Influence of Iris Template Aging on Recognition Reliability

Adam Czajka[1,2](✉)

[1] Institute of Control and Computation Engineering,
Warsaw University of Technology, Warsaw, Poland
[2] Biometrics Laboratory, Research and Academic Computer Network (NASK),
Warsaw, Poland
aczajka@elka.pw.edu.pl
http://www.pw.edu.pl

Abstract. The paper presents an iris aging analysis based on comparison results obtained for four different iris matchers. We collected an iris aging database of samples captured even eight years apart. To our best knowledge, this is the only database worldwide of iris images collected with such a large time distance between capture sessions. We evaluated the influence of the intra- vs. inter-session accuracy of the iris recognition, as well as the accuracy between the short term (up to two years) vs. long term comparisons (from 5 to 9 years). The average genuine scores revealed statistically significant differences with respect to the time distance between examined samples (up to 14 % of degradation in the average genuine scores is observed). These results may suggest that the iris pattern ages to some extent, and thus appropriate countermeasures should be deployed in application assuming large time distances between iris template replacements (or adaptations).

Keywords: Biometric template aging · Iris recognition · Biometrics.

1 Introduction

The statement of a high temporal stability of iris features, in the context of personal identification, appeared as early as in the Flom's and Safir's US patent granted in 1987 [1]. A claim, drawn from a clinical evidence, said that 'significant features of the iris remain extremely stable and do not change over a period of many years'. Although these 'significant features' were not clearly specified in the patent, its context (i.e., recognition of one's identify) suggests that said features should relate to all the iris characteristics having a power to individualize a human within a population. The pioneering work by John Daugman [2] includes more precise suggestion and relates to the high stability of the iris trabecular pattern (as the iris texture is used directly to calculate an iris code). The stability of the iris meshwork is put in contrast to possible changes of other characteristics of the eye, not commonly used in iris recognition, e.g., a melanin concentration

M. Fernández-Chimeno et al. (Eds.): BIOSTEC 2013, CCIS 452, pp. 284–299, 2014.
DOI: 10.1007/978-3-662-44485-6_20

responsible for an eye color. Flom's and Daugman's statements are thus very often cited in the iris recognition literature, fueling a common belief that iris templates, ones determined, are useful for unspecified, yet very long time periods.

Recently this highly desired attribute of biometrics seems to be undermined for iris modality by experimental results revealing an increasing deterioration of a recognition accuracy when the time distance between capturing the gallery and probe images extends significantly, e.g., to a few years. This suggests that the initial claim related to the stability of iris texture might be inaccurate.

However, one should aware that the stability of iris texture only partially contributes to the stability of the iris templates, and many factors may influence the template lifespan. Iris is not exposed to the external environment and it is covered by a transparent fluid (an aqueous humor) that fills the space (an anterior chamber) between the cornea and the iris. Hence, capturing the iris image relies on registering this complicated, three-dimensional structure constituting the frontal, visible part of the eye. Although an iris is the most apparent element of this structure, the aging related to the aqueous humor or the cornea may also influence the aging of iris templates, even if the iris tissue is immune to a flow of time. Moreover, the equipment flux should be considered as an influential element of the template aging phenomenon. This may relate to replacement of the camera between the gallery and probe image capture, or wearing the camera components out. Next, the consequences of the template aging should be distinguished from effects related to inter- and intra-session matching scores. Intra-session comparison scores will typically exhibit a better match among images when compared to the corresponding inter-session results. However, the inter-session changes in imaging conditions, e.g., environmental parameters or subject's interaction with the equipment, usually blur more subtle aging effects related to the eye biology. From the technological point of view, the iris biometric features, not images, are used to finally judge on the extent to which the aging occurs, as we are interested in how this phenomenon transforms from the image space (possible to inspect visually) to the feature space (natural for biometric matching). However, the transformation between these spaces is always proprietary to an iris coding method, and the strength of the template aging effect may depend on the feature extraction methodology. Last but not least, we have no guarantee that the aging effect will be evenly observed across the subjects of different populations. Experiment results obtained with a particular database of images may be a weak predictor of this phenomenon for people of different race, health or dietary culture.

The expected stability of the iris pattern may also be regarded (in a broader sense) as a demand for *stationarity*. Stationary time series is characterized by temporal stability of its statistical properties. However, stability of one property (e.g., the average value) does not guarantee the stationarity, as other properties (e.g., sample variance) may still vary over time. This makes the research of the aging phenomenon even more complicated, as the judgment should not be based solely on properties of a single statistics (e.g. monotonic behavior of the average matching score).

Above aspects related to discovering the truth about the iris aging urgently call for experiments carried out for different populations across the world, performed in different environments, for as many feature extraction methods as possible and for the longest possible time lapse between measurements. Answering this call, we present the iris aging analysis based on comparison results obtained for four iris matchers and the biometric samples captured even eight years apart. To our best knowledge, eight years is the longest time interval characterizing samples used in the iris aging analysis up to date worldwide.

2 Related Work

Iris recognition is relatively young discipline, and thus there is still a shortage of data-bases of iris images collected with adequate time intervals to observe the template aging phenomenon. This is why the literature devoted to the iris template aging is still limited.

Gonzalez *et al.* first addressed a possibility of influence of the time lapse onto iris recognition accuracy [3]. They estimated coding method parameters using a part of the multimodal BioSec database, containing samples of 200 subjects. Final results were generated with the use of the BioSecurID database containing iris images captured for more than 250 volunteers. Although the databases used were reasonably large, the time lapse between image captures in the test database was very short (one to four weeks). As the observation of the aging effects in such a short time period may be difficult, the authors focused on inter- vs. intra-session recognition accuracy analysis. According to the expectations, the genuine intra-session comparisons revealed a better match (e.g. FNMR $\in (0.085, 0.113)$ @FMR = 0.01)[1] when compared to the inter-session results (e.g. FNMR $\in (0.224, 0.258)$ @FMR = 0.01). No significant differences in comparison scores can be observed for inter-session results with respect to these very short time intervals. Thus any conclusions on the iris aging cannot be drawn based on this work, yet it supports the intuition related to the importance of the enrollment procedure that should produce the enrollment samples predicting, to the maximum possible extent, the inter-session variations.

