Statistical analysis of multiple presentation attempts in iris recognition

Adam Czajka, *Senior Member, IEEE*
Institute of Control and Computation Engineering
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
and
Research and Academic Computer Network (NASK)
ul. Wawozowa 18, 02-796 Warsaw, Poland
Email: aczajka@elka.pw.edu.pl

Kevin W. Bowyer, *Fellow, IEEE*
University of Notre Dame
384 Fitzpatrick Hall, 46556 Notre Dame, IN, USA
Email: kwb@nd.edu

Abstract*—This paper presents experimental results showing uneven distributions of selected iris image quality metrics in the consecutive attempts in a biometric system that allows for multiple attempts to complete a transaction. We consider three iris image quality metrics that can be influenced by user behavior: usable iris area, motion blur and margin adequacy. The quality metrics are used to judge the overall quality of the iris image and accept or reject it on each attempt. The experiment simulates a typical physical access scenario with a maximum of three attempts in one transaction. One conclusion is that subjects rejected on the first attempt do, on average, improve the quality of their iris image on their second attempt. If their second image is rejected, the average quality improvement on the third attempt is lower. A second conclusion is that the probability of a subject being rejected on the second try is higher in average than the probability calculated for all subjects delivering their first samples. The latter finding contrasts with a common belief that each try in a single transaction can be assumed to be a draw from the same authentic distribution (and hence the rejection probabilities are equal in each try). A third interesting and surprising observation is that improvement of sample quality is higher for women than for men. To our knowledge, this paper presents the first research explaining the nature of multi-attempt iris recognition system and delivers conclusions that suggest that the default understanding of this process is too simplistic.

Keywords*—*iris recognition, multiple-attempt system, iris image quality, statistical analysis.

I. INTRODUCTION

Typical biometric systems allow for multiple attempts when presenting a biometric characteristic in a single authentication transaction, in order to minimize number of false rejections at the transaction level. This means that if the subject is rejected on the first image, he or she is asked to present the characteristic again. If the rejection occurs again, then a third try is allowed.

A common understanding of this multi-attempt process is that people behave in the same way in all attempts, i.e., delivering biometric samples of the same quality. Expressing this in a statistical language, one could say that the quality metrics calculated for samples in the first, second and third tries come from identical and independent distributions. Hence, if the chance of being accepted in the first try is – say – 0.8, then the chance of being accepted in two tries is $0.8+0.2\cdot 0.8 = 0.96$, and $0.8+0.2\cdot 0.8+0.2\cdot 0.2\cdot 0.8 = 0.992$ when three tries are allowed. However, this paper shows that this common reasoning is incorrect in iris recognition.

Consecutive tries of the same person may not be statistically independent. In other words, it may be the case that the rejection event causes the subject to somehow present the biometric characteristic differently for the next acquisition, and hence deliver a sample of a different quality. This generates two interesting questions:

1) Is a subject able to give a better-quality sample on the next attempt after being rejected? In other words, can we assume that the subject constantly improves his or her iris presentation as a function of the attempt number?
2) Are the means of iris image quality scores obtained in the second attempt (i.e., from those subjects who were rejected in the first attempt) different from means calculated for a general distribution of all quality scores obtained in the first attempt?

To answer the above questions, an experiment was conducted with 173 volunteers giving their iris samples in maximum three attempts (in one transaction). Subjects were informed whether their images were accepted or rejected, but the reason for rejection was not disclosed. Acceptance of the iris image (left and right independently) was based on three quality metrics estimating a usable iris area, motion blur and correct position of the iris within the image frame (all metrics are specified in Sec. III). All images were acquired with the IrisGuard AD100 iris capture system and the acquisition was attended by an operator to collect a dataset of iris images (described in Sec. IV). Results provided in Sec. V suggest that users rejected in the first attempt do indeed improve image quality on their second attempt, but the probability of their rejection in the second try is higher than the average for the population (hence their improvement does not compensate some unknown effects or characteristics of their irides to be at least as good in the second try as in the first one). We found also that women improve to a higher extent than men. Since we did not find any prior work that would present this kind of analysis, the paper seems to be the first explaining what happens with distributions of the quality scores in multiple attempts in an iris recognition system.

