Performance of Humans in Iris Recognition: The Impact of Iris Condition and Annotation-driven Verification

Daniel Moreira, Mateusz Trokielewicz, Adam Czajka, Kevin W. Bowyer, and Patrick J. Flynn *

Abstract

This paper advances the state of the art in human examination of iris images by (1) assessing the impact of different iris conditions in identity verification, and (2) introducing an annotation step that improves the accuracy of people's decisions. In a first experimental session, 114 subjects were asked to decide if pairs of iris images depict the same eye (genuine pairs) or two distinct eyes (impostor pairs). The image pairs sampled six conditions: (1) easy for algorithms to classify, (2) difficult for algorithms to classify, (3) large difference in pupil dilation, (4) disease-affected eyes, (5) identical twins, and (6) post-mortem samples. In a second session, 85 of the 114 subjects were asked to annotate matching and non-matching regions that supported their decisions. Subjects were allowed to change their initial classification as a result of the annotation process. Results suggest that: (a) people improve their identity verification accuracy when asked to annotate matching and non-matching regions between the pair of images, (b) images depicting the same eye with large difference in pupil dilation were the most challenging to subjects, but benefited well from the annotation-driven classification, (c) humans performed better than iris recognition algorithms when verifying genuine pairs of post-mortem and disease-affected eyes (i.e., samples showing deformations that go beyond the distortions of a healthy iris due to pupil dilation), and (d) annotation does not improve accuracy of analyzing images from identical twins, which remain confusing for people.

1. Introduction

The literature of iris recognition has been investigating the performance of humans at tasks such as iris texture perception [18, 2, 10, 11, 17] and identity verification [13, 9]. Understanding how people perceive and analyze iris features is useful not only for inspiring the development of better solutions, but also for making them more *human-intelligible*.

Human-intelligible iris recognition is particularly necessary in forensic applications, where experts often rely on the outputs of algorithms for sustaining conclusions and presenting them in a court of law. As pointed out by Chen *et al.* [3], in spite of the traditional iris recognition solutions providing nearly-perfect false match rates [7], they are yet far from being human-friendly enough to convince people who do not possess image processing expertise.

Moreover, human-intelligible iris recognition helps to meet the need to deploy more transparent and accountable systems [14, 23]. As stated in the new European General Data Protection Regulation (GDPR) [20], citizens have the right to obtain an explanation of decisions made about them by algorithms belonging to either government or industry. As a consequence, iris recognition solutions lacking human-intelligible decision processes may face usage hindrances. Indeed, the accountability discussion is also present in the American scientific community, proven by the recent efforts of the National Science Foundation (NFS) in funding research on the topic [19].

This paper contributes to the understanding of how an everyman performs identity verification based on iris patterns. In each of two experiments, subjects are presented a pair of irises and asked to decide whether the pair belongs to the same eye. In the first experiment, we apply a typical multiple-choice questionnaire [13, 9], with no request for image regions or features that justify the decisions. In the second experiment, subjects are asked to manually annotate matching and non-matching regions in the pair of irises, which support their decisions, as an effort to make them more conscious about the task.

In the experiments, there are image pairs representing six different conditions, which are either commonplace or reportedly known to pose challenges to automated systems or to human examiners: (1) healthy eyes that are easily handled by an example iris recognition software, (2) healthy eyes that are challenging for the same software, (3) disease-affected eyes, (4) iris pairs with extensive difference in pupil dilation, (5) irises of identical twins, and (6) iris images acquired from deceased individuals. This variety of conditions allowed us to observe that pairs of images depicting the same iris but with different pupil dilation, iris images

[*] D. Moreira, A. Czajka, K. Bowyer, and P. Flynn are with the Department of Computer Science and Engineering, Univ. of Notre Dame, USA.

M. Trokielewicz is with the Biometrics Laboratory, Research and Academic Computer Network (NASK), Poland.

Corresponding author: Dr. Daniel Moreira (dhenriq1@nd.edu).

Table 1: Literature of human performance in iris recognition. The present work is added to the last row.

Reference	**Problem**	**Subjects (#)**	**Trials per session (#)**	**Average session time (min)**	**Used images**	**Data details**
Stark *et al.* [18]	Iris texture perception	21	100	30	Segmented iris only	100 images depicting 100 distinct irises from 100 distinct individuals
Bowyer *et al.* [2]	Iris texture perception	55	210	$10^{\dagger}$	Whole eye, segmented iris only, or periocular only	630 images depicting 630 distinct irises from 315 distinct individuals
Hollingsworth *et al.* [10]	Iris texture perception	28	196	$10^{\dagger}$	Segmented iris only or periocular only	392 images depicting 392 distinct irises from 196 distinct individuals (including twins' pairs)
Shen and Flynn [17]	Iris texture perception	$21^{\ddagger}$	$64^{\ddagger}$	270	Strip-normalized iris	124 images depicting 62 distinct irises from 62 distinct individuals
McGinn *et al.* [13]	Identity verification	22	190	26	Whole eye	202 images depicting 109 distinct irises from 109 distinct individuals (including twins' pairs)
Guest *et al.* [9]	Identity verification	32	52	10	Whole eye	208 images depicting 104 distinct irises from 104 distinct individuals
This work	**Identity verification**	**114**	**30^{*}, $24^{\star}$**	**17^{*}, $15^{\star}$**	**Segmented iris only**	**1360 images of 512 distinct irises from 512 individuals (with varied pupil dilation, twins', disease-affected, and post-mortem samples)**