The intra- vs. inter-session variations in iris matching scores were also studied by Rankin *et al.*, who used a database of images captured for 238 subjects [4]. The sessions were separated by three and six months time periods. Results are presented separately for irises grouped in classes depending on the iris texture density, and support a claim on the increase of false rejections when time interval between samples increases. However, this study lies slightly next to the main course of biometrics technology, as the images were captured in visible, not near-infrared light and by a specialized biomicroscope, not typically used in iris recognition.

[1] FNMR (False Non-Match Rate) is an empirical estimate of the recognition method error relying on falsely rejecting a genuine sample; FMR (False Match Rate) is an empirical estimate of the recognition method error relying on falsely accepting an impostor sample.

Baker *et al.* presents the first known to us analysis of the iris aging under long, four-year time lapse [5]. A small database consisted of images captured for 13 volunteers was used in the analysis with the iris segmentation inspected manually (this excludes the segmentation errors from the source of matching score deterioration). As opposed to Gonzalez *et al.*, they eliminated intra-session scores, and the analysis was focused on comparison between short-time-lapse matches (i.e., for images taken a few months apart) and long-time-lapse-matches (i.e., for images taken four years apart). The authors found a statistically significant difference in the average comparison scores between short-time-lapse matches and long-time-lapse matches, namely the genuine comparison scores (based on the Hamming Distance) increased by 3–4 % for long-time-lapse-matches, and the simultaneous change in the impostor scores was not observed. Bowyer *et al.* continue Baker's work presenting the results for slightly enlarged database of iris images captured for 26 subjects [6]. Again, the comparisons between short-time-lapse matches (i.e. for images separated by less than 100 days) and long-time-lapse-matches (i.e. for images taken at least 1000 days apart) are analyzed, and statistically significant deterioration in the genuine comparison scores is reported (increase of the Hamming Distance by approximately 4 %). Later, Baker *et al.* expand their initial work by the use of additional matcher submitted by the University of Cambridge to the Iris Challenge Evaluation 2006 [7]. The authors report an increase in false rejection rate for longer time lapse between images, supporting an evidence of an iris template aging effect. Simultaneously, they concluded that pupil dilation, contact lenses and amount of iris occlusions were not significant factors influencing the aging-related results.

Fenker and Bowyer [8] presented the first study based on comparison results obtained by more than one coding method, with one being a well-recognized commercial product (VeriEye; used also in this paper). The database, consisted of images separated by two years interval, was built for 43 volunteers. The authors, similarly to the previous studies, generated short-time-lapse (from 5 to 51 days) comparisons and long-time-lapse (from 665 to 737 days) comparisons, and the aging effect is studied through observation of the increase of false rejections as a function of time interval. Although we expected an increase of FNMR when the time interval grows, the reported numbers are surprisingly large and alarming. Namely, FNMR increased by 157 % to 305 % for the authors' matcher, and by 195 % to 457 % for a commercial matcher, depending on the acceptance threshold set optimally for short-time-lapse comparison scores. The authors created data subsets with images presenting homogeneous pupil dilation and captured for eyes not wearing contact lenses, yet the results obtained for these data subsets did not show a clear evidence of the significant influence that these factors might have onto the original conclusions. The same authors have broadened their research and used images separated by one-, two- and three-year time intervals captured for 322 subjects [9]. They evaluated four different matchers to select the most accurate one, used finally in their evaluations. Similarly to the prior work, the reduction in the recognition accuracy was observed, as the average false non-match rate increased by 27 %, 82 % and 153 % for one-, two- and three-year intervals between compared samples, respectively.

Current literature delivers also a claim that the iris aging – if exists – is of a negligible significance [10]. However, one should be careful as the linear regression models used in this work explain only partially possible sources and nature of matching scores non-stationarity, as they try to find monotonic deterioration in the selected statistical property (the average matching score). The lack of linear trend in one statistics does not guarantee the statistical stationarity, as still the remaining (and important) statistical properties (e.g. score variance) may vary. Moreover, non-monotonic changes of statistical properties may also be a consequence of aging and might be interesting to the biometric community.

3 Aging Database

3.1 Database Summary

BioBase-Aging-Iris database prepared for this work is a part of a larger set – *BioBase*. The BioBase contains biometric samples of five characteristics collected for the same persons, namely: iris, fingerprints, hand geometry and face images, as well as on-line handwritten signatures registered on the graphical tablet. The BioBase was collected mostly in 2003 and 2004 for more than 200 volunteers. We had repeated in 2010 and 2011 the data collection process for all biometric characteristics a few times for 31 individuals, who agreed to participate in the re-enrollment, building five *BioBase-Aging* datasets, separately for each biometric characteristics. Certainly, we had frozen our database collection environment to use exactly the same equipment and the software, configured identically to minimize the influence of environmental factors onto the biology-related aging effects. To capture samples for all five characteristics, a single measurement session lasted approximately 30 min. In particular, during each session, we realized three iris capture attempts, and each attempt consisted of as many presentations as it was necessary to capture two iris images (per eye). We intentionally did not capture the iris images in immediate series, to introduce some between-attempt variability in intra-session samples. This scenario yields to six iris images for each eye obtained in each measurement session.

The iris images were captured for both subject's eyes. To minimize the influence of poor image quality on the aging-related conclusions, we decided to manually remove poor quality samples, e.g. those showing less than 50 % unoccluded iris texture. Hence, the *BioBase-Aging-Iris* contains 571 iris images for 58 different eyes. The shortest time interval between sessions is 30 days while the longest is 2960 days (i.e., more than 8 years). The resolution of the resulting iris images is 768 × 576 pixels, and the image quality highly exceeds minimum requirements suggested by the ISO/IEC 19794-6:2011 and ISO/IEC 29794-6 standards, Fig. 1.