II. RELATED WORK

The up-to-three-tries methodology is very popular in operational authentication scenarios. A good example is the CANPASS system maintained by Canadian Border Services Agency (CBSA) and providing an efficient entry into Canada for frequent travelers [1]. The dataset resulting from this operational installment has been used in IREX VI report evaluating the iris template aging [2].

Independence of the consecutive tries in one transaction seems to be uncritically assumed (although we feel that it might be incorrect) and hence it is difficult to find literature studies to date analyzing this phenomenon thoroughly. The only paper that we are aware of that addresses differences in multiple attempts is a study by Wayman [3]. The author analyzed histograms of genuine scores for "all first attempts, all second attempts by those that failed the first attempt, and all third attempts by those that failed the first two attempts." The author was "rather surprised by the similarity of the three distributions, although their movement to the right indicates an increasing false non-match rate with subsequent tries after failures." Wayman's observation is consistent with one our findings that the subjects rejected on the first attempt have higher chances to be rejected again when compared to the general population.

III. QUALITY METRICS

A. Selection of quality metrics

There is a large number of possible iris image quality metrics, and the most relevant were collected in ISO/IEC 29794-6 international standard [4]. In this research, it only makes sense to consider those that are plausibly under the subject's control. Hence, we consider the following three quality metrics:

1) **usable iris area** that corresponds to a non-occluded iris portion, usable for feature extraction;
2) **motion blur** that appears when a subject moves the eye suddenly in the plane perpendicular to the lens axis when camera triggers the acquisition;
3) **margin adequacy** that penalizes samples with the iris placed too close to the image boundary, as this may negatively influence an iris segmentation process.

The following subsections describe briefly how the above metrics were implemented in this research.

B. Usable iris area

Each acquired image has been segmented using OSIRIS open source software [5] to find inner and outer circular boundaries of the iris as well as a map of occluded pixels within the iris ring. Usable iris area UIA is calculated as a ratio between the number of non-occluded pixels to the entire number of all pixels found in the iris ring (this follows ISO/IEC recommendation [4]). Note that UIA $\in \langle 0, 1 \rangle$, where 0 denotes totally occluded iris and 1 is for an iris image totally free from occlusion.

C. Motion blur

A blurred image (Fig. 1 left) can be modeled as a convolution of non-distorted image with a distortion kernel: point spread function (PSF). The PSF, if known, delivers crucial information about the motion. Namely, its ellipsoidal shape (in contrast to a point) indicates that the object was moving during the acquisition.

Certainly, neither a perfect (non-distorted) image nor a PSF are known. Hence, we apply a blind deconvolution to find a hypothetical perfect image and a hypothetical PSF, Fig. 1 middle. To calculate the width and orientation of the PSF it is transformed to a binary image, Fig. 1 right. The binarization threshold is set so as to minimize the intraclass variance of the resulting black and white pixels.

Two characteristics of the white shape visible in the binary image are used to assess the image blur: 1) the relative area $\mathrm{MB}_{\mathrm{area}}$ expresses the amount of blurriness (regardless its source) and 2) a ratio between its major and minor axes $\mathrm{MB}_{\mathrm{speed}}$ positively correlates with the speed of eye movement. Direction of eye movement is not important hence it is not calculated. Note that $\mathrm{MB}_{\mathrm{area}} \in (0, 1)$, where values near 0 denote a greater sharpness of the original image (i.e., spike-shaped PSF), and values approaching 1 correspond to a large blur. In theory $\mathrm{MB}_{\mathrm{speed}} \in \langle 1, \infty)$, since if no movement is present when acquiring the iris image then both axes of the white shape are equal (and their ratio is 1). On the other hand, the greater the speed of the eye, the higher the ratio of axis length and hence larger $\mathrm{MB}_{\mathrm{speed}}$. In our experiments $\mathrm{MB}_{\mathrm{speed}}$ does not exceed 3.

D. Margin adequacy

Implementation of margin adequacy MA follows directly ISO/IEC recommendation [4]. Namely, in an ideal case (MA $=$ 1) the outer iris boundary should be distant by at least $0.6R$ and $0.2R$ from vertical and horizontal frame boundaries, respectively, where R is a radius of the circle modeling the outer iris boundary. If the outer iris boundary is tangential to or goes beyond the frame boundary then MA $= 0$. Hence, MA $\in \langle 0, 1 \rangle$.

