Statistical evaluation of up-to-three-attempt iris recognition

Adam Czajka[†,‡], *Senior Member, IEEE*
[†] Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
[‡] Research and Academic Computer Network (NASK)
ul. Wawozowa 18, 02-796 Warsaw, Poland
aczajka@elka.pw.edu.pl

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

Abstract

Real-world biometric applications often operate in the context of an identity transaction that allows up to three attempts. That is, if a biometric sample is acquired and if it does not result in a match, the user is allowed to acquire a second sample, and if it again does not result in a match, the user is allowed to acquire a third sample. If the third sample does not result in a match, then the transaction is ended with no match. We report results of an experiment to determine whether or not successive attempts can be considered as independent samples from the same distribution, and whether and how the quality of a biometric sample changes in successive attempts. To our knowledge, this is the first published research to investigate the statistics of multi-attempt biometric transactions. We find that the common assumption that the attempt outcomes come from independent and identically distributed random variables in multi-attempt biometric transactions is incorrect.

1. Introduction

Biometric systems typically allow for multiple attempts in a single authentication transaction to minimize the number of false rejections. That is, if a subject is not matched on the first sample, he or she is asked to present the biometric characteristics again. After a second non-match, a third presentation is allowed, but a third non-match result causes a rejected transaction in most biometric systems. This up-to-three-tries methodology is very popular in operational authentication scenarios. A good example is the CANPASS system maintained by the Canadian Border Services Agency (CBSA) and providing an entry into Canada for frequent travelers[1]. The dataset resulting from this operational application has been used in, *e.g.*, the NIST IREX VI report on the topic of iris template aging [1]. Another example is the AADHAAR project, in which the subjects are allowed to make five attempts in a single fingerprint recognition transaction[2].

One of the important questions is: what is the acceptance probability when one, two, three, ..., N attempts are permitted in a single transaction? It is common to assume ergodicity of biometric comparisons when calculating the transactional false rejection rate. Assuming that the probability of being accepted in a single attempt is – say – 90%, a common answer would be 90% + (100% - 90%) $\cdot$ 90% = 99% for up-to-two-attempt system, 90% + (100% - 90%) $\cdot$ 90% + (100% - 90%)^2 $\cdot$ 90% = 99.9% for up-to-three-attempt system, and so forth. When making these calculations, one assumes that subjects' behavior is uniform across the populations and attempts, hence the resulting comparison scores achieved in the first, second and third attempts are *i.i.d.* random variables, *i.e.*, are statistically independent and come from the same distribution. This assumption has unknown source. We guess that this may have come from password-based authentication, in which typing errors might be more like a draw from the same distribution in each try. This work shows that the Bayes' equation cannot be used to estimate the transaction-level rejection rate in this case as it is over-simplistic. In particular, the research results presented in this paper are organized around answering the following three questions related to multi-attempt biometric systems.

Question 1: Is the distribution of comparison scores obtained in the second attempt (*i.e.*, from those subjects who were rejected in the first attempt) different from a general distribution of all comparison scores obtained in the first attempt? If so, does this difference still exist when second- and third-attempt comparison scores are analyzed?

Question 2: Is a subject able to give a better comparison score on the next attempt after being rejected? In other words, can we assume that the subject improves his or her iris presentation on each successive attempt allowed after a

[1] http://www.cbsa-asfc.gc.ca/prog/canpass/canpassair-eng.html

[2] https://uidai.gov.in/UID_PDF/Committees/Biometrics_Standards_Committee_report.pdf

previous attempt resulted in a non-match?

Question 3: If there are differences in comparison score distributions among attempts, what causes them? In particular, which dimensions of iris image quality can and can't a subject improve with conscious effort?

To answer the above questions, an experiment was conducted with 120 volunteers enrolled in an iris recognition system and returning after approximately two months to perform an up-to-three-attempt verification transaction. Results suggest that users presenting their irides in the second attempt perform significantly worse than a population of subjects presenting their eyes for the first attempt (re: Question 1), despite the fact that those persons rejected on their first attempt give better comparison scores in their second attempt (re: Question 2). The latter means that this small improvement does not compensate for some other (yet to be identified) reasons making it harder for these subjects to interact with a biometric system. We show that usable iris area and motion blur have some relation with the improvement of subject's performance, yet the changes in quality metrics are not always statistically significant (re: Question 3). This is the only paper that we are aware of to analyze how the distribution of authentic scores changes on the first, second and third attempts in an iris recognition system. We believe that the results presented in this work may apply to other biometrics modalities as well.