† Lower-bound estimated value; each subject had three seconds to inspect each trial — ‡ Average value of three sessions
∗ Conducted at the University of Notre Dame — ⋆ Conducted at the Research and Academic Computer Network (NASK)

of twins, and post-mortem samples are the most challenging to humans. Also, subjects were able to improve their recognition accuracy when they were asked to manually annotate regions supporting their decision. That was not true, however, in the case of iris images of identical twins, which were so confusing that the assessed numbers of improved and worsened decisions were similar.

In summary, this paper advances the state of the art in human examination of iris images with the following contributions:

- Assessment of human skills in verifying the identity of iris images presenting different conditions, including healthy eyes of unrelated individuals, of identical twins, and never used before disease-affected and post-mortem iris samples.
- Employment of custom software to allow subjects to annotate the image regions they rely upon to classify an iris pair, and analysis of how this helps them to provide more accurate decisions.
- Introduction of the notation of *non-matching regions*, besides the typical concept of *matching regions*, in the process of matching pairs of iris images.

The remainder of this paper has four sections. In Sec. 2, we discuss the related work, while in Sec. 3, we detail the configuration of experiments. In Sec. 4, in turn, we report the obtained results, followed by Sec. 5, where we discuss the lessons learned from the experiments.

2. Related Work

There are only a few works related to human examination of iris images. Stark *et al.* [18] studied how people classify iris textures into categories. They used a software tool that allowed subjects to browse a set of segmented near-infrared iris images and use a drag-and-drop scheme to organize the images into groups based on their perception of the iris textures. They found that people consistently identify similar categories and subcategories of irises.

Bowyer *et al.* [2] investigated people's ability to recognize right and left irises as belonging to the same person or not. Through experiments, they discovered that humans perceive texture similarities that are not detected by automated solutions. As a consequence, they can correctly guess, with only three seconds viewing, if left and right eyes belong to the same individual. When evaluating near-infrared images of the whole eye, subjects achieved an accuracy of 86% in the task at hand. When evaluating images with iris portions masked out, subjects achieved an accuracy of 83%, by relying only on the periocular parts of samples. Subjects' ratings of image pairs were collected using a five-level scale, ranging from (1) "same individual (certain)", (2) "same individual (likely)", (3) "uncertain", (4) "different people (likely)", to (5) "different people (certain)".

In a similar fashion, Hollingsworth *et al.* [10] investigated people's skills in deciding if two different iris images depict the eyes of twin siblings or not. Contrary to the typical identity verification pipeline [6], which the authors reported as being useless for the task at hand, human examin-

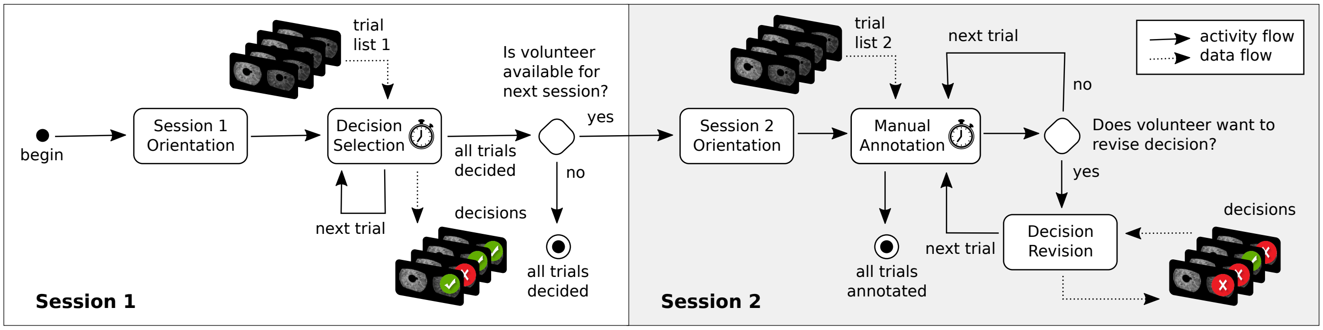


Figure 1: Experimental methodology overview. Rounded rectangular boxes denote subjects' activities, solid arrows represent their precedence, and dashed arrows denote data flow. Experiments always begin with *Session 1*, namely the annotation-less experimental part. Subjects available for *Session 2* then participate in the annotation-driven experimental part.

ers could reach an accuracy of 81% when spending only three seconds analyzing pairs of near-infrared segmented iris images. The accuracy dropped to 76.5% when only periocular regions were available. Again, subjects' responses were collected using a five-level rating. Hollingsworth *et al.* [11] present the combined findings of [2] and [10].