E. Setting acceptance thresholds

To accept a subject's iris image after acquisition, its quality must exceed a minimum quality expressed by the acceptance thresholds set for all the above quality metrics, namely $\mathrm{UIA}^{\mathrm{thr}}$, $\mathrm{MB}^{\mathrm{thr}}_{\mathrm{speed}}$, $\mathrm{MB}^{\mathrm{thr}}_{\mathrm{area}}$, and $\mathrm{MA}^{\mathrm{thr}}$. That is, the i-th image is accepted if $\mathrm{UIA}(i) \geq \mathrm{UIA}^{\mathrm{thr}}$ and $\mathrm{MB}_{\mathrm{speed}}(i) \leq \mathrm{MB}^{\mathrm{thr}}_{\mathrm{speed}}$ and $\mathrm{MB}_{\mathrm{area}}(i) \leq \mathrm{MB}^{\mathrm{thr}}_{\mathrm{area}}$ and $\mathrm{MA}(i) \geq \mathrm{MA}^{\mathrm{thr}}$. In other words, if any of the minimum quality requirements is not met, the image is rejected.

Appropriate thresholds cannot be set arbitrarily. On the one hand we aim at collecting as much data related to the second and third attempts as possible. It could be realized by simply rejecting everyone. But this scenario would be easy to forecast by volunteers who would not be motivated to improve their presentation once being rejected. On the other hand, assuming typical rejection rate for iris recognition (not more than a few percent) would generate a tiny dataset. Hence, a 50% rejection rate was selected as a tradeoff between the operational reality and the anticipated number of samples collected in the second and third attempts.

To set the appropriate thresholds, given the assumed rejection rate, a random subset of 1000 images originating from the existing, not-yet-published dataset was used to calculate a joint distribution of the quality metrics. These samples were acquired by the same camera and in similar circumstances as in this experiment and each distinct eye was represented by a single image (hence we used the data for 1000 different eyes in this step). The inverse joint cumulative distribution analyzed at the rejection rate of 50% resulted in the following threshold values: $\mathrm{UIA}^{\mathrm{thr}} = 0.85$, $\mathrm{MB}^{\mathrm{thr}}_{\mathrm{speed}} = 1.97$, $\mathrm{MB}^{\mathrm{thr}}_{\mathrm{area}} = 0.32$, and $\mathrm{MA}^{\mathrm{thr}} = 1$.

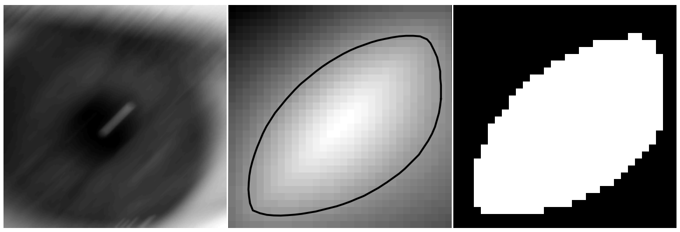

Fig. 1. Illustration of how the motion blur is estimated. **Left:** Center part of the blurred iris image. **Middle:** Point Spread Function (PSF) estimated by way of a blind deconvolution of the blurred iris image. The boundary minimizing the intraclass variance of the 'black' and 'white' pixels (when calculating a binary version of the PSF) is also shown. **Right:** Resulted binary PSF.

IV. Database

A. Collection protocol

Iris images were acquired by the IrisGuard AD100 two-eye camera in an office environment. The acquisition protocol simulated a typical physical access control scenario based on iris recognition. A subject could adjust the height of the camera to align its position with his or her eyes. In each attempt two eyes were photographed and the quality metrics were calculated independently for the left and the right eye. If any of the quality metrics indicated that the image quality is too low, the image and the entire attempt was rejected and the subject was asked to present his or her eyes again. After being rejected three times the acquisition stopped. The subject was not informed about the reason of rejection.

