The Impact of (Segmentation) Quality on Long vs. Short-Timespan Assessments in Iris Recognition Performance

Peter Wild[1], James Ferryman[1], and Andreas Uhl[2]

[1]University of Reading, Computational Vision Group, School of Systems Engineering, Reading RG6 6AY, United Kingdom. Email: {p.wild, j.m.ferryman}@reading.ac.uk
[2]University of Salzburg, Multimedia Signal Processing and Security Lab, Department of Computer Sciences, Salzburg, 5020, Austria. Email: uhl@cosy.sbg.ac.at

July 6, 2015

Abstract

Several researchers have presented studies of temporal effects on iris recognition accuracy, with varying results on severity of observed effects. The sensitive topic continues to be adversely discussed and the difficulty of isolating performance-impacting factors is immanent. The impact of ageing on segmentation versus feature extraction has been largely neglected so far. This paper attempts to shed light on the impact of segmentation quality on observed temporal effects highlighting the critical role of the segmentation module and quality assessment when assessing ageing effects. The lack of large and standardized temporal variation in public datasets as well as additional metadata (age of subject, recording parameters) and strictly enforced unified recording and quality guidelines over time-separated sessions are identified as imminent problems of ageing studies. Results are reported on a long-timespan database of 36,240 images comprising 104 classes and a 4-year time lapse, offering a large variety of recording conditions highlighting the critical role of a transparent recording setup.

1 Introduction

The iris as an epigenetic (not entirely genetically determined) biometric modality is known for its high universality, distinctiveness, and performance attributes [1]. Apart from being reported to achieve remarkably flat receiver operating characteristics (ROC), the human iris is also claimed to be stable over time, except for pigmentation change [2]. Recently, several investigations have been launched to further verify claimed stability based on new datasets with long-timespan data [3, 4, 5, 6] investigating ageing effects.

Biometric ageing refers to *permanence* of biometric features [7], questioning manifold properties: the temporal stability of features, impact of time-spans on recognition accuracy, and effect of subject's age on error rates and recording conditions. It is important to differentiate between temporal impact on biometric features and potentially observable degraded recognition accuracy. Related work sensibly distinguishes between physical ageing (biometric data change over time) and template ageing (changes between enrollment and verification/identification) [8]. Whereas a change in features is likely to imply a degradation of recognition accuracy, observed decreased recognition accuracy is not necessarily a result of modified features. As an example [9, 4, 5] identify the subject's ageing to impact on user-specific average pupil dilation, which is reported to decrease over time. While smaller available iris area is likely to be a dominant factor for recognition accuracy, for ageing studies many more impacting variables on recognition accuracy can be identified, which need to be taken into consideration when designing ageing experiments. To give examples, physiological changes in shape of cornea and rigidity of the lens [10], as well as changes in physical behaviour [11] can impact, especially on segmentation. Therefore, in existing ageing studies, there is high risk of inconsistent conclusions. The contribution of this paper is a systematic evaluation of adverse segmentation and quality variability effects on recognition accuracy for long-timespan versus short-timespan iris comparisons. It illustrates the feasibility to observe positive and negative longitudinal effects by just varying and controlling the segmentation quality environment. The aim of this paper is to contribute a set of precondition predicates impacting on ageing studies, rather than looking for a simple answer to the question of existence of irreversible ageing effects. It highlights the need for controlling these measures in ageing studies towards understanding the inter-relation of quality and time-span on observable effects. Knowing the contribution to aging of (quality) factors (as in [12]) that can be controlled can help isolating factors, yet their interaction is often complex and any not included quality impacting factor is likely to attribute to observed "ageing" effects. Thereby the study gives further evidence of potential reasons for inconclusive large deviations in reported ageing studies.

The remainder of this paper is organized as follows: Section 2 reviews related work highlighting main reported iris ageing results and available databases as well as research work assessing the impact of quality on recognition. The employed reference system under test and recorded ageing dataset are presented in Section 3. Experimental results are outlined and discussed in Section 4, introducing the tested quality predicates impacting on observed temporal change regarding recognition accuracy. Section 5 forms the conclusion.

2 Related Work

With the availability of age as meta-information for biometric records and release of long-timespan iris databases, several researchers have investigated ageing issues, see [4, 13] for recent surveys of approaches.

Focusing on the time-span aspect, Tome-Gonzales et al. [14] are among the first iris reports to survey ageing effects, revealing an impact of time separation on iris recognition accuracy for timespans of several weeks using using BioSecure iris data (254 subjects, 8128 images). They also report the largest effect considering the session-methodology employed in this paper. Baker et al [3] confirm an average increase of 0.018 HD value per person comparing 4-year with 1-year time lapse data. They identify a larger difference in the pupil dilation for long-term comparisons using the NotreDame iris ageing dataset (13 subjects, 1809 images, taken in 2004 and 2008). Also for 2-years vs. half-year [15] and 5-year vs. 2-year time lapse [16] statistically significant increases of genuine scores have been reported. Fenker et al. [5] confirm a shift of the genuine score distribution towards the impostor score distribution. All these reports are referring to the agreed definition of ageing in ISO/IEC 19795-1:2006 as "increase in error rates caused by time-related changes in the biometric pattern, its presentation and the sensor".

Focusing on the age-group aspect, Fairhurst and Erbilek [9] give evidence on an improvement of iris recognition accuracy with age (1.4% EER for age <25 vs. 1.1% EER age 25-60 and zero EER for age >60). However, they report a negative impact on simple Hough transform-based iris segmentation algorithms without optimizations (binary masks), highlighting the critical role of preprocessing. As reason for the observed behaviour decreased pupillary dilation with age is identified (which is confirmed in [4, 5]). It has also been claimed recently, that age information can be deduced from iris texture data [17].