3.2 Equipment Used

Limited availability of commercial cameras offering raw iris images in 2003, when the experiment was initiated, encouraged us to construct a complete hardware

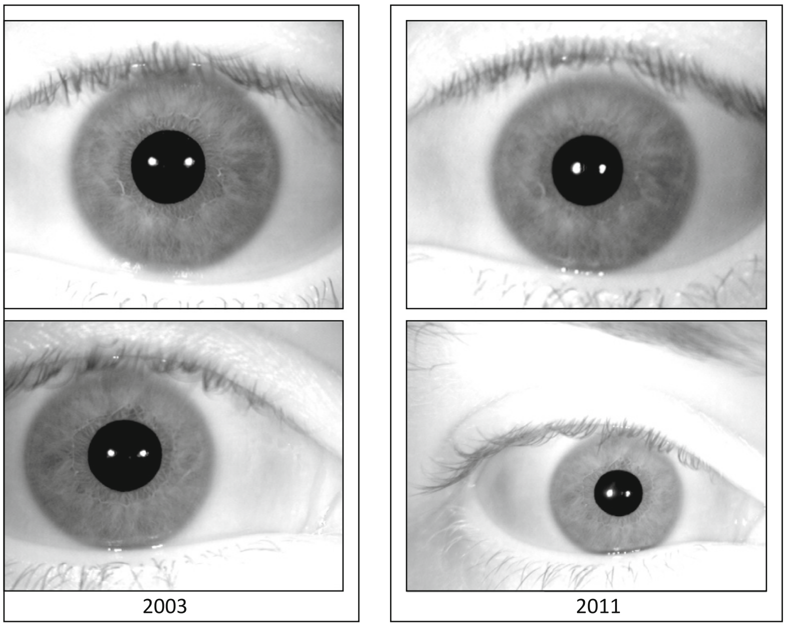

Fig. 1. Example iris images captured in 2003 (left) and the corresponding images of the same eyes captured in 2011 (right). Visual inspection of the upper samples reveal no serious differences within the iris pattern, yet the bottom samples show slight differences in the iris size and in the distribution of illumination. These possible differences may also influence the matching scores, incorrectly obscuring an aging phenomenon resulted from the biology. Images originate from *BioBase-Aging-Iris* database.

setup for the iris capture: the IrisCUBE camera. This prototype equipment captures the iris from a convenient distance of approximately 30 cm with the desired speed and quality. The camera was equipped with optics that actively compensates for small depth-of-field, typical in iris recognition systems, through an automatic focus adjustment. Two illuminants of near infrared light (with maximum power set at 850 nm) are placed horizontally and equidistantly to the lens, what guarantees consistent and sufficient scene illumination. IrisCUBE uses TheImagingSource DMK 4002-IR B/W camera that embeds SONY ICX249AL 1/2" CCD interline sensor with enlarged sensitivity to infrared light. Camera parameters such as shutter speed and gain may be adjusted manually or automatically.

During lifespan of the database collection project, new equipment with iris capture capability emerged on the market, and nowadays we may select among dozen of iris capture devices. However, due to high quality of the images captured by the constructed camera and due to the aim of guarantying homogeneous data collection environment, we used IrisCUBE to capture all the images in *BioBase-Aging-Iris*, thus also in 2010 and 2011 re-captures.

3.3 Database Variants

We observed that iris images captured after a few years might have different iris diameters when compared to these captured at the beginning of this project. Although each recognition methodology should be iris-diameter agnostic and normalize its size prior to feature extraction, we prepare a second variant of the

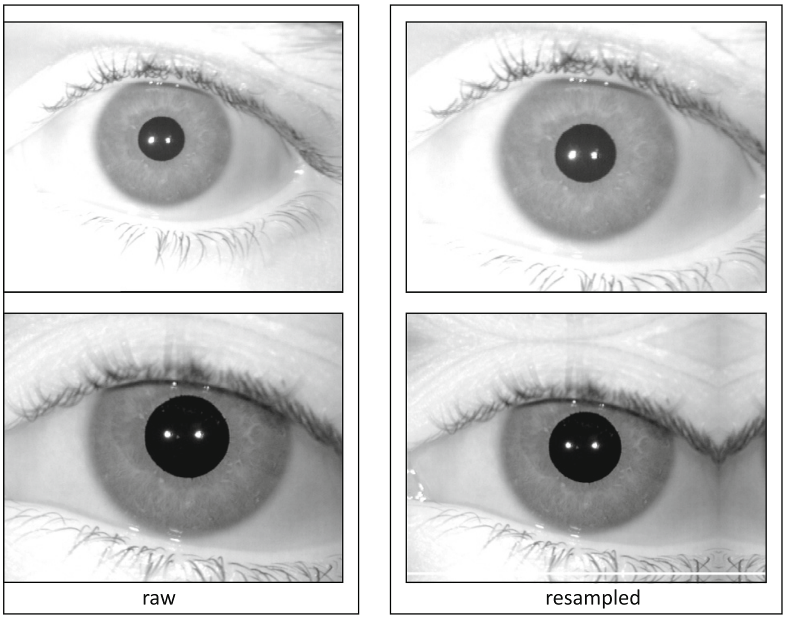

Fig. 2. Examples of raw database images (left) and the corresponding size normalization results (right) after an increase (top) and a decrease (bottom) of the image resolution through bicubic interpolation. We may observe the effects of cropping and filling up with neighboring elements when changing the image resolution. Normalization is performed to center the iris within the image, what should help the matchers in correct data segmentation.

database with iris diameters normalized to the intra-class average using bicubic interpolation.

Images with iris diameter smaller that the intra-class average diameter are enlarged and cropped to the original resolution (768 × 576 pixels). If the iris diameter is larger that the average, the image resolution is decreased and the missing parts at the image borders are filled up with a mirror copy of the neighboring parts, again to keep the original resolution. We use the iris segmentation parameters which were calculated at this stage, and the cropping or filling up the image are realized to to center the iris within the image (Fig. 2). Further in the paper we refer to these two variants as the *raw* and *resampled* versions.