We analyze each eye independently (like in a single-eye iris recognition system). However, using a two-eye imaging system generates some excess data that are not used in this paper. That is, if – for instance – the left eye is accepted and the right is rejected in the first try, the entire attempt is rejected and the second try is started. But in this case we cannot use the image of the left eye acquired in the second try, since the subject would be accepted in the first attempt if the access control system would be based on the left eye only.

B. Data censoring

UIA quality metric depends on the segmentation result. In order not to introduce samples with incorrect segmentation, all the acquired images were manually inspected and those with bad localization of the iris boundaries (Fig. 3 left) or occlusions (Fig. 3 right) were discarded from further processing (26 images out of 796 collected samples were removed).

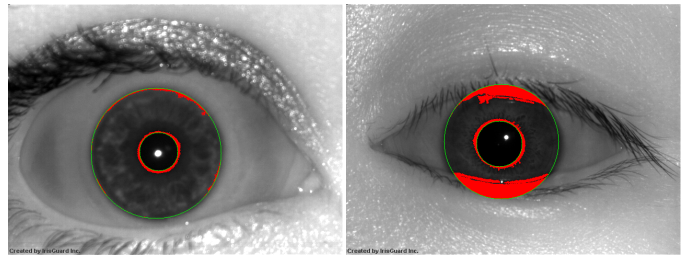

Fig. 2. Example iris images acquired during data collection. Segmentation results are also shown: the usable iris part is all that is enclosed by green contours and not marked in red.

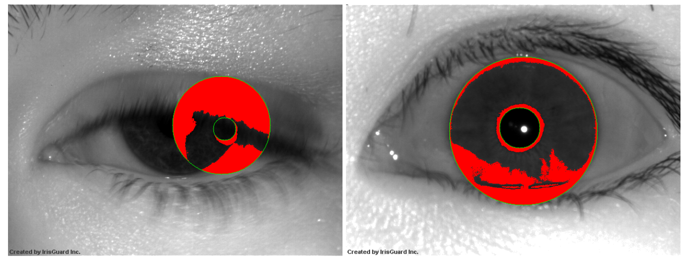

Fig. 3. Examples of the iris images manually rejected (and not further processed) due to bad localization of inner/outer boundary (**left**) or bad marking of occlusions (**right**).

C. Database summary

173 subjects participated in the experiment, hence presenting 346 distinct eyes. Minimum subject age was 18 years, maximum 67 years, and the average was 25 years. 86 subjects (49.7%) were female and 87 subjects (51.3%) were male, giving almost perfect gender balance.

104 left and 104 right eyes out of 173 (i.e. 60%) were rejected in the first attempt. Among those eyes rejected in the first attempt, 83 left and 89 right eyes out of 104 eyes (i.e. 79% and 85% left and right eyes, respectively) were rejected again in the second attempt. In the third attempt still 80 out of 83 left eyes (96%) and 85 out of 89 right eyes (95%) were rejected. These simple statistics suggest that the probability of rejecting a sample in the next attempts after being rejected in the first attempt grows significantly.

V. Results

This section presents results of statistical testing relevant to the questions formulated in the Introduction. For each question the appropriate null and alternative hypotheses are formulated. One-tailed paired-sample t-test is used when answering the question 1, and one-tailed unpaired t-test is used when answering the question 2. To decide on the acceptance of null or alternative hypotheses a significance level $\alpha = 0.05$ is used in all cases. Note that the *margin adequacy* is not used since a maximum value of this quality metric (1) was obtained for all samples. This is due to a mechanism implemented in the AD100 camera, which efficiently controls the position of the iris during acquisition. To simplify the calculations, two quality metrics related to motion blur were replaced with a geometric mean of $\mathrm{MB}_{\mathrm{speed}}$ and $\mathrm{MB}_{\mathrm{area}}$. It makes sense since these single metrics are scaled differently and an increase in either of them should denote an increase in the motion blur in general.

A. Question 1: Is a subject able to give a better-quality sample on the next attempt after being rejected?

To answer this question let's assume that the answer is negative. It means that the average quality in the second try is statistically equal to the average quality in the first try (for those rejected). Let's formulate a null hypothesis H0 as: 'the mean value of the quality score obtained in the first try for those rejected in the first try is equal to the mean value of the quality score obtained in the second try'. Hence, the alternative hypothesis H1 related to the *usable iris area* should state that the mean value of the *usable iris area* is higher in the second attempt. The alternative hypothesis H1 related to the *motion blur* should state that the mean value of the *motion blur* is lower in the second attempt.