Focusing on the feature stability aspect, in contrast to reported decreased accuracy due to ageing effects, the IREX VI report [4] comes with several conclusions underlining the thesis that there is no evidence of significant "ageing" effects in iris recognition. First, the observed average increase of 0.003 HD per person in the employed OPS-XING dataset due to "ageing" is an order of magnitude lower than other influencing factors, like camera change or user familiarity with the system. Second, large-scale (622,464 subjects) genuine distributions are found to be stable over a period of 6 years. Third, they mention a concentration of errors and varied pupil dilation in specific subsets of the ND-Iris-Template-Aging dataset to be likely the reasons for reported degradation in their 2004-2008 and 2008-2010 datasets. Finally, the study comes with suggestions to let individual-specific measurements and mixed-effect regression models guide ageing estimation - emphasizing the need for tight acquisition control and mitigation of other temporal effects. Therefore, they propose a classification of ageing effects into systematic (e.g. insufficient illumination), subject-specific behaviour-remediable (e.g. squinting eyes), subject-specific design-irremediable (e.g. motion symptoms of Parkinson's disease), and subject-specific information-source-irremediable (changes in the texture) understanding iris ageing as "irreversible changes to the healthy iris or neighboring anatomy that yield mated dissimilarity scores that increase monotonically with time-separation of the compared images" explicitly excluding the effect of natural dilation. Recently, Bowyer et al. [6] presented a report claiming "methodological errors" in IREX VI: The removal of all matches with HD values exceeding 0.27 suggested that a truncated regression should have been done. Further, it is argued that the interpretation of the regression model did not take correlation among independent variables (e.g., dilation as a function of age) into account, interpreting one coefficient in the predictive model isolated from the others. Finally, merging of recording sessions, unavailability of the raw image dataset for reproducibility and a non-ISO definition of ageing are identified as problematic factors, therefore likely leading to biased results. When comparing ageing studies, differences in experimental configurations (e.g., Grother et al. [18] recently stated the specific operational setup of the employed Nexus dataset used token-less "1-to-first" comparison) should be considered. Related to this observation also criticism with regards to lacking time-analysis of all first, second and third matching attempts is outlined in [6]. The potential introduction of bias into ageing experiments by filtering data based on predicates is subject to this paper.

Given that the literature has no agreed view on the significance of ageing effects in iris recognition, it seems reasonable to further investigate reasons for observed deviations in results. Mehrothra et al. [13] investigate ageing on ND-Iris-Template-Aging-2008-2010 and ND-TimeLapseIris-2012 databases and find, that the reported increased false rejection rate may apart from pupil dilation also be attributed to other factors, such as occlusion, blur, rotation, illumination and noise. Also Ortiz et al. [19] find, that obviously, a high difference in pupillary size and low absolute pupillary values create low HD scores. The experiments in this paper augment existing studies observing the recognition behaviour of long-timespan vs. short-timespan comparisons under elimination of specific variation conditions. Introducing a new dataset with aged iris images focusing on image quality, in addition to considerations in [4], this work tries to highlight segmentation-related quality measures and further illustrates the feasibility to observe positive and negative effects when filtering short and long-timespan data based on quality-related predicates. Recently, Trokielewicz [20] examined the predictability of changes in Hamming distance in relation to time by testing 29 regression models on 6 predictors, including time and geometrical parameters, and come up with different regression models for each of the tested iris SDKs. While such prediction models are a means to target expected intra-personal variability, the authors raise that miscellaneous factors were not taken into account. While the present work can not provide an exhaustive evaluation of factors either, the added value of our study is an investigation of even more (10) parameters, and identifies the possibility to observe positive and negative impacts on accuracy under different quality preconditions. While previous work has identified the impact of quality on recognition behaviour, our experiments aim to investigate whether in presence of such variation, indeed different "ageing" effects (as per the open ISO/IEC 19795-1:2006 definition) can be observed.

3 Experimental Setup

Let $k \in \{G, I\}$ denote the class label (*genuine, imposter*), y be the score and q be a (parameterised) quality predicate indicating if a comparison pair $\langle g, s\rangle$ fulfills a criterion. As an example, $q(\langle g, s\rangle, \alpha) \in 0, 1$ could indicate, if both gallery and sample templates g, s exhibit a pupil radius less than α. In Bayesian decision theory, $P(k = G|y)$ in presence of quality pre-estimation can be extended to $P(k = G|y, q)$ calculated via density $p(y, q|G)$ as follows [21]:

$$P(G|y,q) = \frac{p(y|G,q)P(G)}{\sum\limits_{k\in\{G,I\}} p(y|k,q)P(k)}. \tag{1}$$

It is clear that depending on a suitable choice of quality predicates $p(y|G, q_1)$, $p(y|G, q_2)$ can be rather different. However, this paper highlights by experimental analysis, that for combined quality-timespan predicates $q \wedge t$ comparing short-timespan t_S and long-timespan t_L data, $\exists q_1, q_2$ such that for a system error estimator e (we use equal error rate) $e(t_S \wedge q_1) > e(t_L \wedge q_1)$ and $e(t_S \wedge q_2) < e(t_L \wedge q_2)$. Implications of this observation would be the dedicated need to make any influencing quality predicates q visible, while questioning the universality of claimed effects.

3.1 Test System

In order to test the impact of quality predicates on observed iris ageing effects, we employ a publicly available iris recognition system with exchangeable segmentation and feature extraction modules: USIT [22]. Performance reports on several reference datasets and detailled discussion of the employed techniques are available in [22].

The following two different feature extraction techniques are employed to extract stable binary features, which are compared using fractional Hamming Distance as comparator taking noise masks into account:

- *Quality Assessment and Selection of Spatial Wavelets [24] (QSW)*: The min-max wavelet-based feature extraction technique creates a feature vector of alternating binary sequences of 0s and 1s whenever minima and maxima are obtained in the 10x2 concatenated 1-D signals from 512x5-sized texture stripes in wavelet domain yielding 10,240 bits.

- *Log-Gabor (LG)*: The second approach is a reimplementation of Daugman's iris code feature extraction using Log-Gabor filters on the same ten 1-D signals quantising the complex phase angle responses, based on a reimplementation of [25]. The result is again a 10,240 bits feature vector.

Two segmentation algorithms are used for localising iris boundaries and noise masks mapping the iris into its normalised Faberge coordinates [2]:

- *Contrast-adjusted Hough Transform (CAHT)* [22] uses Hough transform after contrast adjustment enhancing pupillary and limbic boundaries, to robustly detect iris boundaries in the Canny edge detection result. It is used as classical reference technique.

- *Weighted Adaptive Hough and Ellipsopolar Transforms [23] (WAHET)* is a representative of two-stage iris segmentation algorithms decoupling coarse centre detection (using a weighted adaptive Hough transform), and fine boundary localisation in the polar image (using an ellipsopolar representation to assist detection of the second, less pronounced boundary using the first detected boundary via simple windowed-search in polar domain).