4 Iris Coding Methods

We use four different, commercially or publicly available iris matchers in this work, namely Neurotechnology VeriEye [11], Monro Iris Recognition Library – MIRLIN [12], Open Source for IRIS – OSIRIS [13], as well as the BiomIrisSDK, which is based on the methodology developed by this author [14]. In the following paragraphs we briefly characterize all four methods and provide rationale for their employment.

Neurotechnology VeriEye employs a proprietary and not published iris coding methodology. The manufacturer claims a correct off-axis iris segmentation with the use of active shape modeling, in contrast to typical circular approximation of the iris boundaries. VeriEye was tested for a few standard iris image databases, it was used in the NIST IREX project and presents pretty good accuracy. It is also the only – known to us – commercial matcher available for free (for a month period). The resulting score corresponds to the similarity of samples, i.e. a higher score denotes a better match. For the sake of simplicity, we use further the *NT* acronym for the VeriEye matcher.

MIRLIN derives the iris features from the zero-crossings of the differences between Discrete Cosine Transform (DCT) calculated in rectangular iris image subregions [15]. The coding method yields to binary iris codes, thus the comparison requires to calculate a Hamming Distance (a lower score denotes a smaller distance between samples, i.e. a better match). The *ML* acronym is used for the MIRLIN matcher further in the paper.

BiomIrisSDK employs the Zak-Gabor wavelet packets and the binary iris feature vectors are derived by one-bit coding of the Zak-Gabor transform coefficients' signs. The uniqueness of this method relies on the fact, that it does not employ image filtering, popular in iris recognition, and produces iris features that reveal global character with respect to the iris regions used in coding. As for the MIRLIN matcher, a Hamming Distance is used to calculate the matching score. We use the *ZG* acronym further in the paper when referring to the Zak-Gabor-based matcher.

OSIRIS implements Daugman's iris recognition idea of using 2D Gabor filtering to calculate a binary feature vector (iris code). Pairs of bits code one of the four phase quadrants, thus roughly approximating the signal phase information. A Hamming Distance, normalized by the number of non-occluded code bits, is used to calculate the dissimilarity between irises. A few programming bugs had to be fixed prior to the OSIRIS application, and after these improvements it finally offers a good reliability.

This is noteworthy that we had a full control over the ZG and OS methods. This allows to separate segmentation and coding processes, and in consequence to feed the matchers by images correctly segmented. This possibility is highly beneficial, as the results depend only on the properties of iris pattern fluctuations, and not on segmentation errors. We thus manually inspected the segmentation results achieved by ZG method for all samples in *BioBase-Aging-Iris*, and – if needed – corrected positions of the pupil, iris and sector positions used in ZG to generate the feature vector. For OS matcher, we transformed all iris images to the polar coordinates applying Daugman's rubber sheet model [2], and manually corrected occlusion masks. As the OS matcher employs normalized images expressed in polar coordinate, there is no use to examine its performance for resampled database variant, thus only raw subset was used for this method.

We had however no chance to separate the segmentation and the coding procedures in the commercial matchers (NT and ML), thus this part of the aging assessment encompasses the entire performance of the methods (i.e., eventual segmentation errors and changes in the iris tissue).