Shen and Flynn [17] asked people to manually annotate *iris crypts*, oval-shaped iris regions with strong edges and darker interior, over near-infrared images. Using annotation software, subjects were asked to outline the borders of the crypts, finding "easy" and "challenging" samples, depending on the clarity of crypts. Presented images comprised strip-normalized iris images, with non-iris-texture regions masked out. The aim of the research was to figure out the utility of crypts for developing more human-interpretable iris-based identity verification. For that, they assessed the repeatability of annotated crypts across subjects, finding that it was possible in the case of "easy" samples.

McGinn *et al.* [13] assessed the performance of human examiners in iris-based identity verification. For that, they asked subjects to classify pairs of irises as either genuine (two images depicting the same eye) or impostor (two images depicting different eyes), again with a five-level rating scale. Presented images comprised near-infrared samples, containing whole eyes of either close-age and same-ethnicity unrelated individuals, or of identical twins. They concluded that identical twins pose a challenge to human performance. In spite of that, the overall accuracy was very high: 92% of the time subjects were successful in classifying iris pairs. Finally, results suggested that subjects improved skills as they gained experience.

Guest *et al.* [9], in turn, investigated the performance of humans in deciding if two distinct infrared whole-eye images depict the same eye or not. In the experiments, subjects presented an overall decision accuracy of 83.2%.

Table 1 summarizes these previous works and the work described in this paper. To our knowledge, there are no other publications about human examination of iris images.

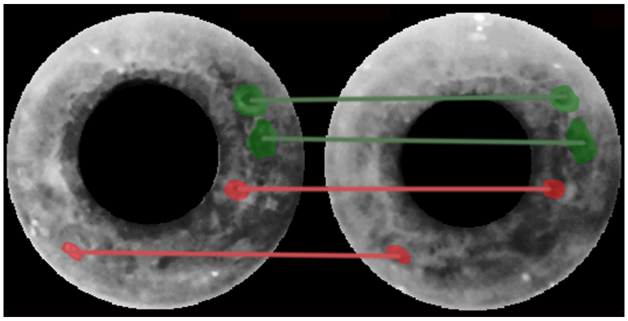


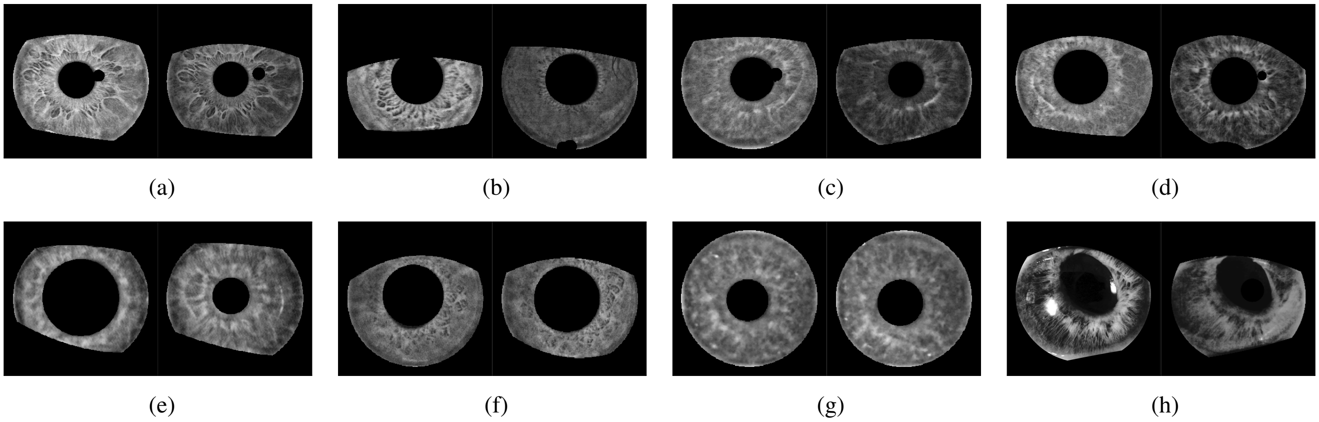
Figure 2: An example of manual annotation containing matching (in green) and non-matching (in red) regions between two post-mortem irises.

3. Experimental Setup

The experimental setup is described in five parts. In Sec. 3.1, we introduce the two-session experimental methodology, while in Sec. 3.2 we explain the chosen categories of iris pairs, including data sources and image preprocessing. In Sec. 3.3 and 3.4, respectively, we describe the experiments conducted at the University of Notre Dame and at NASK headquarters. Finally, in Section 3.5, we detail the experimental setup for employing *OSIRIS* [15], which contains an open-source implementation of Daugman's method for iris recognition [6], and *IriCore* [12] and *MIRLIN* [8], two state-of-the-art solutions for iris recognition. The idea is to provide, along with the performance of humans, the results of fully automated strategies.

