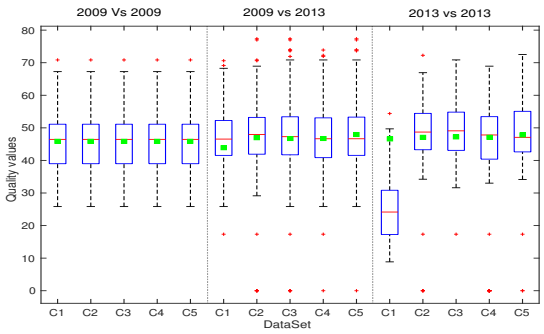


Figure 5: BRISQUE quality of NBIS false rejected matches based on decision threshold 10.0.

is hardly any difference between the three depicted cases, except for the C1 dataset in 2013 vs 2013 matches. The box is positioned much lower (higher quality) compared to the others. This indicates better quality, but including some very poor quality values as well because the average quality

value is similar to the other average quality results.

4. Conclusion

In the first part of this study we have clearly observed template ageing for 4 years separated FP data acquired with off-the shelf commercial optical FP scanners. Results are consistent for two different minutiae-based FP matching schemes. In particular, we detect clearly increased EER, FAR_{100}, ZeroFAR and correspondingly decreased AGS for time separated data, while the AIS remains remarkably stable. These results confirm earlier findings on FP template ageing on forensic FP datasets [1, 19] and on a dataset acquired with a flatbed scanner [17].
A subsequent analysis of the quality of imprints involved in false matches (type 1 and type 2 errors, respectively) did not give clear evidence that reduced quality of time-separated data can be made responsible for the observed template ageing effects. Of course, the question remains, which effects cause template ageing and how this can be mitigated or avoided at all. Based on the observed results, effects are caused by a phenomenon not reducing (but eventually increasing) FP specific NFIQ quality but slightly reducing generic image quality as measured by BRISQUE. In this context, corresponding systematic acquisition setting differences or even subject ageing effects and corresponding changes in the finger tips' physiology are candidates to cause the observed effects. Of course, the robustness of employed FP recognition schemes plays an important role – while for less robust schemes template ageing might be observed, the results of a more robust scheme eventually will not be influenced at all. Finally, it has to be stated that observing a decrease in FP image quality would not rule out subject ageing effects as possible reasons for template ageing – subject ageing effects might just impact on FP quality as well.

References

[1] M. Arnold, C. Busch, and H. Ihmor. Investigating performance and impacts on fingerprint recognition systems. In *Information Assurance Workshop, Proc. from the 6th Annual IEEE SMC*, pages 1–7, 2005.

[2] A. Bevilacqua and A. Gherardi. Age-related skin analysis by capacitance images. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 703 – 706 Vol.2, aug. 2004.

[3] M. Drahansky, M. Dolezel, J. Urbanek, E. Brezinova, and T.-H. Kim. Influence of skin diseases on fingerprint recognition. *Journal of Biomedicine and Biotechnology*, 2012:Article ID 626148, 2012.

[4] F. Galton. *Finger Prints*. Macmillan, London, 1892.

[5] M. Gniadecka and G. Jemec. Quantitative evaluation of chronological ageing and photoageing in vivo: studies on skin echogenicity and thickness. *British J. of Dermatology*, 139:815–821, 1998.

[6] 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.

[7] J. Hämmerle-Uhl, M. Pober, and A. Uhl. General purpose bivariate quality-metrics for fingerprint-image assessment revisited. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'14)*, Paris, France, Oct. 2014.

[8] C. Kauba and A. Uhl. Fingerprint recognition under the influence of sensor ageing. In *Proceedings of the 4th International Workshop on Biometrics and Forensics (IWBF'16)*, pages 1–6, Limassol, Cyprus, 2016.

[9] 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.

[10] A. Lanitis. A survey of the effects of ageing on biometric identity verification. *IET Computer Vision (Special Issue on Future Trends in Biometric Processing)*, 5(6):338–347, 2011.

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

[12] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[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] N. B. Nill. IQF (Image Quality of Fingerprint) software application. Technical report, 2007. MTR 070053.

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

[16] 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.

[17] 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.

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

[19] 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.