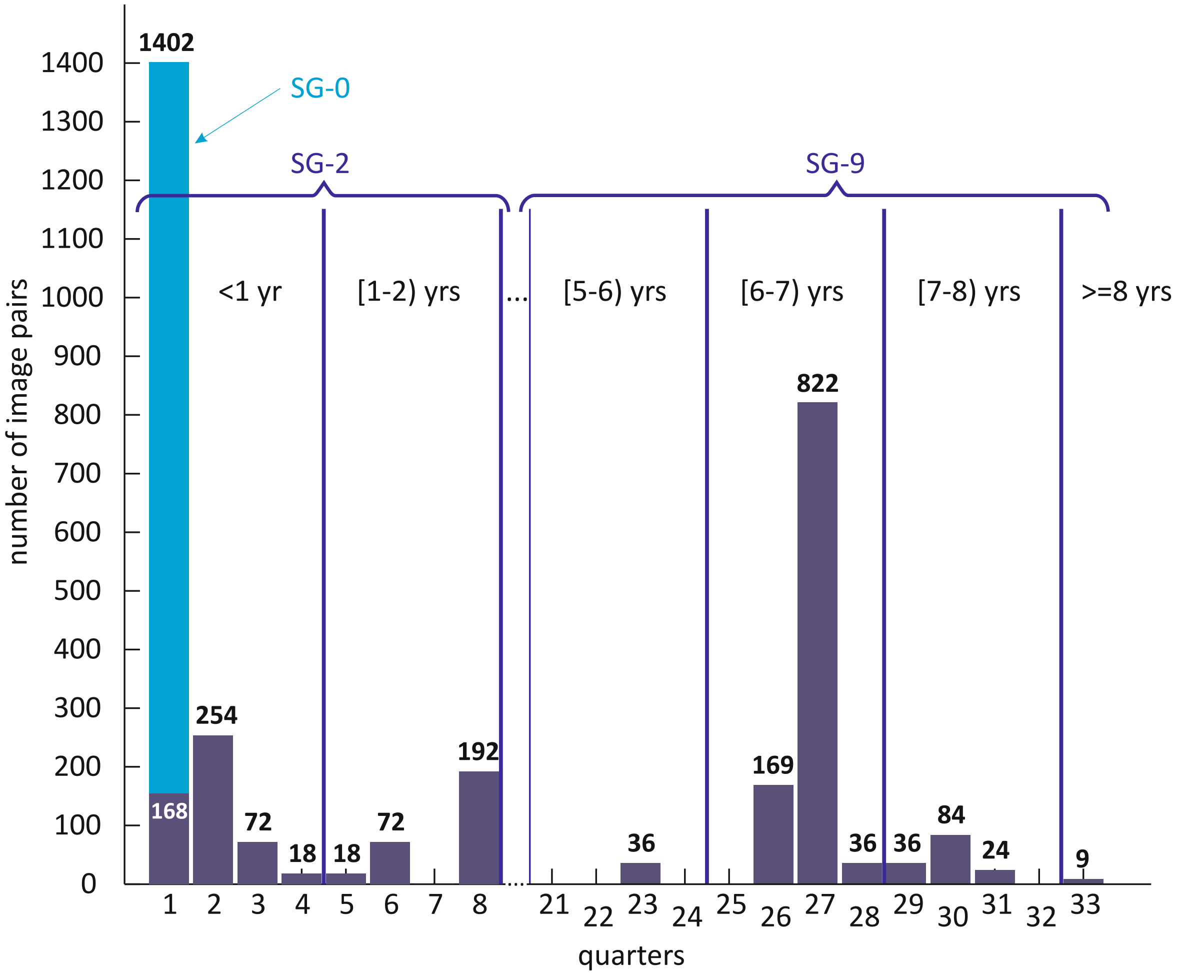

Fig. 3. Numbers of possible genuine pairs that can be constructed with the images of *BioBase-Aging-Iris* database plotted as a function of time lapse between the samples (gathered in quarters, i.e., three-month periods). Different-session pairs are marked with dark blue color, and the light blue color shows the number of pairs including same-session ones. Segmentation of pairs into three groups (SG-0, SG-2 and SG-9) is also shown, where SG-0 contains all the scores for intra-session comparisons, SG-2 groups all the scores generated for time lapse not greater than 2 year (excluding the intra-session scores) and SG-9 gathers all the scores calculated for samples distant by at least 5 years.

5 Results

5.1 Matching Score Generation

We inspected the *BioBase-Aging-Iris* database to construct a distribution of all possible pairs of the same-eye images with respect to the time lapse between image captures, Fig. 3. The number of possible genuine comparisons is equal to twice the number of possible iris image pairs, as the matchers may not return a symmetrical scores (i.e., the score between the iris image A and the iris image B may be unequal to the score between B and A). We may generate 3 244 image pairs in *BioBase-Aging-Iris*, thus the total number of all genuine comparison is

6 488. Among these comparisons, we have 2 468 results of comparing the iris images captured in the same session, and 4 020 scores of matching inter-session images. *BioBase-Aging-Iris* allows to construct 51 654 impostor comparisons for all the time intervals observed during genuine comparison generation.

NT, ZG and OS matchers allow to generate all the above mentioned genuine and impostor scores for both database variants (raw and resampled). The ML matcher generated a smaller number of scores (5 948 and 44 514 of genuine and impostor scores, respectively) due to the template generation errors, yet the numbers of scores for resampled database is slightly greater (6 328 and 49 162 of genuine and impostor scores, respectively), what may mean that normalization of the iris size increases the accuracy of the ML matcher for this database.

5.2 Matching Score Grouping

It was impossible to encourage all volunteers to participate in the experiment on each day we organized the re-capture, thus the number of sample pairs with respect to the time interval is uneven, Fig. 3, yielding highly uneven numbers of comparison scores possible to be generated in short periods. To obtain statistically significant results, we decided to gather comparison scores into three groups that can be identified when observing the distribution of sample pairs shown in Fig. 3. The first score group, denoted by *SG-0*, contains all the genuine and impostor scores for intra-session comparisons. The second – inter-session – subset, denoted by *SG-2*, groups all the scores generated for the samples with time lapse not greater than 2 year, certainly *excluding* the intra-session scores. The third subset, referred to as *SG-9*, gathers all the scores calculated for samples distant by at least 5 years and up to 2960 days, i.e. more than eight years. Table 1 details the numbers of genuine and impostor scores obtained for all the matchers used in this work with respect to all three score groups.

5.3 Evaluation Results

To answer the question related to the existence of aging effects in iris recognition, we present the average genuine and impostor scores calculated for each score group: SG-0, SG-2 and SG-9. Note that the conclusions related to aging should be based solely on inter-session comparisons (i.e. scores classified as SG-2 and SG-9), and the intra-session results are presented only for a completeness to show the potential influence of intra- vs. inter-session captures on the recognition accuracy.

The NT matcher was the most accurate for images in *BioBase-Aging-Iris* as it reached zero EER[2] for samples of raw database variant, and only for one time period a non-zero EER was observed after resampling the data. We may clearly observe a deterioration in the average genuine scores within the SG-2 and SG-9 subsets, Fig. 4 on the left. The genuine scores are approximately 14 % lower when the time lapse between samples starts from 5 years and reaches more than

[2] EER (Equal Error Rate) is the value of FNMR (or FMR) at the operating point of Receiver Operating Curve yielding equal values of FMR and FNMR.

Table 1. Number of genuine ξ_g and impostor ξ_i scores in score groups (SG-0, SG-2 and SG-9) for four iris recognition methods used in this work, namely VeriEye (NT), Zak-Gabor-based (ZG), MIRLIN (ML) and OSIRIS (OS). As the latter matcher (ML) behaves differently for raw and resampled data, numbers for ML are presented separately for these database variants.

	Score group → Coding method ↓	Same session (SG-0)	≤ 2 years (SG-2)	From 5 to 9 years (SG-9)
$\|\|\xi_g\|\|$	NT, ZG &OS for each DB	2468	1588	2432
	ML for raw DB	2292	1548	2108
	ML for re-sampled DB	2394	1588	2346
$\|\|\xi_i\|\|$	NT, ZG &OS for each DB	7988	10 186	33 480
	ML for raw DB	7188	9362	27 964
	ML for re-sampled DB	7690	9646	31 826