3.2 Test Database

The data set used in this work contains iris texture images acquired with the *Irisguard H100 IRT* sensor. Note, that the IrisGuard family of sensors use visible light for controlling the observed pupil dilation, see UK Patent 2495328, which is likely to even minimise ageing impact on pupil dilation. The images are divided into 2 data sets, the first one containing images from 52 (29 male and 23 female) subjects, aged 21 to 45 (as of 2009) and acquired in 2009 (480 - 1561 images per subject), the second one containing images from exactly the same subjects acquired in 2013 (40 images per subject), resulting in a time gap of four years between the acquisition of the two image sets. These data sets[1] are subsequently denoted as *H100-2009* and *H100-2013*. *H100-2009* is a subset of the CASIA cross sensor iris database [26]. There are 6 different recoding sessions in 2009 with quite varying underlying image characteristics regarding pupillary dilation, occlusion, presence of glasses, and blur, Table 2 lists basic information with further characteristics on occurrence of closed eyes, glasses, over-/underexposure and the

[1] the employed dataset is to be released as part of the upcoming CASIA V5.0 Iris Dataset at `http://biometrics.idealtest.org/`

obtained frequency of consistent normalisation results for both tested (WAHET and CAHT) segmentation algorithms, thereby indicating the segmentation difficulty. Further information given covers the gender balance listing the ratio of male and female samples, as well as the average age (with standard deviation) of images. As sessions 0 and 3 contained a single user only, they have been merged with sessions 2 and 5 of closest characteristics, respectively. Whereas the first 2 sessions (1 and 0+2) are free of any glasses, all images in sessions 4 to 6 contain glasses and therefore also significant reflections. Note, that these different grades in quality are by intention to see their impact and potential shadowing of observable time separation effects. Images recorded in 2013 follow a rather controlled setup (highest segmentation consistency, lowest average pupillary dilation and standard deviation, very low frequency of glasses compared to other sessions), indicating the presence of a strong light source with setup with particularly large available iris area. Figure 1 presents samples illustrating the type of variability across sessions, Figure 2 illustrates age distribution in the dataset. Further properties of individual sessions with regards to quality predicates are discussed in the following subsection. A total of 36,240 images have been employed in experiments, randomly drawing 10K pairs for each class for each of the tested quality setups.

4 Experimental Results

In experiments we target a range of questions related to the impact of segmentation quality on ageing assessments in iris recognition, which are covered in the following subsections.

4.1 On the Impact of Quality Predicates on Accuracy

Table 1 lists employed quality assurance parameters, which are grouped into setup-related (α), segmentation-consistency (β to ϵ), and image-quality measures (ϕ to ω). There is a predicate associated with each parameter, which is evaluated for each image and only if it evaluates to true, the corresponding image is included in evaluations with regards to a particular parameter.

Any precondition based on filtering of data (as claimed to be done in some ageing studies, see Sct. 2) might bias observed temporal effects, especially if intra-personal variation is larger than impact due to subject ageing. In a first experiment, we therefore investigated the impact of each of the image-quality measurements in Tab. 1 on recognition accuracy (rates refer to WAHET segmentation using LG as feature). It is important to note that measures ϕ to ω themselves depend on the accuracy of the employed segmentation algorithm. While ground truth would be the ideal reference, unfortunately such information is usually not provided - especially for ageing databases, see [27]. In this study we employ two different segmentation algorithms (one based on circular HT, another using iterative centre search) on the used dataset and restrict evaluations to images where both algorithms coarsely agree on the segmented region (we employ the predicate $\alpha = 5 \wedge \beta = 10 \wedge \gamma = 10 \wedge \delta = 10 \wedge \epsilon = 10$).

Figure 3 lists box plots of quality score values (left column) for each of the different recording sessions. For the prominent quality parameter 'pupil dilation' (ϕ) primarily determining the size of the available iris texture, there is huge variability between recording sessions in 2009. From the graphs it is evident that session 4 is closest in configuration to the session recorded in 2013. When looking at intra-2009 comparisons, we can see, that this parameter has a clear impact on recognition accuracy (4.55% vs. 3.78% EER) despite employed rubbersheet normalisation, when splitting up the dataset in two sets based on the mean ($\phi = 35.74$). While pupil dilation might be affected by the subjects increasing age, (e.g., in [9] an effect on pupil dilation is observed), other influencing factors were too pronounced and recording timestamps not supplied with the employed database to verify such a claim.

For the parameter 'occluded iris area' (ψ) being measured as the absolute difference between the area occupied between pupillary and limbic boundaries, and the netto area masking out eyelids and reflections, sessions 0+2 and 4 are closest to the 2013 setup. Note, that there are still many outliers present accounting for consistent detections of 'squinting eyes' which might affect detections in a quite drastic way. However, the quality parameter follows a similar distribution in all tested sessions. Indeed, this parameter turned out to have an immersive impact on recognition accuracy (5.31% vs. 3.71% EER), tested using split point $\psi = 4600$.

The parameters 'pupillary contrast' and 'limbic contrast' (μ, ν) were introduced to judge for the difficulty of segmentation, assessing the contrast of the reported segmentation boundary (presumably, if the contrast is low it is more difficult to assess the pupillary or limbic boundary correctly and due to over- or undersegmentation alignment problems might arise, which can not easily be accounted at the matching stage [23]). As an indicator of boundary contrast the absolute difference in average intensity of the circular window (5 pixels height) outside and inside the boundary were assessed. The 2009 Sessions 1,4 and 3+5 all exhibit similar distributions of these parameters with 2013 which indicate the bias on ageing assessments should be limited (however, its overall impact is comparable to pupil dilation, causing EER to be reduced from 6.96 to 2.71% for μ and 4.93 to 2.74 % for ν using split values $\mu = 7.46, \nu = 1.33$). Compared to the little differences between sessions this has a drastic impact on accuracy raising the question how segmentation algorithms affect ageing studies, as these algorithms typically look for boundaries maximising local boundary contour energy functions [23].

Finally, a challenge of on-the-move iris recognition is finding the right balance between illumination, aperture, gain and f-stop to achieve the right depth of field for iris images. We computed the energy of high-frequency content (using a Sobel filter sized k=3 and estimating the total edge energy in the image) in the image to assess in-focus capture. It is interesting to see, that sessions 1 and 0+2 versus 4 and 3+5 exhibit quite different in-focus characteristics. While one would assume that in-focus has a drastic impact on recognition, the degradation in performance is merely visible, especially not over the entire operational range (3.19 vs. 4.72% EER). This result is certainly also attributed to data capture quality assurance checking the acquired image and in line with previous research on iris image compression, showing that low frequency information in iris images has enormous discrimination ability. Even more, this raises the question on whether any observable ageing effects might affect low-frequency or high-frequency content, or both.

Eventually, testing different quality measures we can confirm claims in [13] that quality plays a critical role in accuracy. Even the age of the person may be an impacting quality attribute itself, so we tested this impact splitting up the dataset by its average value, 28.13 years. Interestingly, this selection turned out to have a clear impact on performance, with the older 29-45 years age group exhibiting clearly better EER (2.06% for LG) than the younger age group 21-28 years (5.03%). This large difference can certainly be also attributed to the larger inclusion of more challenging images from sessions 4 and 3+5 for the younger age group (with lower average age, see Tab. 2). However, even restricting evaluations to "easier" sets 1 and 0+2 (4.76% vs. 1.56%, see also Figure 2) and another recognition algorithm (4.35% vs. 2.03% for QSW) yielded similar results, confirming observations in [9]. Yet, we believe it is important to mention, that the density of the dataset in terms of subjects is not very high and should be considered in the interpretation of results.