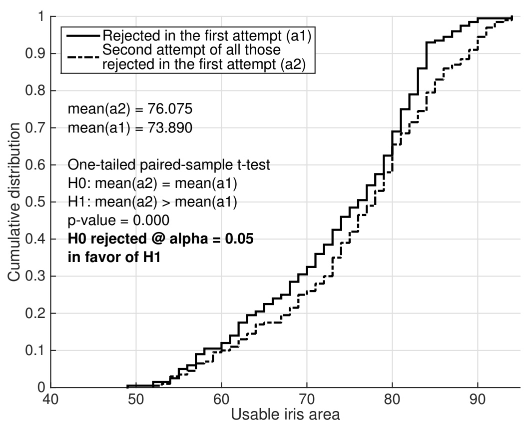

Fig. 4. Cumulative distribution functions for the rejected first-attempt quality scores vs. second-attempt quality scores from those who had a first attempt rejected. Quality score: *usable iris area*.

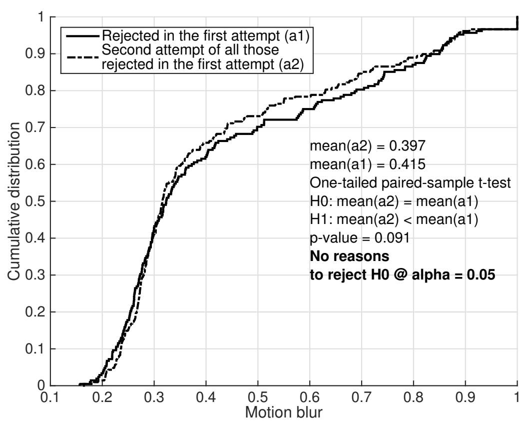

Fig. 5. Same as in Fig. 4 except that *motion blur* is the quality score.

It turns out that the answer to question 1 is affirmative: the average *usable iris area* is higher on the second attempt, **Fig. 4**) and the average *motion blur* is lower on the second attempt, **Fig. 5**). Statistical tests suggest that this phenomenon is stronger for *usable iris area* (we get a statistically significant increase in usable iris area on the second attempt) and weaker for *motion blur* (difference in motion blur in two first attempts is not statistically significant).

An interesting observation is that further improvement on average for a third attempt is small for both quality metrics

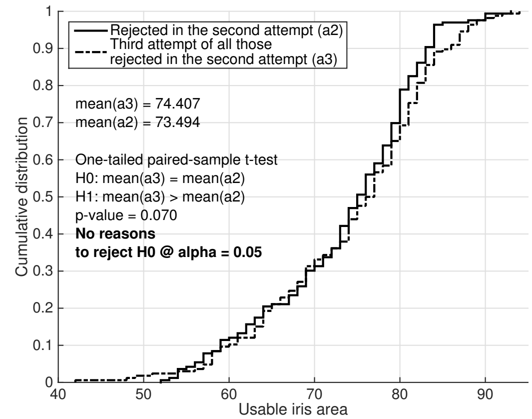

Fig. 6. Cumulative distribution functions for the rejected second-attempt quality scores vs. third-attempt quality scores from those who had a second attempt rejected. Quality score: *usable iris area*.

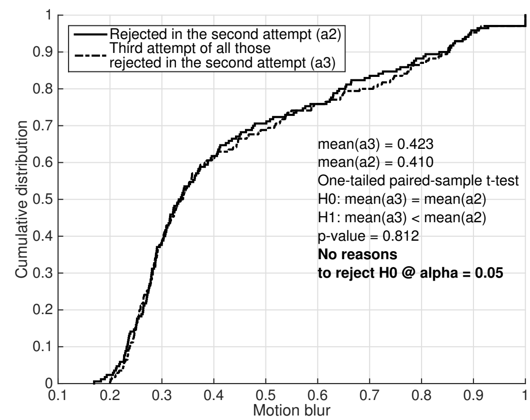

Fig. 7. Same as in Fig. 6 except that *motion blur* is the quality score.

analyzed (**Figs. 6 and 7**). It means that the rejected subjects give better samples in the second try, but we should not expect an identical improvement on the third attempt.