3.1. Experimental methodology

We propose a two-session experimental method that allows humans to perform identity verification through the examination of iris patterns. For that, we collect subjects' decisions on whether iris image pairs depict the same eye or not. In the first session, subjects are expected only to provide their decision, with no need for clarification. In the second session, subjects are asked to provide a manual an-



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3: Examples of iris pairs presented to subjects: (a) pair with the same iris generating *low* Hamming distance, (b) different irises generating *high* Hamming distance, (c) pair with the same iris image generating *high* Hamming distance, (d) different irises generating *low* Hamming distance, (e) the same iris before and after *visible light stimulation*, (f) different irises of *identical twins*, (g) the same *post-mortem* iris captured five and 16 hours after death, (h) the same disease-affected iris captured in two different sessions.

notation of the image regions that they see as matching or diverging between the pair of iris images, in order to justify their classification of the image pair. This serves as an effort to make them more conscious about the task at hand. Fig. 1 provides an overview of the proposed experimental method, with the activities that we envision for each subject.

Session 1 starts with an orientation and signing of the consent form (*Session 1 Orientation* in Fig. 1). The subject then views a sequence of image pairs and judges whether or not they represent the same eye (*Decision Selection* in Fig. 1). The list of trials (*trial list 1* in Fig. 1) is previously generated through a pre-processing step, in which irises are selected and genuine and impostor pairs are created. Details about the selected samples are presented in Sec. 3.2. Decisions are then recorded and stored for further analysis. Subjects are given as much time as they need to make a decision, and the decision times are collected for each trial (explaining the chronometer icons in Fig. 1). This first session was followed in experiments at two different institutions, which allowed us to get more diverse results across distinct subject populations.

Session 2 is an annotation-driven part of the experiment that subjects are asked if they have time to complete. Subjects who elect to do this session receive additional instructions on how to provide manual annotation. A subset of iris image pairs used in the first session is selected for annotation in this second session (*trial list 2* in Fig. 1). Details about the selected samples are given in Sec. 3.2. For each trial, the subject is asked to annotate and connect matching and non-matching regions between the two irises. Subjects are also given access to their decision made in *Session 1*, and allowed to change such decision. As in *Session 1*, subjects can spend as much time as they need in this task, and the time intervals are recorded. Fig. 2 depicts an example of manual annotation of matching and non-matching regions between two postmortem irises.

3.2. Dataset of iris image pairs

To assess the influence of different conditions on the accuracy of human iris recognition, we conducted experiments with six categories of irises, which are either commonplace or reportedly known to pose challenges to automated systems or human examiners:

1. ***Healthy and easy:*** images depicting apparently healthy eyes that pose no challenge to the OSIRIS iris recognition software used in this study [15]. That is, the case of genuine pairs generating low Hamming distances between them (Fig. 3(a)), or of impostor pairs whose iris codes yield large Hamming distances (Fig. 3(b)).

2. ***Healthy but difficult:*** apparently healthy eyes that pose challenges to the OSIRIS software. That is, the case of genuine pairs generating unexpectedly large Hamming distances (Fig. 3(c)), or of impostor pairs generating unexpectedly small Hamming distances (Fig. 3(d)).

3. ***Large difference in pupil dilation:*** images of the same eye with significantly difference in pupil dilation, as representatives of the natural iris transformations that occur due to variations in environment lighting (Fig. 3(e)).

4. ***Twins:*** images depicting different eyes, one from each of a pair of identical twins, which are reportedly recognizable by humans, in opposition to being indifferent to automated systems [11] (Fig. 3(f)).

5. ***Post-mortem:*** images depicting either the same or different eyes, captured from deceased individuals, which are known to be surprisingly useful for iris recognition [22] (Fig. 3(g)).
6. ***Disease-affected:*** images depicting the same eye, which suffers from varied eye diseases that may deteriorate the recognition reliability of automated systems [21] (Fig. 3(h)).

Fig. 4 illustrates the distributions of genuine and impostor comparison scores generated by OSIRIS to image pairs of healthy eyes. This information was used to select "easy" and "difficult" cases. Additionally, we generated both genuine and impostor pairs for disease-affected and post-mortem eyes. With respect to twins' samples, it was obviously not possible to generate genuine pairs. To balance the number of impostor and genuine trials, we did not generate impostor pairs from images presenting a large difference in pupil size. Also, when generating the genuine and impostor pairs of healthy, post-mortem, and disease-affected irises, we neither mixed different categories, nor created pairs of images that were captured on the same day.

Given that our intent was to focus on the iris texture and that the dataset was very diverse, we manually segmented all the images and masked out the regions that should not be used by subjects in their judgment, such as eyelashes, eyelids, specular reflections, and severe effects from disease or post-mortem deterioration (*e.g.* corneal wrinkles). In addition, contrast-limited adaptive histogram equalization (CLAHE [16]) was used to enhance contrast for image display, as illustrated in Fig. 3.

Images of healthy eyes were collected from the *ND-CrossSensor-Iris-2013* dataset [4]. Disease-affected iris images were picked from *Warsaw-BioBase-Disease-Iris v2.1* database [21]. Post-mortem iris images were selected from *Warsaw-BioBase-Post-Mortem-Iris v1.0* dataset [22]. Iris images of twins and images presenting high difference in pupil dilation were selected from datasets of the University of Notre Dame, including the one used by Hollingsworth *et al.* [11].