4.2 On the Impact of Quality Checks on Ageing Effects

Whereas previous research has mainly focused to either show the presence of template ageing or illustrate its minor impact, investigations in this section try to focus on making the impacting conditions clearer and illustrate how they can affect the outcome of ageing observations when naively assuming the presence of just a single impacting quality parameter (the dataset is still too small to normalize all of the quality parameters at the same time).

Concentrating again on WAHET-based segmentation and LG as feature we investigate the impact of ageing using each a single image-quality based measure as filtering predicate and compare high-quality with low-quality results. See Figure 4 (right column) for ROCs using different quality predicates on intra-year versus 4-year comparisons.

An unexpected result is obtained for the first tested quality parameter, pupil dilation. For the high-quality setup (low pupil dilation), the 4-years separated (2009 versus 2013) evaluation returned higher (1.6% vs. 3.78% EER) instead of lower accuracy compared to the intra-year (2009) iris pair evaluation. While this result at first sight raises questions on the presence of systematic errors in the evaluation, it becomes clear when interpreting it from a quality perspective: when filtering the 2009 dataset for low pupils, images from sessions other than 1 and 4 are mostly rejected. As session 4 contains more challenging images (glasses with reflections, see low focus values of Sct. 4 and 5) and since they can be selected for either reference or gallery in 1-year comparisons, but play a minor role in 4-year comparisons, this is biasing the result. Consequently, for the low quality setup (high pupil dilation) the effect is no longer present (e.g. comparisons between sessions 0 and 2 possible). Overall, this illustrates the critical selection of preconditions and is a very good example of how generalisation for ageing is very difficult with datasets captured under varying recording conditions. This is especially an issue for datasets where emphasis on short-timespan data has likely been put on intra-session effects and the 2013 session consists of a standard recording setup to augment long-timespan data.

Interestingly, for all the other quality measures, the long-timespan comparisons overall yielded inferior results, yet the impact was not always pronounced. For iris occlusion within the EER range, long vs. short timespan comparisons' difference turned out to be almost as pronounced as the impact of the quality parameter (6.28 vs. 3.71% EER and 6.06 vs. 5.31% EER for high and low-quality setups, respectively). Pupillary and limbic contrast experiments yielded similar results with EERs degrading from 4.93-6.96% for low-quality setups to 5.72-8.09% and from 2.71-2.74% to 3.62% EER in high-quality configuration. Finally, focus-based filtering turned out to have the most pronounced impact on observed long. vs. short-timespan comparisons. If just highly-focused images are considered, temporal effects more than doubled EER from 3.19% to 7.26%, whereas low-focus had a minor impact from 4.72% to 6.4% EER. This result is interesting, as it confirms observations in [15, 16] where the larger ageing-effect is noted on a higher-quality subset.

Generally, when limiting data to high-quality setups, the impact on recognition rates observed when comparing long-timespan versus short-timespan data was more pronounced compared to low-quality setups and overall, temporal impact (including any other systematic or non-systematic bias not eliminated by the quality criterion) turned out to be inferior to the high versus low quality measurements when comparing ROC performance. This would justify looking into both predictability models, as in [20], as well as overall quality metrics for datasets setting studies into context, which however is not the scope of this work.

4.3 On the Impact of Segmentation Algorithms on Observed Ageing Effects

Finally, we also compared the impact of segmentation algorithms on observed ageing effects, taking the diversity introduced by quality predicates into account. Figure 4 illustrates the different behaviour of the traditional HT-based CAHT (left) versus the iterative 2-step segmentation algorithm WAHET (right) with regards to individual quality predicates.

While the choice of feature extraction algorithms had only a minor impact on the general shape of ROC curves, for different segmentation techniques certain quality restrictions had an even clearer, distinct impact on accuracy and therefore also on induced intra-session variability co-occurring with ageing effects. Table 3 lists all obtained experimental EERs in various tested configuration. This leads to observations as follows: when investigating ROCs compared for low-quality setups it is interesting to see, that within the high-security part of the ROC (requested low FARs), using the less accurate CAHT algorithm, minor segmentation errors are likely to be the primary cause for an obtained higher accuracy of long-timespan comparison compared to short-timespan comparisons.

However, while segmentation algorithms turned out to have some impact on the magnitude of the observed longitudinal effect and on accuracy at low FARs, they did not change the overall picture of observable impact of an increased timespan. However, depending on the choice of quality-based filtering, this impact is either more pronounced or less pronounced than other variability.

4.4 On the Impact of Experimental Configuration on Observed Ageing Effects

In order to finally test typical setups, we defined 4 test scenarios:

- **Controlled, high quality**: Quality Setup with complete exclusion of images from sessions with glasses or reflections, setting a low threshold for pupillary and limbic boundary deviation of two employed segmentation algorithms and excluding images with large occlusion areas (noise mask) and maximum allowed pupillary dilation ($\alpha = 2, \beta = 10, \gamma = 10, \delta = 10, \epsilon = 10, \phi = 61, \psi = 6000$).
- **Controlled, low quality**: Like high quality setup, but without restrictions regarding occlusion areas (noise mask) and maximum allowed pupillary dilation ($\alpha = 2, \beta = 10, \gamma = 10, \delta = 10, \epsilon = 10, \phi = \infty, \psi = \infty$).
- **Uncontrolled, high quality**: This setup includes challenging sessions with glasses and reflections, setting a low threshold for pupillary and limbic boundary deviation of two employed segmentation algorithms and excluding images with large occlusion areas (noise mask) and maximum allowed pupillary dilation ($\alpha = 5, \beta = 10, \gamma = 10, \delta = 10, \epsilon = 10, \phi = 61, \psi = 10000$).
- **Uncontrolled, low quality**: Quality Setup allowing glasses/reflections with very modest segmentation quality checks ($\alpha = 5, \beta = 20, \gamma = \infty, \delta = \infty, \epsilon = \infty, \phi = \infty, \psi = \infty$).

For simulation of ageing effects under controlled recording conditions, sessions are restricted to sets 0 to 2 without images containing glasses and more rigorous specular reflections, whereas uncontrolled involves images from all sets 0 to 5, including increased intra-personal variability.

ROCs in Figure 5 illustrate, how recognition rates generally gradually decrease from intra-session comparisons to intra-year all-session (including intra-session comparisons) to inter-session experiments, and finally, long-timespan comparisons. However, by varying the quality setup similar effects to the one described for the pupil dilation predicate can be observed, where 1-year performance is reduced compared to 4-year performance. Again, the quality filter kept high intra-session variability in the 1-year comparisons, where the 4-year comparsions are able to benefit from the less varying recording conditions in the 2013 dataset. Together with the unbalanced session size this is likely to be a major cause for the observed adverse effects in the uncontrolled, low-quality setup. Further, it is interesting to see the effect of full cross-comparison (all session) versus an exclusion of low-challenge intra-session comparisons. Finally, it turned out that within the most restrictive quality scenario (controlled, high-quality) long-timespan recognition degradation was most pronounced.