B. Question 2: Are the means of iris image quality scores obtained on the second attempt (i.e., from those subjects who were rejected on the first attempt) different from a mean of a general distribution of all quality scores obtained on the first attempt?

To answer this question let's again assume that the answer is negative. Let's formulate a null hypothesis H0 as: 'the mean value of the quality score obtained in the first try for all subjects is equal to the mean value of the quality score given by those rejected in the first try'. Let's assume that the rejected subjects perform worse in their second try. Hence, the alternative hypothesis H1 related to the *usable iris area* states that the mean value of the *usable iris area* is lower for those rejected. The alternative hypothesis H1 related to the *motion blur* states that the mean value of the *motion blur* is higher for those rejected.

Indeed, when observing the results it is clear that iris image quality expressed as *usable iris area* and *motion blur* is **worse** for those rejected on the first attempt. That is, usable iris area is smaller (**Fig. 8**) and motion blur is higher (**Fig.**

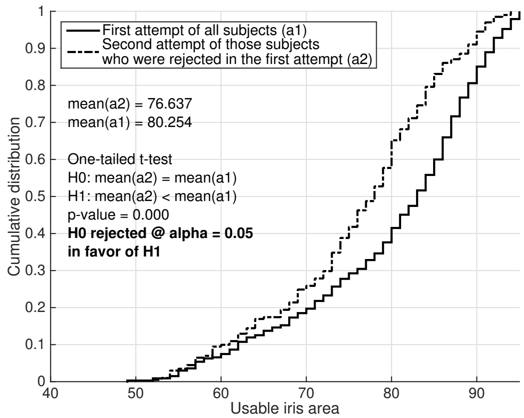

Fig. 8. Cumulative distribution functions for all first-attempt quality scores vs. second-attempt quality scores from those who had a first attempt rejected. Quality score: *usable iris area*.

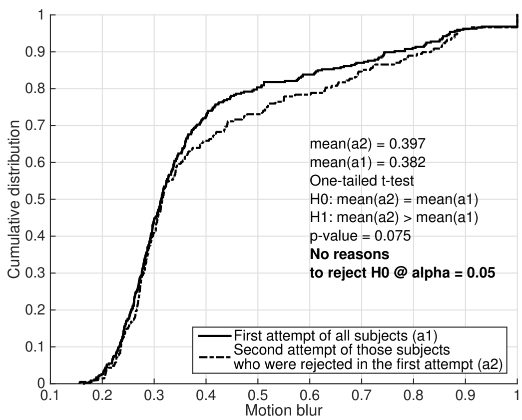

Fig. 9. Same as in Fig. 8 except that *motion blur* is the quality score.

9) for those repeating their attempts. Statistical tests suggest that this phenomenon is stronger for *usable iris area* (we get a statistically significant decrease in usable iris area on the second attempt) and weaker for *motion blur* (difference in motion blur on two first attempts is not statistically significant).

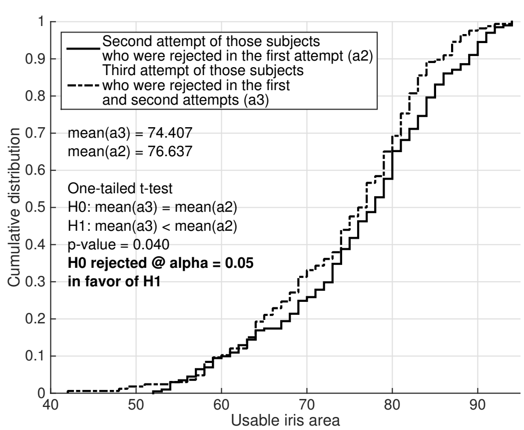

Fig. 10. Cumulative distribution functions for the second-attempt quality scores from those who had a first attempt rejected vs. third-attempt quality scores from those who had first and second attempts rejected. Quality score: *usable iris area*.