2. Related work

One of a few papers addressing differences in comparison score distributions calculated for different attempts is an evaluation of the INSPASS hand geometry system [5]. The authors "were 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.**" This statement perfectly harmonizes with one of our findings related to a higher probability of being rejected for those who did not succeed on previous attempts when compared to a general population of users having their first attempt. This shift in genuine score distributions can be observed in Fig. 6 of [5]. Kukula and Elliot [3] present deployment of the commercial hand geometry system at the Purdue University's Recreational Sports Center, implementing a three-attempt decision rule. The authors report that "the 1-try false reject rate was 2.26%, the 2-try rate was 1.18%, and the 3-try rate was 0.98%." These estimates are much higher than theoretical values calculated under the assumptions on statistical independency and identical distributions of the comparison scores in each attempt, *i.e.*, $1.18\% > (2.26\%)^2 = 0.067\%$, and $0.98\% > (2.26\%)^3 = 0.002\%$. Similar theoretical underestimation of the error rates is clear when studying AADHAAR report[3]. Namely, "using the residents best finger single-attempt gives an accuracy of 85%", and "using multiple (up to 3) attempts of the same best finger improves the accuracy to 91%." However, if the comparison scores in the consecutive tries are *i.i.d.* random variables, then the accuracy in the three-attempt system should be $99.66\% = 85\% + 15\% \cdot 85\% + 15\% \cdot 15\% \cdot 85\%$, which is higher than reported 91%.

All these papers and reports are silent on statistics of the multi-attempt systems.

3. Database

3.1. Collection protocol

The acquisition protocol simulated a typical physical access control scenario based on iris recognition and it was organized in an office environment. The ambient conditions were stable for all acquisitions. The IrisGuard AD100 two-eye camera was installed on a tripod and each participant could adjust its height to align camera position with his or her eyes. The camera controlled basic properties of the image (such as brightness and contrast) as well as the size and location of the iris within the image. If an excessive pupil dilation is detected, the visible light is automatically turned on to get the pupil constricted. Two acquisition sessions were organized and separated by approximately two months. In the first session volunteers were asked to enroll to the system presenting their eyes once. In the second session the same subjects presented themselves for a verification transaction which allowed up to three iris recognition attempts prior to being definitely rejected. That is, in each verification attempt two eyes were photographed and the comparison scores were calculated independently for the left and the right eye. If either the left iris or the right iris comparison score indicated non-match, the attempt was rejected and the subject was asked to present their irides again. Acquisition was stopped when both the left and the right iris images resulted in acceptable comparison scores or after the third attempt. Subjects were not informed why they had been rejected.

It is important to note that the left and right eyes were processed independently as they would be collected in a single-eye acquisition system. This means that if, for instance, the left eye was rejected but the right eye was accepted in the first try, then only the results related to the left eye were analyzed in the second try for this subject.

3.2. Iris matching methodology used in this work

To perform analyses related to both the iris matching and iris image quality we use the OSIRIS (*Open Source for IRIS*) software [4], which follows Daugman's well-known

[3] https://authportal.uidai.gov.in/static/role_of_biometric_technology_in_aadhaar_authentication.pdf

concepts for the iris image normalization, Gabor-based filtering and quantization of the filtering results to build a fixed-length iris binary code, and comparison between iris codes by calculating a fractional Hamming distance (HD). HD $\in \langle 0; 1 \rangle$, where 0 denotes identical eyes, and 0.5 should be expected on average for a match between two different irides. OSIRIS returns also circular approximations of the pupil and iris, as well as irregularly shaped approximations of occlusions.

To collect samples acquired on the second and third attempts, subjects must be rejected on prior attempts. An aspiration to collect as large a dataset as possible suggests to reject everyone on each attempt. This simple scenario, however, would be easy to guess by volunteers who would not be motivated to improve their presentations once being rejected. On the other hand, using a "real" operational-scenario rejection rate for iris recognition, *i.e.* no more than a few percent, would generate a tiny dataset. Hence, a 50% rejection rate was targeted as a tradeoff between the operational reality and the anticipated dataset volume. To set the acceptance threshold for the OSIRIS method we used an internal, not-yet-published dataset of samples collected in the same office environment and with the same camera as in this work for 1000 distinct eyes. Each iris was represented by two non-same-session samples. This allowed to calculate 1000 genuine and statistically independent comparisons. Analyzing an inverse cumulative distribution of the resulting comparison scores at the 50% level ended up with the acceptance threshold equal to 0.2547. That is, any comparison score that is above this value resulted in the rejection of a given attempt.