Given the observable adverse effects it is evident, that any ageing study might suffer from similar quality impact unless fair data eliminates all biasing effects. If access to images is ungranted and in the absence of universal notions of quality indicators for the datasets employed, conducted experiments showed the high risk of obtaining controversial results, supporting criticism in ageing studies based on closed-access data. It is therefore essential to outline the configuration of experiments and pay special attention towards high-quality and low-quality setups, where long-timespan versus short-timespan comparisons can lead to quite controversial conclusions.

5 Conclusion

Reasons for observed increased error rates over long temporal distance between biometric recordings are manifold: Biological ageing, changed behaviour, modified recording conditions, aged hardware. This paper investigated the impact of ageing on segmentation versus feature extraction, highlighting the critical impact of quality predicates, experimental configuration, and employed datasets on evaluations. Using CASIA's database of 36,240 images of 104 classes with intra-year and 4-year comparisons we identified the critical impact of segmentation quality predicates on recognition accuracy. While quality-based filtering based on individual predicates could increase/decrease accuracy by large factors up to 3 (1.97 vs. 6.8% EER based on pupil contrast for CAHT), long vs. short-timespan comparisons turned out to have an almost similar impact on recognition accuracy (e.g 3.19% to 7.26% EER for high-focus intra-year vs. 4-year). Due to the potential attribution to many factors, this paper does not aim to answer the question whether there are irreversible ageing effects (therefore, certainly all other influencing factors have to be excluded). Instead, this paper illustrates the feasibility to observe positive and negative longitudinal effects by just varying and controlling the (segmentation) quality environment. This highlights an imminent problem of ageing studies: the lack of large and standardized temporal variation in public datasets as well as additional metadata (age of subject, recording parameters) and strictly enforced unified recording and quality guidelines over time-separated sessions. While several databases exist, partly with quite different conditions, it is difficult to keep environmental conditions unmodified over large timespans. Further, biometric processing consists of several processing steps with possible ageing impacts on different parts of a system, not necessarily related to a modification of the biometric signal itself questioning uniqueness. Especially the intransparency of failure to process / failure to acquire images is an important issue and should be stated in papers.

This paper verified via experiments, that traditional biometric quality estimates employed as segmentation checks (e.g. restricting recording samples to only samples where little segmentation errors can occur) can significantly affect the perceived "aged" genuine score distribution. While originally conducted to be used to improve decisions based on quality estimates, the same quality differences might lead to differently observed effects which might be attributed to ageing but instead reflect deficiencies in the data not sufficiently eliminating biasing (quality) factors. On the other hand, a study could fail to observe an aging effect when in fact one exists, because of not-controlled quality conditions (ageing impact might be shadowed in datasets with strong intra-personal variability). As an outcome and suggestion this paper highlights the need for transparent recording conditions and pre-evaluation of quality in underlying ageing databases, clear separation of effects on different biometric processing modules, and the provision of further metadata in ageing databases. Iris ageing according to IREX focuses on FNMR changes over time that can not be attributed to non-pupillary-dilation. Given the results in this paper it seems that there are even further impacts which should be considered for coming to more convincing conclusions on the presence of any "permanent" change. When it comes to a definition of ageing it is hard to justify any assumption of monotonicity which further raises the question whether it should be better be seen as a stochastic process. It is important to consider correlation between variables in any employed regression model. Further work will try to continue the exploration in separation of related factors to further progress in this question.

Acknowledgements

The authors gratefully acknowledge the support of Prof. Zhenan Sun, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, for providing access to the dataset employed in this work and to be released as part of the next version (5) of the CASIA iris image dataset. This work was supported by the EU FASTPASS project under grant agreement 312583.

References

[1] Bowyer, K.W., Hollingsworth, K., and Flynn, P.J. Image understanding for iris biometrics: A survey. *Computer Vision and Image Understanding*, 2007, 110, (2), pp. 281–307.

[2] Daugman, J. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2007, 37, (5), pp. 1167–1175.

[3] Baker, S.E., Bowyer, K.W., and Flynn, P.J. Empirical evidence for correct iris match score degradation with increased time-lapse between gallery and probe matches. In *Proc. Int'l Conf. on Biometrics (ICB)*, Alghero, Italy, June 2009, pp. 1170–1179.

[4] Grother, P., Matey, J.R., Tabasi, E., Quinn, G.W., and Chumakov,M. Irex vi temporal stability of iris recognition accuracy. Interagency report 7948, (NIST, 2013).

[5] Fenker, S.P., and Bowyer, K.W. Analysis of template aging in iris biometrics. In *Proc. IEEE Comp. Vision and Patt. Recog. Workshop (CVPRW)*, Providence, RI, USA, June 2012, pp. 45–51.

[6] Bowyer, K.W. and Ortiz, E. Making sense of the irex vi report. Cvrl technical report, (Univ. of Notre Dame, 2013).

[7] Lanitis, A. A survey of the effects of aging on biometric identity verification. *Int. J. Biometrics*, 2010, 2, (1), pp. 34–52.

[8] Erbilek, M. and Fairhurst, M. Framework for managing ageing effects in signature biometrics. *IET Biometrics*, 2012, 1, (2), pp. 136–147.

[9] Fairhurst, M., and Erbilek, M. Analysis of physical ageing effects in iris biometrics. *IET Computer Vision*, 2011, 5, (6), pp. 358–366.

[10] Michael, R. and Bron, A. The ageing lens and cataract: a model of normal and pathological ageing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 2011, 366, (1), pp. 1278–1292.

[11] Watson, D.G., Maylor, E.A., and Bruce, L.A.M. Search, enumeration, and aging: Eye movement requirements cause age-equivalent performance in enumeration but not in search tasks. *Psychology and Aging*, 2005, 20, (2), pp. 226–240.

[12] Kalka, N.D., Zuo, J., Schmid, N.A., and Cukic, B. Estimating and fusing quality factors for iris biometric images. *IEEE Trans. Systems, Man and Cybernetics, Part A*, 2010, 40, (3), pp. 509–524.

[13] Mehrotra, H., Vatsa, M., Singh, R., and Majhi, B. Does iris change over time? *PLoS One*, 2013, 8, (11), pp. e78333.

[14] Tome-Gonzalez, P., Alonso-Fernandez, F., and Ortega-Garcia, J. On the effects of time variability in iris recognition. In *Proc. Int'l Conf. Biometrics: Th., App., and Syst. (BTAS)*, Arlington, VA, USA, September 2008, pp. 1–6.