Going further, the population of third-attempt quality

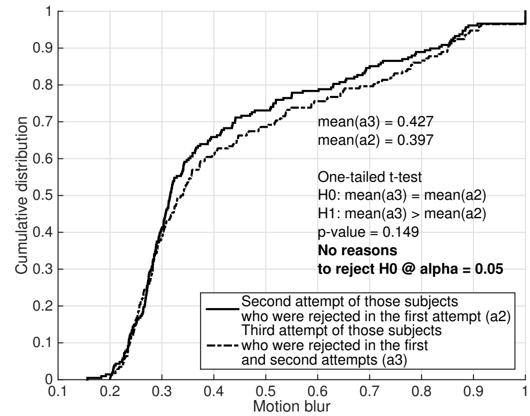

Fig. 11. Same as in Fig. 10 except that *motion blur* is the quality score.

scores from those who had first and second attempts rejected is also different from the population of the second-attempt quality scores. Specifically, iris image quality expressed by *usable iris area* and *motion blur* is again **worse** for those rejected on the first and second attempts: usable iris area is smaller (**Fig. 10**) and motion blur is higher (**Fig. 11**) for those presenting their eyes for the third time. However, observed differences seem to be smaller than earlier, since a statistically significant result is obtained for *usable iris area* and not for *motion blur*.

Answers to the questions 1 and 2 show that the improvement in presentation of the iris does not compensate for some unknown properties of those subjects that are rejected in the first try (as they perform worse on the second attempt when compared to a general population). This strongly suggests that the set of people who are rejected on the first image is not a random subset of the people who present for the first image. The set of people rejected on the first image is a subset of the first presenters that are more likely than the average to have a reason that causes them to be rejected. So, on a second image, this subset may still not do as well as the average presenter for the first image, even though they on average do better than they did on the first image.

C. Dependance on gender

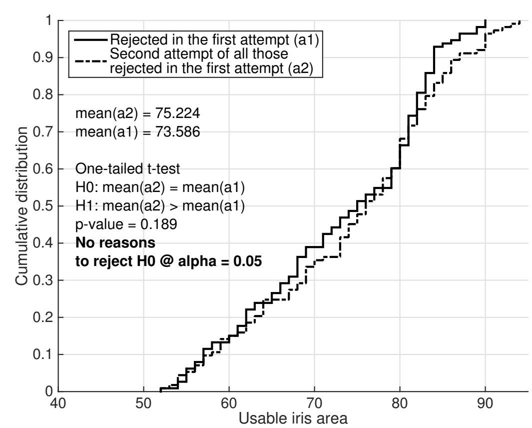

Fig. 12. Same as in Fig. 4 except that only men-related data are analyzed.

It is interesting to note that there is a difference in improvement of *usable iris area* quality metric between women and

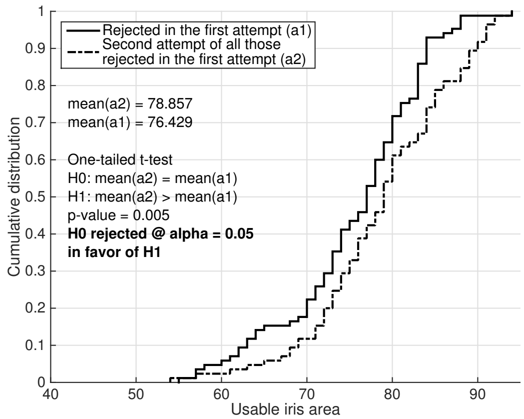

Fig. 13. Same as in Fig. 4 except that only women-related data are analyzed.

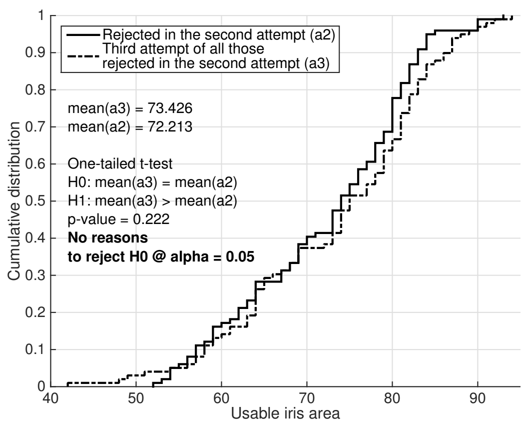

Fig. 14. Same as in Fig. 6 except that only men-related data are analyzed.