3.3. Database statistics

173 subjects participated in the first, enrollment session, and 120 subjects returned after approximately two months to attend the second, verification session. Further statistics and analysis in this paper are provided only for those returning subjects, 67 of whom (56%) were female and 53 of whom (44%) were male, giving a good gender balance. Minimum subject age was 19 years, maximum was 60 years, and the median was 22 years.

50 left eyes and 39 right eyes (out of 120, *i.e.*, 41.7% and 32.5%, correspondingly) were rejected in the first verification attempt. Among those eyes rejected in the first try, 34 left and 31 right eyes (*i.e.*, 68% and 79% of the left and right eyes, respectively) were rejected again in the second try. In the third attempt, still 32 out of 34 left eyes (94%) and 24 out of 31 right eyes (77%) were rejected. It is thus evident that the **probability of rejecting a sample in the next attempts after being rejected in the first try grows significantly**.

We have collected 240 enrollment images in the first session (a single image for the left and right eyes of each subject) and 474 images in the second session (up to three images for each eye, depending on the matching decision). An example good quality image with correct segmentation is presented in Fig. 1A. All the acquired images were visually inspected and those with bad localization of the iris boundaries (Fig. 1 C and E) or occlusions (Fig. 1 B and D) were discarded from further processing (10 and 22 images were removed from those collected in the first and second session, respectively). Since the quality of the enrollment samples is crucial for the verification performance, one of the aims of this visual inspection was to minimize the influence of bad quality enrollment samples on the final conclusions. However, one should be aware that discrepancies in sample quality across the subjects may still be present. Finally, we have **230 enrollment samples** and corresponding **452 verification samples** acquired for **230 unique eyes** available for our analysis.

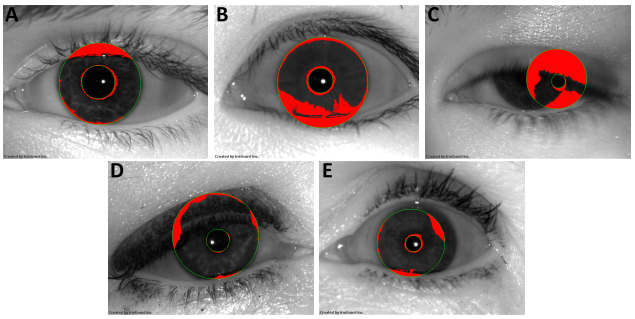

Figure 1. Example of a good quality image (A). Images rejected in visual inspection due to overestimation of occlusions (B), failure in localizing the iris (C), underestimation of the occlusions (D), and imprecise localization of the iris boundary (E).

4. Experimental results

4.1. Answer to question 1

To answer the first question the following sets of comparison scores were calculated:

a1-all: all comparison scores (fractional Hamming distances) obtained from all subjects on the first attempt,

a2-all: all comparison scores obtained from those subjects who were rejected on the first attempt, and presented their irides for the second time,

a3-all: all comparison scores obtained from those subjects who were rejected twice, and presented their irides for the third time.

Cumulative distributions of all the above sets of comparison scores are shown in Fig. 2. One may see a clear shift of the distributions towards worse values in consecutive attempts. This means that the population of subjects rejected in the first attempt and presenting their eyes for the second time performs much worse than a general population of all subjects in their first try. A similar effect, although to

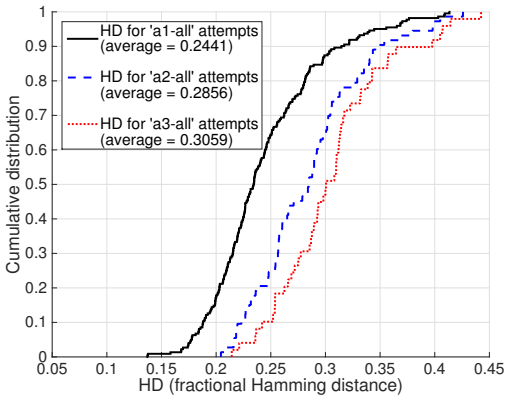

Figure 2. Cumulative distributions of comparison scores obtained in all first attempts (black solid line), all second attempts (blue dashed line) and all third attempts (red dotted line).

some lesser extent, can be observed when comparing those rejected twice with the entire population having second attempt.