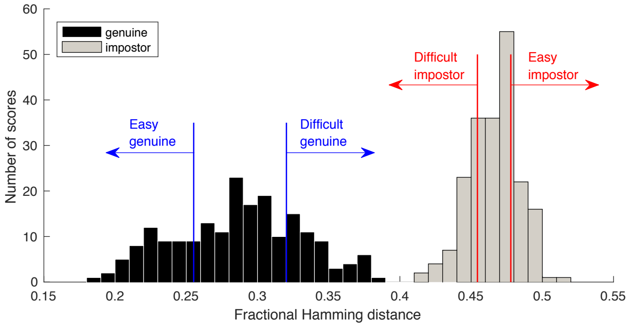


Figure 4: Distributions of the fractional Hamming distance between iris codes. Codes were calculated using the OSIRIS software. These distributions were used to select "easy" and "difficult" cases to use in experiments.

3.3. Notre Dame Experiments

Custom software was prepared for both annotation-less and annotation-driven sessions. In the annotation-less *Session 1*, 86 adult individuals (between 18 and 65 years old) from the university community (including students, staff, and faculty) volunteered to participate, with no constraints related to gender and ethnicity. All were subject to the same protocol, approved by the internal academic *Human Subjects Institutional Review Board*. Each volunteer was asked to evaluate a set of 20 iris image pairs, which always contained the following distribution of image pairs, presented in randomized order for each subject:

- four healthy easy pairs, with two impostor and two genuine samples;
- four healthy difficult pairs, with two impostor and two genuine samples;
- four genuine pairs of irises with large difference in pupil dilation;
- four impostor twins' pairs;
- four genuine post-mortem pairs.

In each trial, the software displayed a pair of iris images and asked the subject to select one of the following: *"1. same person (certain)"*, *"2. same person (likely)"*, *"3. uncertain"*, *"4. different people (likely)"*, and *"5. different people (certain)"*.

Fig. 5 depicts the interface of the software, showing one of the 20 trials and possible answers. Each subject could spend as much time as necessary before selecting the answer. The following trial was only displayed after the acceptance of the current selection. In total, the software used 20 disjoint sets of 20 image pairs, leading to 400 available trials composed from 800 manually segmented iris images. As a consequence, each trial set was submitted, on average, to four subjects, but never with the same order, since, for each individual, the tool randomly shuffled the 20 trials to be presented. On average, each subject spent seven minutes participating in the annotation-less first session.

In the second session, 85 of the 86 subjects continued to provide manual annotations for both matching and non-matching regions between irises. For each person, from the previous 20 trials they have opined, the software automatically selected 10 trials in an iris-category balanced manner. In addition, the tool tried to present, if possible, at least one hit and one miss of each category. Subjects were allowed to annotate as many pairs of matching or non-matching regions as they want, however annotating from 2 to 5 feature pairs was recommended. Also, they were advised to avoid using masked out black regions. Fig. 6 depicts the annotation interface of the custom software, showing an example annotation. Subjects could freely change and update their decisions while annotating a particular pair. Each subject spent between 10 and 20 minutes participating in the annotation-driven session.

3.4. NASK Experiments

These experiments consisted of only the annotation-less (first) session. In total, 28 subjects (different from those attending the experiments at Notre Dame) participated, committing themselves to the exact same protocol, locally approved by the NASK data-protection office. Each subject was asked to evaluate a set of 24 image pairs, which always contained the following setup:

- five genuine post-mortem iris pairs;
- five impostor post-mortem iris pairs;
- five genuine disease-affected iris pairs;
- five impostor disease-affected iris pairs;
- four repeated pairs, each one being randomly selected from one of the above subsets.

In each trial, a custom software displayed a pair of iris images and asked the subject to provide a binary decision on whether images depicted the same eye or not. For the 28 subjects, the software had 10 disjoint sets of 24 trials available, leading to a total of 240 available iris pairs. As a consequence, each trial set was presented to at least two subjects. On average, each subject spent 15 minutes participating in this experiment.

3.5. OSIRIS, IriCore, and MIRLIN Setup

We used OSIRIS [15], IriCore [12], and MIRLIN [8] as representatives of automated iris-matching algorithms. OSIRIS implements a Daugman-style solution [6], therefore relying on Gabor filters to generate iris codes that are compared through fractional Hamming distance. IriCore and MIRLIN, in turn, comprise two commercial iris recognition solutions, which together represent the current state of the art in this area. All three methods generate genuine comparison scores that should be close to zero. Since OSIRIS does not apply Daugman's score normalization [7], we assumed an acceptance threshold equal to 0.32, as earlier suggested in [6]. With respect to IriCore, we adopted an acceptance threshold of 1.1, as suggested in its documentation. For MIRLIN, in turn, we used a threshold of 0.2, as recommended in [5].