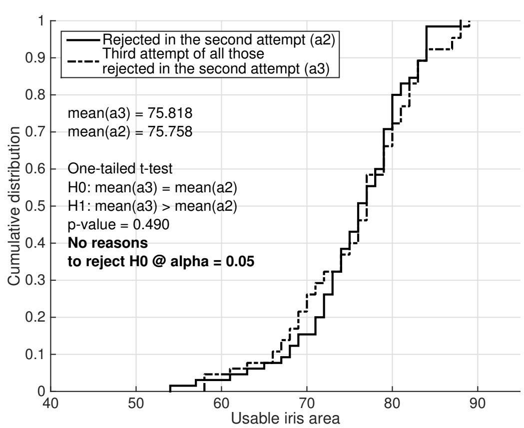

Fig. 15. Same as in Fig. 6 except that only women-related data are analyzed.

men. Namely, improvement on the second attempt for women is stronger than for men (**Figs. 12 and 13**). This may be possibly explained by different reaction of men and women when being rejected. Women could reveal their emotions by changing the facial expression to a greater extent when compared to men. However, gender-dependent differences in improvement on the third attempt are small (**Figs. 14 and 15**). But the difference in improvement between male and female on the third attempt may be partly harder to see because the improvement on the third attempt is small in general.

VI. CONCLUSIONS

Biometric applications commonly allow multiple acquisition attempts to complete a transaction. (In fact, this is not limited to biometric applications, as password applications typically also do this.) One prominent example in the area of iris recognition is the Canadian Border Services Agency frequent-traveler border-crossing application [1]. The common informal understanding of such multiple-attempt transactions implies an assumption that the different subsequent attempts are independent samples from the same underlying authentic distribution. This is embodied in the line of reasoning that says that if 5% of persons are rejected based on their first image, then only 5%·5% = 0.25% would be rejected if up to two images were allowed, and only 0.0125% would be rejected if up to three images were allowed. Our research suggests that the common informal understanding of multiple-attempt biometric transactions is over-simplistic in various ways. People whose first image is rejected are able to give a better quality image, on average, on their second acquisition. However, the overall rejection rate for second images is still significantly larger than the overall rejection rate for first images. We interpret this result as saying that the underlying reason for a failed first image may be reduced to some degree but is not eliminated for the second image. People whose first and second images are rejected have a further reduced probability of a successful image on the third attempt. In effect, the sequence of multiple attempts seems to serve to isolate people who have some inherent difficulty providing a good quality image. In addition to the basic points about probability of success on subsequent images, we also, surprisingly, found a gender difference in the results. It appears that women are better able than men to give an improved second image.

Our results suggest a number of productive future questions. What image quality metrics is the average subject most able to consciously control? What is the most effective way of prompting a subject to give a good quality image on the first acquisition? What is the cause for the apparent gender difference in the results?

ACKNOWLEDGMENT

The authors would like to cordially thank Diane Wright of University of Notre Dame for a diligent dataset acquisition.

REFERENCES

[1] Canadian Border Services Agency. (2015, January 1) CANPASS – Air. Accessed: February 27, 2015. [Online]. Available: http://www.cbsa-asfc.gc.ca/prog/canpass/canpassair-eng.html

[2] P. Grother, J. Matey, E. Tabassi, G. W. Quinn, and M. Chumakov. (2013, July 24) IREX VI: Temporal Stability of Iris Recognition Accuracy. NIST Interagency Report 7948. Accessed: February 27, 2015. [Online]. Available: http://biometrics.nist.gov/cs_links/iris/irexVI/irex_report.pdf

[3] J. Wayman, "Evaluation of the INSPASS Hand Geometry Data," in *National Biometric Test Center Collected Works, v.1.2*. San Jose State University, August 2000.

[4] ISO/IEC 29794-6, "Information technology – Biometric sample quality - Part 6: Iris image data (FDIS)," August 2014.

[5] G. Sutra, B. Dorizzi, S. Garcia-Salitcetti, and N. Othman. (2013, April 23) A biometric reference system for iris. OSIRIS version 4.1. Accessed: October 1, 2014. [Online]. Available: http://svnext.it-sudparis.eu/svnview2-eph/ref_syst/Iris_Osiris_v4.1/