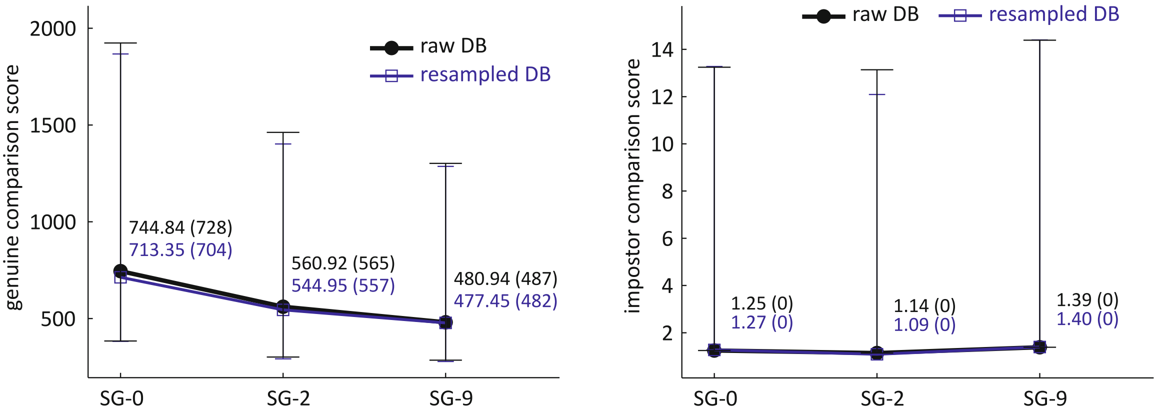

Fig. 4. Average and median scores (in brackets) for raw (circles, black color) and resampled (rectangles, blue color) database variants, shown with respect to the score groups for NT matcher. The whiskers show the 95 % boundaries of the sample distributions in each combination of the SG and database variant. Result for genuine and impostor scores are shown on the left and right, respectively. A higher score denotes a better match.

8 years, and this observation is supported by the outcome of one-way unbalanced analysis of variance (ANOVA). Namely, we cannot accept the null hypothesis that all samples in SG-2 and SG-9 subsets are drawn from populations with the same mean, as they obtained p-value is near to zero ($p < 10^{-47}$ for raw database

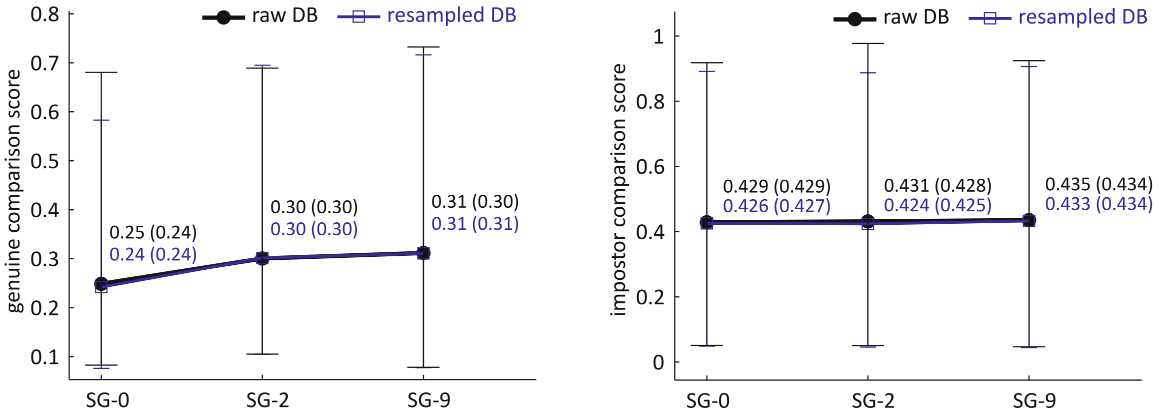

Fig. 5. Same as in Fig. 4 for the ZG matcher. A lower score denotes a better match.

variant and $p < 10^{-37}$ for resampled database variant). When comparing the intra- vs. inter-session scores, we encounter even higher accuracy deterioration, namely 25 % decrease of average score in SG-2 when compared to the SG-0 average, and a decrease by 35 % of SG-9 scores when compared to the SG-0 average. Certainly, also in these cases the analysis of variance casts doubt on the null hypothesis, as the obtained p-values are below machine accuracy when compared SG-0 vs. SG-2 and SG-0 vs. SG-9 scores for both variants of the database. Note that resampling of the iris images has a little influence on the average genuine scores, what may suggest that NT matcher is iris-size agnostic and the aging of the NT templates seems to occur independently of this factor.

Analogously to the presentation of the NT matcher results, we show the average genuine scores for ZG method, Fig. 5 on the left. The intra- vs. inter-session scores show 16 % and 19 % increase in the average Hamming Distance for raw datasets (20 % and 22 % increase for resampled datasets) when compared SG-0 average score with the SG-2 and SG-9 averages, respectively. These changes are statistically significant, as obtained p-values are below machine accuracy (for all combinations of a database variant and a time period). Comparing SG-2 and SG-9 scores show only 3 % of the average score increase, yet still $p < 10^{-8}$ that suggests rejecting of the null hypothesis on equal means. We may observe that the ZG matcher is robust to the absolute iris size, as the genuine scores for raw and resampled database variants do not differ significantly.

The ML matcher average genuine scores are presented in Fig. 6 on the left. As for NT and ZG methods, we may encounter statistically significant differences when comparing average intra- vs. inter-session comparison scores (p-value not exceeding 10^{-10} for all combinations of SG and a database variant), namely the decrease reaches 27 % and even 45 % when compared SG-0 vs. SG-2 and SG-0 vs. SG-9 average scores, respectively. However, when compared the average scores between SG-2 and SG-9 we obtain a low p-value ($p < 10^{-15}$) only for the raw database variant, and $p = 0.19$ for the resampled data, although the increase of the average score in the latter case reaches 4 %. This may suggest, that the aging effect related to the ML templates is somehow compensated by the unifying of the iris diameters inside the iris classes (but in a sense of statistical significance