4. Results

Table 2 shows the performance of human subjects in assessing the comparison type (genuine or impostor) of iris pairs during the first annotation-less session of experiments, combined for both *Notre Dame* and *NASK* experiments. Reported accuracy expresses the percentage of correctly classified trials, across all subjects. A subject's response was considered correct, or a *hit*, if the subject selected "same person", with either "certain" or "likely" as their confidence, and the image pair was in fact from the same iris. A "different people" response, with either "certain" or "likely" confidence, was considered correct, or also a *hit*, if the image pair was in fact of different irises. All other responses, including the "uncertain" option, were treated as a mistake, *i.e.* a *miss*. Given that people's decisions were discrete, we do not provide ROC-like curves in the results. In addition, OSIRIS, IriCore, and MIRLIN were used according to their respective recommended operating points.

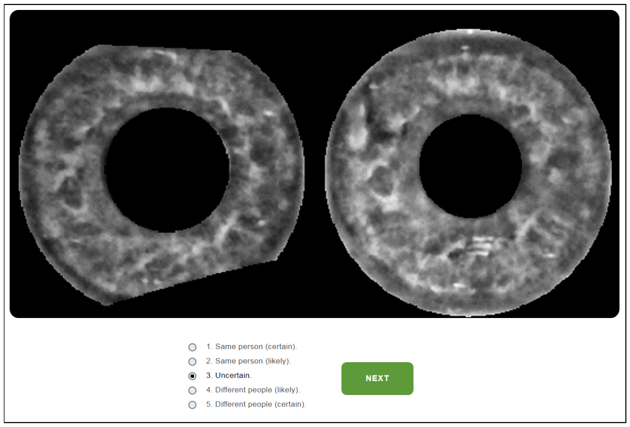
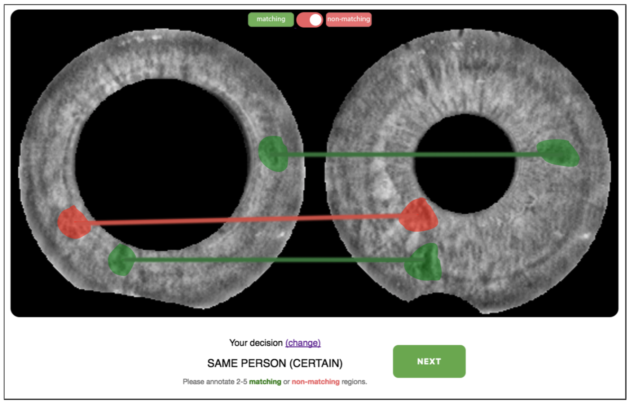

○ 1. Same person (certain).
○ 2. Same person (likely).
● 3. Uncertain.
○ 4. Different people (likely).
○ 5. Different people (certain).

Figure 5: An example screen of a *Session 1* trial. To proceed to the next image pair, the subject had to select one decision and click "Next".



Your decision (change)
SAME PERSON (CERTAIN)
Please annotate 2-5 matching or non-matching regions.

Figure 6: An example screen of a *Session 2* trial. Subjects were allowed to freely annotate and connect matching (green) or non-matching (red) features over the two presented irises, to support their decision. They could update their previous decision by clicking on the "change" option.

Overall, subjects were correct nearly 70% of the time, while the best algorithm (IriCore) achieved a higher overall accuracy of 89.06%. Both human subjects and software tools were more successful in identifying impostor than genuine pairs, with all the three tools not making a

Table 2: Annotation-less performance of human subjects in iris identification. Subjects were only asked to select their decisions. For comparison sake, we report results of the OSIRIS, IriCore, and MIRLIN software, with acceptance thresholds equal to 0.32, 1.1, and 0.2, respectively.

	Iris category	Accuracy (%)			
		Humans	OSIRIS	IriCore	MIRLIN
Genuine pairs	Healthy easy	91.28	95.00	100.00	97.50
	Healthy difficult	79.07	90.00	97.50	97.50
	Pupil-dynamic	43.90	61.25	95.00	97.50
	Post-mortem	51.95	33.57	73.57	47.14
	Disease-affected	70.80	25.00	53.33	25.00
	Combined	**60.60**	**58.86**	**80.56**	**65.83**
Impostor pairs	Healthy easy	84.30	100.00	100.00	100.00
	Healthy difficult	76.16	100.00	100.00	100.00
	Twins	55.81	100.00	100.00	100.00
	Post-mortem	83.90	100.00	100.00	100.00
	Disease-affected	91.00	100.00	100.00	100.00
	Combined	**74.41**	**100.00**	**100.00**	**100.00**
Overall		**70.11**	**79.43**	**89.06**	**80.78**

single mistake in recognizing impostors. Nonetheless, in the particular case of genuine samples, humans performed on par with OSIRIS and MIRLIN, exceeding their results in face of genuine post-mortem samples. Moreover, people presented a much superior performance when analyzing disease-affected samples, surpassing all the three tools. Indeed, in such cases, software was always biased towards classifying samples as impostors, justifying close-to-chance (IriCore) or worse-than-chance (OSIRIS, MIRLIN) accuracies for genuine pairs, and perfect hit rates for impostors.