We use a two-sample one-sided t-test at the significance level $\alpha = 0.05$ to judge the statistical significance of the observed differences. The null hypothesis (H0) is that the comparison scores obtained in different attempts come from normal distributions with equal means (equality of variances is not assumed). The alternative hypothesis (H1) is that the mean comparison score is worse for those presenting their irides again when compared to those having one attempt less. The test rejected H0 in favor of H1 in both cases, *i.e.*, mean fractional HD in **a2-all** subset of scores is greater than mean fractional HD in **a1-all** subset of scores (p-value=10^{-9}), and mean fractional HD in **a3-all** subset of scores is greater than mean fractional HD in **a2-all** subset of scores (p-value=0.016). Hence, **the answer to the question 1 is: the distribution of comparison scores obtained in the second attempt is different from, and worse than, a general distribution of all comparison scores obtained in the first attempt**. This is also true when all third-attempt comparison scores are compared to all second-attempt comparison scores, suggesting that the probability of being accepted in the next attempt is significantly lower than in the previous attempt.

4.2. Answer to question 2

To answer the second question, we compare the scores of those rejected in the first attempt with their subsequent second-attempt scores. Extending this research to the next attempts, the rejected second-attempt scores are also compared to the corresponding third-attempt scores. Thus, besides **a2-all** and **a3-all** score subsets, we need the following two additional subsets:

a1-rejected: subset of all comparison scores obtained on the first try that were rejected,

a2-rejected: subset of all comparison scores obtained on the second attempt that were rejected.

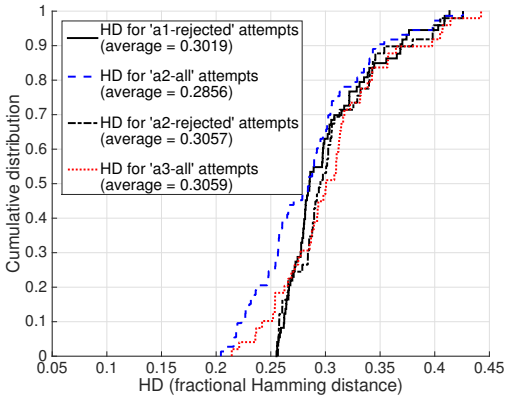

Figure 3. Cumulative distributions of comparison scores in the first rejected attempts (black solid line), all second attempts (blue dashed line), second attempts that were rejected (black dotted-dashed line) and all third attempts (red dotted line).

Cumulative distributions of the appropriate subsets of comparison scores are shown in Fig. 3. To check whether rejected subjects improve in the next attempt there is a need to compare **a1-rejected** subset of scores with **a2-all** subsets of scores for those being rejected in the first try, and **a2-rejected** subset of scores with **a3-all** subset of scores for those rejected twice. Difference in average values for the former case (0.3019 vs. 0.2856, Fig. 3) suggests an improvement in the second try. However, further improvement (in the third try) is marginal (0.3059 vs. 0.3057, Fig. 3). Again, appropriate statistical tests were performed (a two-sample one-sided t-test at the significance level $\alpha = 0.05$). The null hypothesis (H0) corresponds to no improvement in the next try after being rejected. The alternative hypothesis (H1) stands for the improvement (lower mean HD is observed in the next try). We get statistically significant improvement in the second attempt (p-value=0.0174). Not surprisingly, the mean value of scores in the **a3-all** subset is not statistically greater than mean value of scores in the **a2-rejected** subset (p-value=0.51). Hence, **the answer to the question 2 is: a subject is able to give a better comparison score on the second attempt after being rejected in the first try**. However, there is no further improvement observed if the third try is allowed after being rejected twice. A possible interpretation for these results is that the second attempt just more strongly separates those who can give a good-quality iris image and those who cannot, so that those who fail the second time just can't do any better. There may be some non-random reason that subjects rejected twice can't give a good sample. But those who are accepted on the second attempt just had some random thing go wrong on the first attempt and they correct it on the second attempt.

4.3. Answer to question 3

There are various metrics estimating the quality of iris images that should explain the fluctuations in matching performance observed in different attempts. In this work we decided to use only those metrics that depend on subject's behavior, namely *usable iris area* (UIA) related to the eyelid coverage, and *motion blur* (MB) corresponding to the stability of one's head when acquiring the iris image. Eye gaze, being the next possibility, was not considered since it was controlled by the operator during the acquisition.

UIA is the ratio between the number of iris pixels marked as non-occluded to the overall number of iris pixels, and its calculation follows directly the ISO/IEC 29794-6 recommendation [2]. Circular approximations of the inner and outer iris boundaries are used. UIA $\in \langle 0, 100 \rangle$, where 0 denotes totally occluded iris, and 100 corresponds to a clear iris image.