[15] Sazonova, N., Hua, F., Liu, X., Remus, J., Ross, A., Hornak, L., and Schuckers, S. A study on quality-adjusted impact of time lapse on iris recognition. *Proc. SPIE 8371*, 2012, pp. 83711W–9.

[16] Czajka, A. Template ageing in iris recognition. In *Proc. Int'l Conf. on Bio-Inspired Systems and Signal Proc. (BISSP)*, Barcelona, Spain, February 2013, pp. 1–8.

[17] Sgroi, A., Bowyer, K.W., and Flynn, P.J. The prediction of young and old subjects from iris texture. In *Proc. Int'l Conf. on Biometrics (ICB)*, Madrid, Spain, June 2013, pp. 1–5.

[18] Grother, P., Matey, J., Quinn, G., and Tabassi, E. Iris Permanence. What We Know, What We Don't, and How to Find Out More. Presentation at *Global Identity Summit, Iris Workshop*, Tampa, USA, September 2014, p. 17, Retrieved from `http://www.biometrics.org/bc2014/presentations/Tues_1819_Grother_1400.pdf`.

[19] Ortiz, E., Bowyer, K.W., and Flynn, P.J. A linear regression analysis of the effects of age related pupil dilation change in iris biometrics. In *Proc. Int'l Conf. on Biometrics: Th., App., and Syst. (BTAS)*, Washington DC, USA, September 2013, pp. 1–6.

[20] Trokielewicz, M. Linear regression analysis of template aging in iris recognition. In *3rd Int'l WS on Biom. and Forensics*, Gjovic, Norway, March 2015, pp. 1–6.

[21] Poh, N. and Kittler, J. A unified framework for biometric expert fusion incorporating quality measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012, 34, (1), pp. 3–18.

[22] Rathgeb, C., Uhl, A., and Wild, P. *Iris Recognition: From Segmentation to Template Security*, volume 59 of *Advances in Information Security*. (Springer, 2013).

[23] Uhl, A. and Wild, P. Weighted adaptive hough and ellipsopolar transforms for real-time iris segmentation. In *Proc. Int'l Conf. on Biometrics (ICB)*, New Delhi, India, March 2012, pp. 1–8.

[24] Ma, L., Tan, T., Wang, Y., and Zhang, D. Efficient iris recognition by characterizing key local variations. *IEEE Trans. on Image Proc.*, 2004, 13, (6), pp. 739–750.

[25] Masek, L. Recognition of human iris patterns for biometric identification. Master's thesis, (Univ. of Western Australia, 2003).

[26] Xiao, L., Sun, Z., He, R., and Tan, T. Coupled feature selection for cross-sensor iris recognition. In *Proc. Int'l Conf. Biometrics: Th., App., and Syst. (BTAS)*, Washington DC, USA, September 2013, pp. 1–6.

[27] Hofbauer, H., Alonso-Fernandez, F., Wild, P., Bigun,J., and Uhl, A. A ground truth for iris segmentation. In *Proc. 22th Int'l Conf. on Pattern Recognition (ICPR)*, Stockholm, Sweden, August 2014, pp. 1–6.

Table 1: Segmentation quality parameters and predicates (subscripts denote segmentation: 1=WAHET, 2=CAHT).

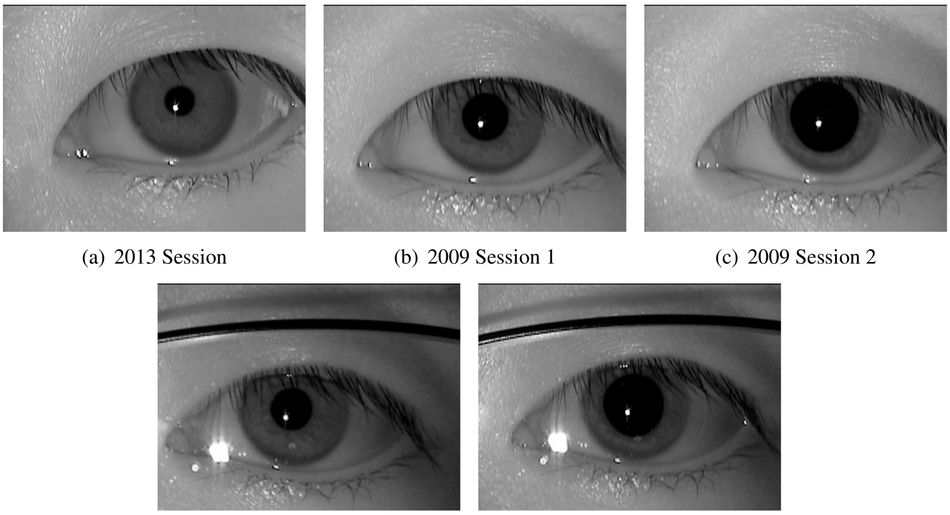
Par.	**Predicate**	**Description**
α	$Session \in \{0, 1, \ldots, \alpha\}$	constraint on 2009 intra-session variability
β	$\|PupilCenter_1 - PupilCenter_2\| < \beta$	both segmentations yield similar pupil center
γ	$\|Iris_1 - Iris_2\| < \gamma$	both segmentations yield similar iris center
δ	$\|PupilRadius_1 - PupilRadius_2\| < \delta$	max. pupillary radius deviation of δ for both segmentations
ϵ	$\|IrisRadius_1 - IrisRadius_2\| < \epsilon$	max. limbic (outer iris) radius deviation of δ for both segmentations
ϕ	$\|PupilRadius_1\| < \phi$	constraint on max. absolute pupillary dilation (using first segmentation)
ψ	$\|IrisArea_1 - IrisNettoArea_1\| < \psi$	occlusion constraint limiting squinting eyes (using first segmentation)
μ	$\|PupilContrast_1\| > \mu$	constraint on absolute pupillary contrast.
ν	$\|IrisContrast_1\| > \nu$	constraint on absolute limbic contrast.
ω	$\|ImageEnergy\| > \omega$	constraint on in-focus assessment.

Table 2: Metadata (year of recording, session, number of classes / images, percentage of images with closed eyes / glasses), errors (exposure problems, consistent segmentations) and characteristics (m/f ratio and age distribution).

Year	**Session**	**Classes**	**Images**	**Closed Eyes**	**Glasses**	**Over/under exposed**	**Same Seg. Result**	**Gender Balance**	**Age Distr.**
2013		96	1920	0.16%	2.08%	0.05%	89.58%	56:44	31.7 ±6.2
2009	1	100	12000	0.85%	0%	1.09%	83.82%	52:48	28.1 ±6.5
2009	0+2	100	12240	1.24%	0%	0.67%	73.91%	52:48	28.2 ±6.5
2009	4	42	5040	0.56%	100%	0%	34.40%	71:29	25.7 ±3.5
2009	3+5	42	5040	0.58%	100%	0%	30.60%	71:29	25.7 ±3.5

Table 3: Impact of quality predicates on accuracy for different segmentation methods and feature extraction algorithms.