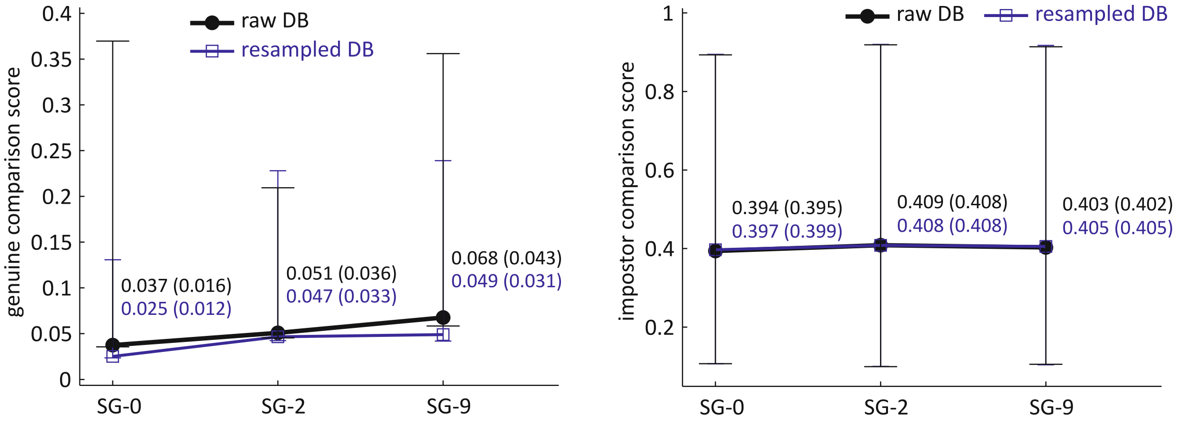

Fig. 6. Same as in Fig. 4 for the ML matcher. As for ZG matcher, a lower score denotes a better match.

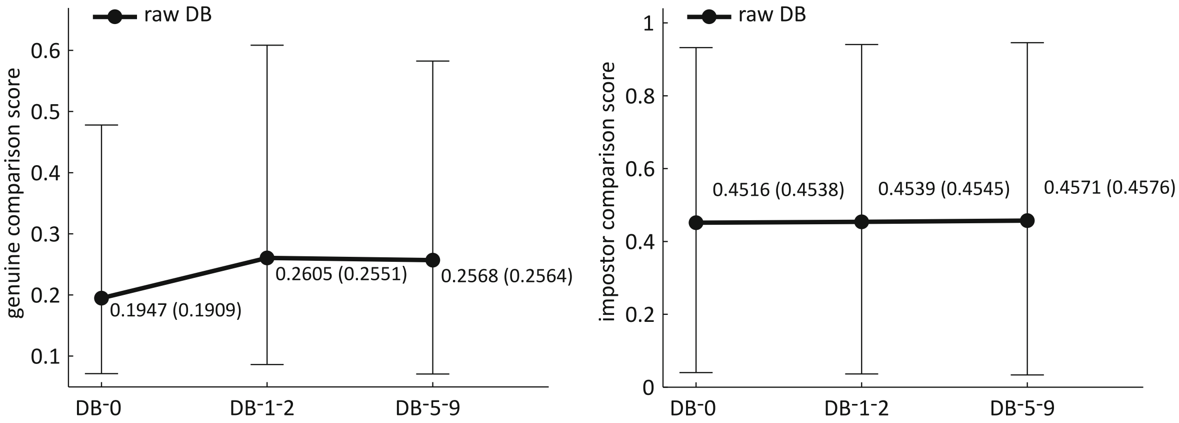

Fig. 7. Same as in Fig. 4 for the OS matcher. As for ZG and ML matchers, a lower score denotes a better match.

of the ANOVA test). One, however, should note that resampling allowed for a better accuracy of the ML matcher, manifesting by a greater number of correctly calculated templates and a lower average of the genuine scores in each score subgroup.

As for all the matchers presented above, also the OS matcher presents statistically significant differences in average genuine scores when comparing intra- and inter-session results (p-value near zero). The genuine average scores increases by about 25 % in both cases, i.e. SG-0 vs. SG-2 and SG-0 vs. SG-9). However, unexpectedly the OS method seems to be resistant to aging effects as the average of inter-session genuine scores (i.e. SG-2 vs. SG-9) slightly decreases by 1.5 % (opposite to the expectations), yet the median slightly increases by 0.5 % (suggesting aging effect), Fig. 7 on the left. These changes are statistically significant as $p < 0.0064$. To investigate this result in more details, we plot cumulative distributions of the genuine scores, Fig. 8. We can see, that the OS matcher behaves differently in regions of higher and lower similarity of iris images. Namely, the stronger is the comparison score (higher similarity between images), the more

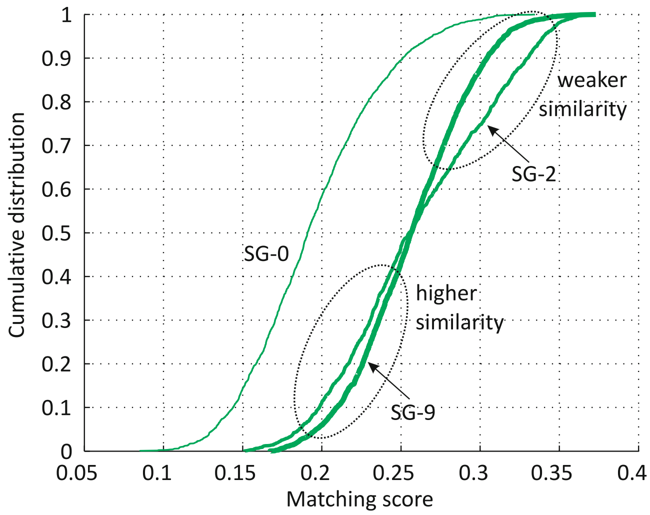

Fig. 8. Cumulative distribution of genuine scores for OS matcher. Regions of weaker and higher similarities between images are intentionally shown to highlight different behavior related to the sensitivity of the OS matcher to aging effects.

aging effect is visible. However, if the images are not matched well (image similarity is lower), then the order of average comparison score is opposite to the expectation: average SG-2 result is greater than average SG-9 score. This suggests that more accurate matching present the aging effect to the more extent, and supports the observation related to the most accurate (on this DB) NT matcher, presenting the highest sensitivity to the time lapse.