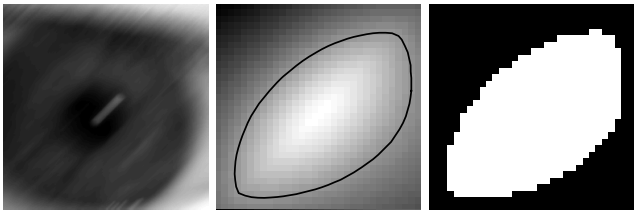

Figure 4. Illustration of how the motion blur is estimated. **Left:** Center part of the iris image with a motion blur. **Middle:** Estimated point spread function (PSF). The boundary minimizing the intraclass variance of the 'black' and 'white' pixels (in a binary version of the PSF) is also shown. **Right:** Binary PSF.

Calculation of the MB is based on a principle that a blurred iris image (Fig. 4, left) can be modeled as a convolution of a non-distorted image with a distortion kernel (point spread function, PSF). The ellipsoidal shape of the PSF indicates that the blurring has a directional character, *e.g.*, the object was moving when being photographed. Also, the wider the PSF the more blurred the image. Both the image and the PSF are unknown, so a blind deconvolution was applied to find a hypothetical perfect image and a hypothetical PSF (Fig. 4, middle). The PSF is transformed into a binary image by minimizing the intraclass variance of the resulting black and white pixels (Fig. 4, right). The MB quality metric is calculated as a geometrical mean of two components derived from a white shape visible in the binary image: 1) the relative area that expresses the amount of blur (estimating also the image sharpness) and 2) a ratio between its major and minor axes, which positively correlates with the speed of the movement. In our experiments MB $\in \langle 0; 1 \rangle$, where lower values are obtained for sharp images.

When finding the answer to question 1 (subsection 4.1) we found that the probability of being rejected in the next attempt is significantly higher than in the previous attempt. Let's analyze what happens with the UIA and MB quality metrics in this case. Average values and cumulative distributions of the usable iris area and motion blur are shown in Figures 5 and 6, respectively. Surprisingly, those attempting for the second time open their eyes wider (higher value of UIA) and stand more still (lower value of MB). It means that those rejected in the first attempt try, in average, to improve in the second attempt. However, this improvement is not observed for UIA in the third attempt, while the motion blur is further reduced in the third try. Statistical testing (analogous to those performed in subsections 4.1 and 4.2) suggests unfortunately that the observed differences in average quality metrics are not statistically significant. Thus, we conclude that there is probably some other significant source of difficulty in generating a good comparison scores for some isolated group of people, even if a little improvement in the selected quality metrics is observed.

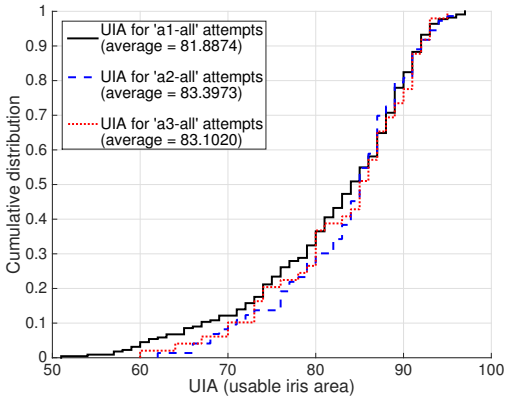

Figure 5. Same as in Fig. 2 except that the cumulative distributions for *usable iris area* (UIA) are plotted.

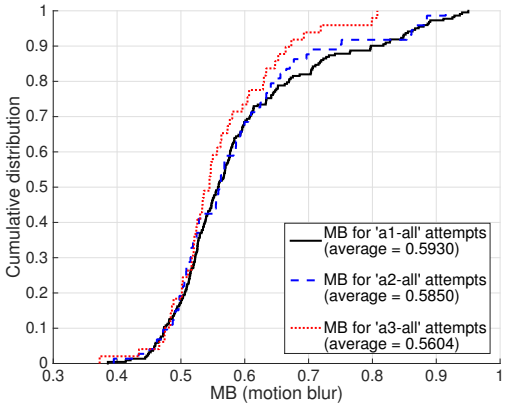

Figure 6. Same as in Fig. 2 except that the cumulative distributions for *motion blur* (MB) are plotted.