Parameter Constraint	**Segmentation Method**	**QSW 1-year**	**QSW 4-years**	**LG 1-year**	**LG 4-years**
$\phi < 35.74$	Wahet	5.51	1.62	3.78	1.60
	CAHT	4.89	1.63	2.42	1.69
$\phi \geq 35.74$	Wahet	3.74	7.71	4.55	7.91
	CAHT	4.80	6.45	4.62	7.17
$\psi < 4600$	Wahet	3.27	6.79	3.71	6.28
	CAHT	3.39	5.80	3.87	5.82
$\psi \geq 4600$	Wahet	5.06	6.92	5.31	6.06
	CAHT	8.59	8.54	5.76	5.79
$\mu < 7.46$	Wahet	5.95	8.82	6.96	8.09
	CAHT	7.05	9.05	6.80	8.38
$\mu \geq 7.46$	Wahet	2.29	2.77	2.71	3.62
	CAHT	2.39	2.64	1.97	3.12
$\nu < 1.33$	Wahet	4.59	6.52	4.93	5.72
	CAHT	5.95	7.18	5.19	6.26
$\nu \geq 1.33$	Wahet	1.78	5.70	2.74	3.62
	CAHT	2.83	4.83	3.08	3.65
$\omega < 4190$	Wahet	5.12	6.82	4.72	6.40
	CAHT	4.10	5.72	3.86	5.99
$\omega \geq 4190$	Wahet	2.91	8.40	3.19	7.26
	CAHT	3.61	7.16	3.20	6.83



(a) 2013 Session

(b) 2009 Session 1

(c) 2009 Session 2

(d) 2009 Session 4

(e) 2009 Session 5

Figure 1: Sample session variation of iris images of user #0728 in the test ageing dataset.

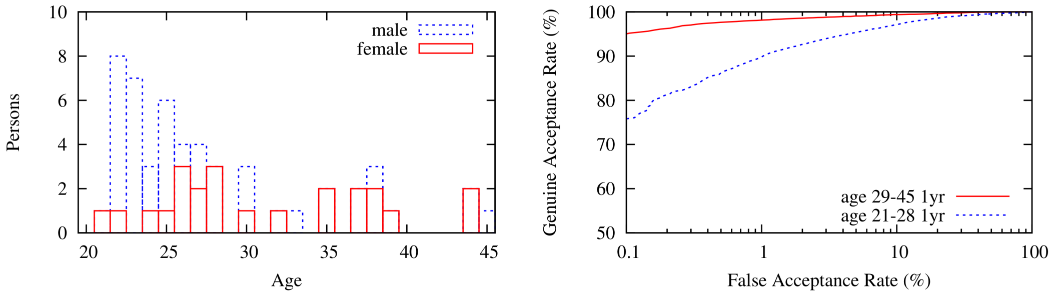


Figure 2: Age distribution (left) and impact on recognition accuracy using LG on Sessions 0+1+2 (right).

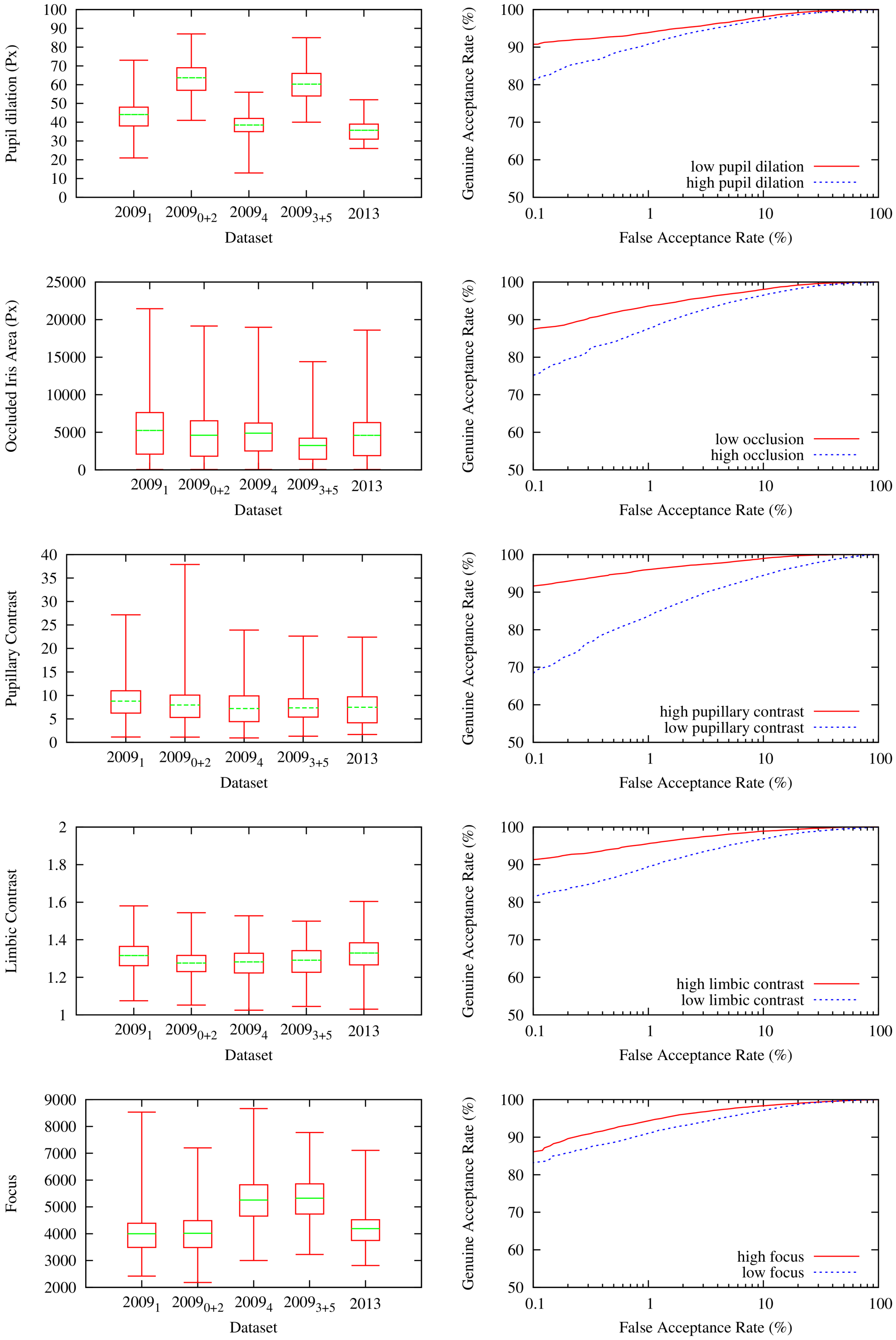


Figure 3: Box plots of quality score values (left) in different subsets and their impact on ROCs using *LG* and *WAHET* (right).