Subjects performed better than chance, *i.e*. the accuracy was higher than 50%, in most of the iris categories. However, for the subset composed of iris images with large difference in pupil dilation, the accuracy was only 43.9%. Variations in pupil dilation were the most challenging cases for subjects, impairing their ability to recognize different versions of the same eye.

Subjects also had difficulty in analyzing post-mortem iris pairs. In general, they tended to classify post-mortem samples as impostors, leading to a low accuracy in genuine cases (51.95%, slightly better than chance), and a higher accuracy in impostor cases (83.90%). Similar to the observations of Bowyer *et al.* [2], irises of twins also revealed themselves as challenging for people, but easy for automated solutions. Among impostor samples, they are the category where subjects had the lowest accuracy (55.81%).

Fig. 7 shows the subjects' confidence level when classifying the iris pairs, during the first session of *Notre Dame* experiments. Bars depict the normalized frequencies of each response; as a consequence, they sum up to 1.0. According to the adopted groundtruth color notation, black-bar regions represent genuine pairs and gray regions represent impostor pairs. Therefore, black regions are expected to occur mostly on the left side of the chart (which is respective to people's claims of seeing genuine pairs), while gray regions are expected to occur mostly on the right side (which is respective to impostor pairs). Gray regions on the left side and black regions on the right side are all errors, as well as any answer in the center ("uncertain" option).

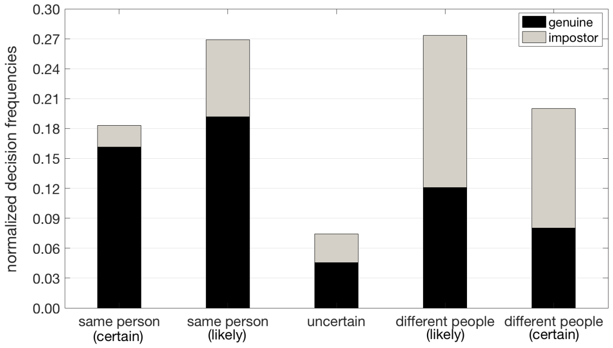


Figure 7: Normalized frequencies of the decisions of the 86 subjects of the *Notre Dame* experiments, according to their decisions and groundtruth. Genuine iris pairs are represented by black bars, while impostor pairs are represented by gray ones. In an ideal classification output, black bars should happen only on the left part of the chart, while gray bars should happen only on the right side.

As one might observe, the "uncertain" option was the least selected choice (being taken in less than 8% of the trials), agreeing with the reports of McGinn et al. [13]. Among the "same person (certain)" answers (18% of all answers), one in each nine (nearly 11%) was wrong. In opposition, among the "different people (certain)" answers (20% of all answers), nearly one third was wrong, revealing more errors in people's convictions of seeing impostor pairs. In accordance to the data presented in Table 2, this indicates that people had more problems in recognizing genuine pairs, wrongly classifying many of them as impostors with high confidence.

Table 3 provides a comparison of the performances of 85 subjects who participated in the first session of *Notre Dame* experiments (without annotations), and were able to return to the second session, when they were asked to provide annotations for the matching and non-matching regions between the irises of each presented pair. Reported accuracy regards the correctness of the decisions for the subset of iris pairs they have already seen in the first session.

The annotation feature helped subjects to improve their decisions in all iris categories, except for impostor healthy difficult pairs (in which accuracy slightly dropped from 80.00% to 78.88%). Iris image pairs with large difference in pupil dilation were the category that benefited the most from annotations, with an improvement in accuracy from 41.18% to 52.35%. Accuracy for post-mortem cases also was significantly improved.

Table 3: Comparison of 85 subjects' accuracy when performing iris identification without annotations versus with annotations, over exactly the same iris samples.

		Accuracy (%)	
Pair class	Iris category	without annotations	with annotations
Genuine	Healthy easy	87.06	96.47
	Healthy difficult	75.29	84.71
	Pupil-dynamic	41.18	52.35
	Post-mortem	45.29	54.12
	Combined	**55.88**	**65.69**
Impostor	Healthy easy	85.88	90.59
	Healthy difficult	80.00	78.82
	Twins	59.41	60.59
	Combined	**71.18**	**72.65**
Overall		**62.00**	**68.47**

Fig. 8 details how decisions were revised when subjects provided manual annotation. Black bars express the absolute number of revised decisions that were worsened (*i.e.*, a correct decision after the first session, updated to an incorrect decision during the second session). Conversely, gray bars express the number of revised decisions that were fixed (*i.e.*, they were originally a miss after first session, but then were updated to a correct decision during the second one). In general, more decisions were fixed (74 decisions) than worsened (19 decisions). Interestingly, post-mortem samples presented only improvements (15 decisions were revised), suggesting that people perceived new details on them while providing annotations. Twins' samples, in turn, once more revealed how confusing they appear to people; 11 incorrect decisions were corrected, but 9 correct decisions were changed to be incorrect.