When answering the question 2 (subsection 4.2) we found that those rejected in the first attempt generate better quality scores than in the first try, and further improvement in the third try is not observed. This seems to perfectly

coincide with the *usable iris area* quality metric, which is better for those rejected in the first try and repeating their attempt for the second time, Fig. 5. This difference is statistically significant (p-value= 0.016). Decrease of the *usable iris area* in the third attempt for those rejected once (Fig. 5), and differences in the *motion blur* (Fig. 8) are not statistically significant (p-values exceed 0.1 in all those cases).

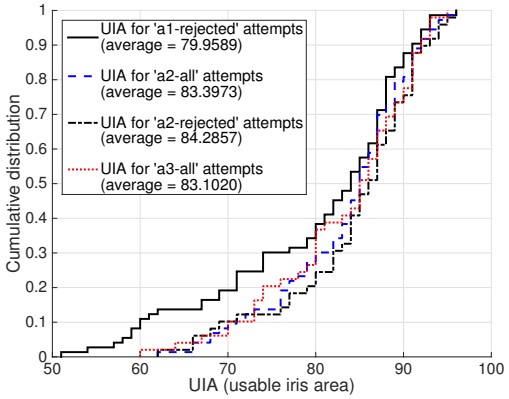

Figure 7. Same as in Fig. 3 except that the cumulative distributions for *usable iris area* (UIA) are plotted.

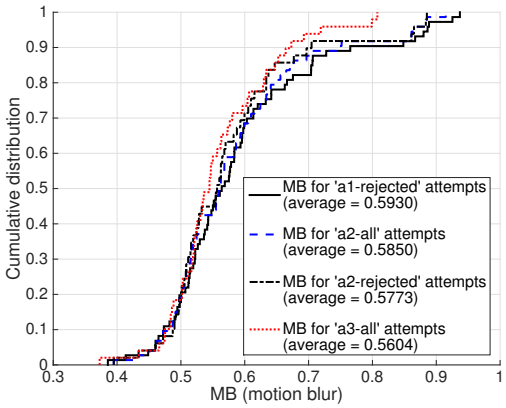

Figure 8. Same as in Fig. 3 except that the cumulative distributions for *motion blur* (MB) are plotted.

Hence, we can provide the following, multi-part **answer to question 3: selected quality metrics (dependent on the subject's behavior) only partially explain the differences in the average comparison scores**. Namely, those rejected in the first try and improving their comparison score in the second attempt open the eyes wider and are not able to stabilize their head more. Those rejected twice and presenting their eyes for the third time do not provide samples of significantly better quality.

5. Conclusions

Our results show that persons who make a second attempt (due to the first attempt not succeeding) will have a lower success rate than those who make a first attempt (some of whom succeed and some of whom do not). At the same time, those persons who make a second attempt do, on average, improve the quality of their biometric sample relative to that provided on their first attempt. And they do improve their probability of success. But this improvement is relative to the 0% success that resulted from this group on the first attempt. Even with improved biometric sample quality on their second attempt relative to their first attempt, the group of persons whose first attempt failed does not achieve second-attempt success equal to that of the overall group's first-attempt success.

Subjects are able to improve some dimensions of biometric sample quality more than other dimensions. For iris recognition, it appears that, on average, subjects are more able to improve the usable iris area dimension of quality than the motion blur dimension. This suggests that iris recognition users could potentially be "coached" to give better quality samples on the first attempt by giving instructions related to the dimensions most under user control. In the iris recognition context of our experiments, allowing a third attempt appears to have marginal value.

Perhaps the major result from our work is that the common understanding of how a multi-attempt transaction will increase the overall transaction success rate is simply incorrect. Allowing a third or a fourth attempt within a transaction is not an effective means to increase the overall transaction success rate. Another important result is that at least some dimensions of sample quality are under conscious control of the user. Explicitly prompting the user for a high-quality sample on these dimensions could result in a higher first-attempt success rate.

Acknowledgements

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

References

[1] P. Grother, J. Matey, E. Tabassi, G. W. Quinn, and M. Chumakov. IREX VI: Temporal Stability of Iris Recognition Accuracy. NIST Interagency Report 7948, July 24, 2013.

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

[3] E. Kukula and S. Elliott. Implementation of hand geometry: an analysis of user perspectives and system performance. *Aerospace and Electronic Systems Magazine, IEEE*, 21(3):3–9, March 2006.

[4] G. Sutra, B. Dorizzi, S. Garcia-Salitcetti, and N. Othman. A biometric reference system for iris. OSIRIS version 4.1, April 23, 2013. Accessed: October 1, 2014.

[5] 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.