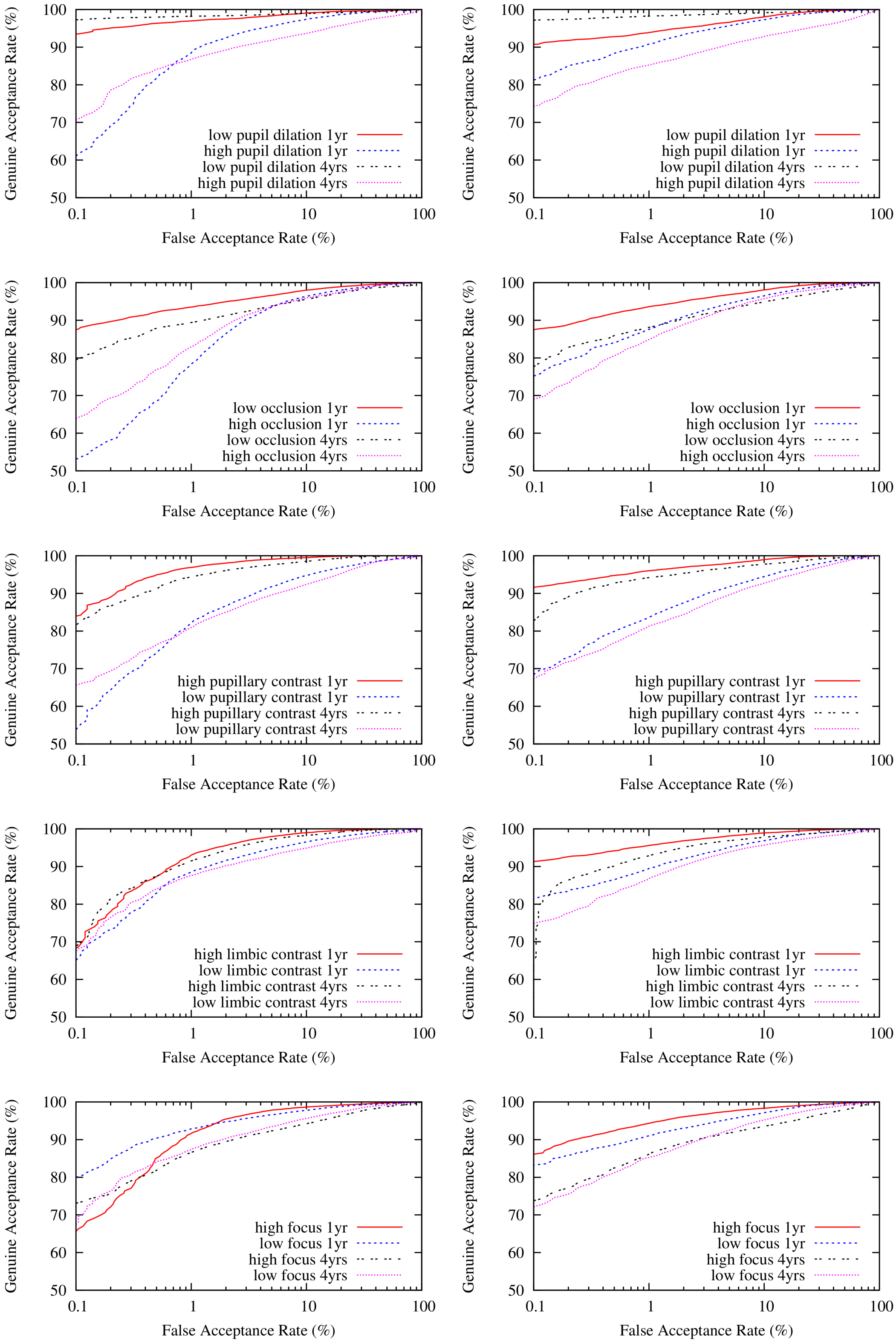
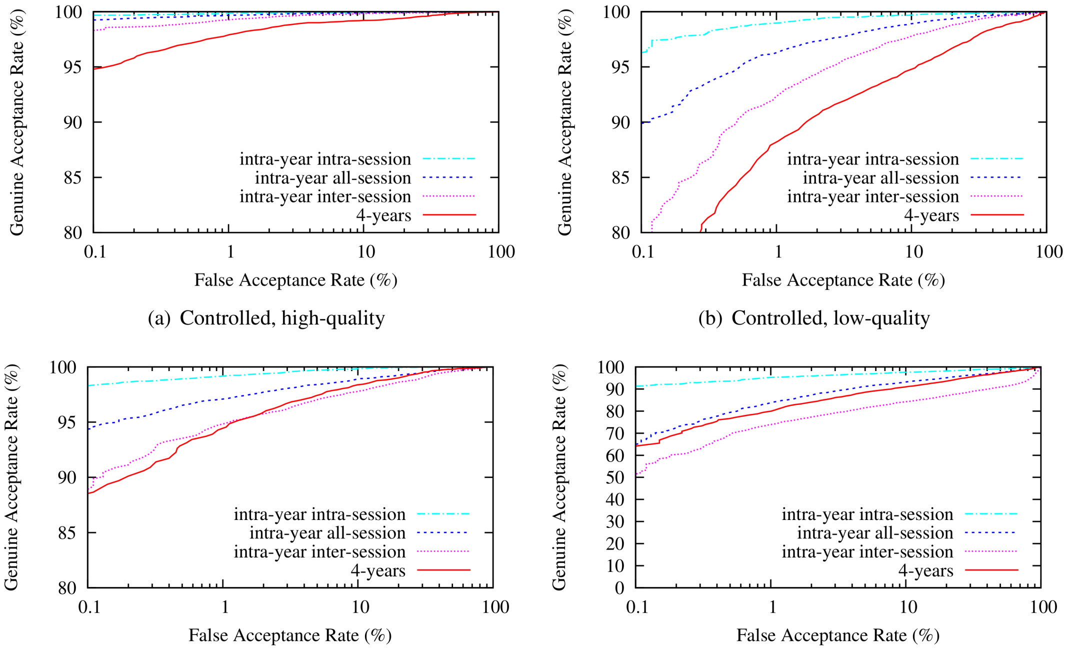


Figure 4: *CAHT* (left) vs. *WAHET* (right) impact on ageing ROC assessments using *LG*.



(a) Controlled, high-quality

(b) Controlled, low-quality

(c) Uncontrolled, high-quality

(d) Uncontrolled, low-quality

Figure 5: ROCs for different quality setups of intra-year vs. 4-year comparisons using *LG* as feature extraction.