As of now we discussed the genuine scores only, and now we turn toward the matching between impostor samples. Figures 4, 5, 6 and 7 on the right present the average impostor scores for NT, ZG and ML matchers, respectively. As it was already suggested in the literature related to the iris aging effect, we may observe lower differences in average scores in groups when compared to the result of genuine comparisons. Namely, the changes range from 0.5 % for ZG and OS matchers (intra- vs. inter-session averages for raw samples) to 22 % for NT matcher (inter-session averages for resampled images). The ANOVA test casts doubt on accepting the hypothesis that investigated sample subsets are drawn from populations with the same mean ($p < 0.021$ for all combinations of SG and database variant). We may also see that iris absolute diameter size has no influence on the impostor scores.

6 Discussion

Average values of the genuine scores obtained for all the tested matchers may suggest that iris templates age, what partially supports earlier findings determined for different matchers and different databases, yet collecting samples with a shorter time lapse between captures than in this work. The extent to which the template aging phenomenon is observed is however uneven across different matchers, in particular we observe a higher influence of the time flow for more accurate methods. This observation may be explained in two ways. On the one hand, less accurate matcher may encounter a higher number of segmentation errors, or improper iris image mapping, when comparing irises with different pupil dilation, iris diameter or occlusion extent. These factors may be classified as ‘aging factors’ (as the resulting template ‘ages’, independently of the ‘aging’ source), yet we rather would like to answer the question if the aging relates also (or first of all) to the iris texture, i.e. a direct donor of the biometric features. If the answer is affirmative, then assessment to what extent it tackles the elements of the complicated iris tissue would be of a great value. So, on the other hand, we may assume that high accuracy of the matcher relates to a higher accuracy of the segmentation process. If so, more comparison scores result from an appropriate matching of the iris patterns (with occlusions appropriately removed and iris texture appropriately mapped), which – according to the experimental results – exhibits significantly different nature after a few year time lapse. Note that the aim of each coding method is to be sensitive for iris features which guarantee individualization of subjects (e.g., frequency bands in wavelet-based coding routines). Collecting these thoughts, we would hazard a guess that the iris aging relates also to the iris characteristics that are responsible for our individual biometric features, i.e. the iris pattern. Simultaneously, we stress again that difference in average comparison scores is only one indicator of the inter-session variability, suggesting the non-stationarity in iris recognition.

The fact of iris aging, if finally confirmed by a series of additional experiments exploiting a large number of matchers and big, heterogeneous datasets, should under no circumstances devalue the strength of the iris recognition. Next research step should be focused on the assessment of the extent to which the aging phenomenon deteriorates the accuracy of this modality, allowing for introducing precise rules of template usage, in particular adequate time periods which call for re-enrollment, what may only increase an accuracy of this prominent and very accurate authentication technology.

Acknowledgements. The author would like to thank Mr. Mateusz Trokielewicz for initial validation of the segmentation results calculated for the *BioBase-Aging-Iris* samples. The author is also cordially indebted Dr. Joanna Putz-Leszczyska, Dr. Lukasz Stasiak, Mr. Marcin Chochowski and Mr. Rafał Brize, with whom he had collectively built the *BioBase-Aging-Iris* database. The author also appreciates Mr. Marcin Chochowski for necessary bugs fixing in the original OSIRIS matcher, and making a new, corrected version available for this work. This work was partially funded by The National Centre for Research and Development (grant No. OR0B002701: “Biometrics and PKI techniques for modern identity documents and protection of information systems – BIOPKI”).

References

1. Flom, L., Safir, A.: Iris recognition system. US Patent 4,641,349, February 1987
2. Daugman, J.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1148–1161 (1993)

3. Tome-Gonzalez, P., Alonso-Fernandez, F., Ortega-Garcia, J.: On the effects of time variability in iris recognition. In: IEEE Conference on Biometrics: Theory, Applications and Systems, pp. 1–6. IEEE (2008)
4. Rankin, D., Scotney, B., Morrow, P., Pierscionek, B.: Iris recognition failure over time: the effects of texture. Pattern Recognit. **45**, 145–150 (2012)
5. Baker, S., Bowyer, K.W., Flynn, P.J.: Empirical evidence for correct iris match score degradation with increased time lapse between gallery and probe images. In: International Conference on Biometrics, pp. 1170–1179 (2009)
6. Bowyer, K.W., Baker, S.E., Hentz, A., Hollingsworth, K., Peters, T., Flynn, P.J.: Factors that degrade the match distribution in iris biometrics. Identity Inf. Soc. **2**(3), 327–343 (2009)
7. Baker, S., Bowyer, K.W., Flynn, P.J., Phillips, P.J.: Template aging in iris biometrics. In: Burge, M., Bowyer, K.W. (eds.) Handbook of Iris Recognition. Advances in Computer Vision and Pattern Recognition, pp. 205–218. Springer, London (2013)
8. Fenker, S.P., Bowyer, K.W.: Experimental evidence of a template aging effect in iris biometrics. In: IEEE Computer Society Workshop on Applications of Computer Vision, pp. 232–239 (2011)
9. Fenker, S.P., Bowyer, K.W.: Analysis of template aging in iris biometrics. In: CVPR Biometrics Workshop, pp. 1–7 (2012)
10. Shchegrova, S.: Analysis of iris stability over time using statistical regression modeling. In: Biometric Consortium Conference & Technology Expo, September 18–20, 2012, Tampa, Florida, USA (2012)
11. Neurotechnology: VeriEye SDK, version 4.3, revision 87298, July 2012
12. SmartSensors: MIRLIN SDK, version 2.23.2, August 2012
13. BioSecure: OSIRIS, version 2.01 (2009)
14. Czajka, A., Pacut, A.: Iris recognition system based on Zak-Gabor wavelet packets. J. Telecommun. Inf. Technol. **4**, 10–18 (2010)
15. Monro, D.M., Rakshit, S., Zhang, D.: DCT-based iris recognition. IEEE Trans. Pattern Anal. Mach. Intell. - Special Issue on Biometrics: Progress and Directions **29**(4), 586–595 (2007)