Last but not least, we could not find correlation between time spent by subjects and accuracy. Fig. 9 depicts the distributions of time spent by subjects to decide each trial in the first annotation-less sessions (combining both *Notre Dame* and *NASK* experiments, shown in the left side of the chart), and to annotate each trial in the second annotation-driven sessions (shown in the right side of the chart). As one might observe, regardless of iris pairs being genuine or impostor, and of decisions being hits or misses, distributions were not significantly different. As expected, annotation-driven trials were, on average, longer than annotation-less trials.

5. Conclusions

This paper presents results of a unique study estimating the accuracy of human subjects in comparing iris images of different levels of difficulty, including healthy and disease-affected eyes, and images acquired from cadavers.

The first observation from this study is that we may expect people to be worse than automated iris-recognition methods when comparing healthy eyes. However, they can be better in cases not yet considered in the development of automated algorithms, such as eyes suffering from diseases or post-mortem deformations.

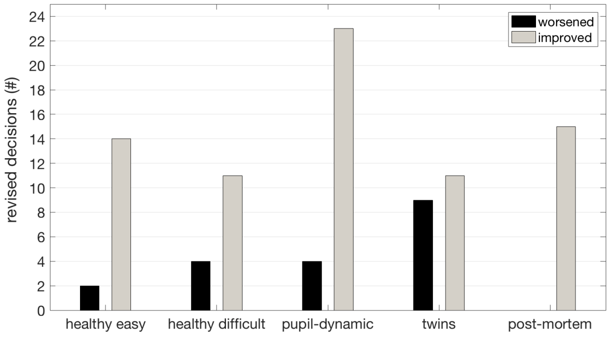


Figure 8: Numbers of revised iris pairs grouped by iris category. While manually annotating an iris pair, subjects could change their decision, either improving it (*i.e.* making it right, depicted in gray), or worsening it (*i.e.* making it wrong, depicted in black).

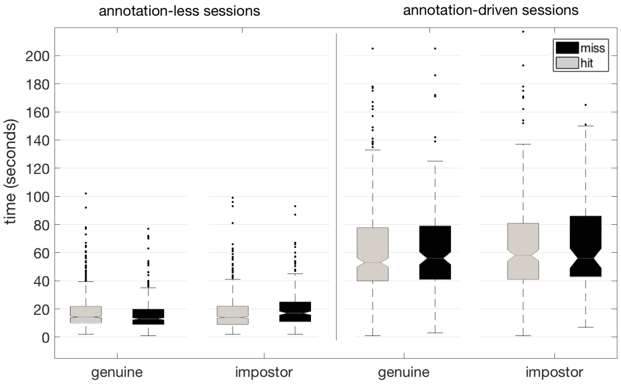


Figure 9: Times spent by subjects to answer each trial. Left side: times spent in the annotation-less sessions. Right side: times spent in the annotation-driven sessions.

The second observation is that human examiners on average improve their accuracy when they are asked to annotate matching and non-matching features that support their decision. Although this improvement is larger for genuine pairs than for impostor pairs, it still suggests that a comparison of iris images performed by humans should be organized in a way that allows them to annotate the features they are using in their judgment. As future work, this may help in the development of a method for the examination and documentation of irises that is analogous to *ACE-V* [1], originally proposed for fingerprints.

The third observation is that different categories of iris images result in significantly different performance of human subjects. Three categories of samples that seem to be particularly challenging to humans: irises of identical twins, iris images showing large difference in pupil dilation, and irises of deceased individuals.

References

[1] D. Ashbaugh. The Identification Process. In *Quantitative-qualitative friction ridge analysis: an introduction to basic and advanced ridgeology*, chapter 4, pages 87–148. CRC Press, 1999. 8

[2] K. Bowyer, S. Lagree, and S. Fenker. Human Versus Biometric Detection of Texture Similarity in Left and Right Irises. In *IEEE Intl. Carnahan Conference on Security Technology (ICCST)*, pages 323–329, 2010. 1, 2, 3, 7

[3] J. Chen, F. Shen, D. Chen, and P. Flynn. Iris Recognition Based on Human-Interpretable Features. *IEEE Transactions on Information Forensics and Security*, 11(7):1476–1485, 2016. 1

[4] Computer Vision Research Laboratory at the University of Notre Dame. Collection ND-CrossSensor-Iris-2013. Available at https://www3.nd.edu/~cvrl/LicenseAgreements/UNDLicenseAgreementCrossSensorIris2013.pdf, 2013. Accessed Mar 13, 2018. 5

[5] A. Czajka, W. Kasprzak, and A. Wilkowski. Verification of iris image authenticity using fragile watermarking. *Bulletin of the Polish Academy of Sciences, Technical Sciences*, 64(4):807–819, 2016. 6

[6] J. Daugman. How Iris Recognition Works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):21–30, 2004. 2, 3, 6

[7] J. Daugman. Probing the uniqueness and randomness of IrisCodes: Results from 200 billion iris pair comparisons. *IEEE Proceedings*, 94(11):1927–1935, 2006. 1, 6

[8] FotoNation. MIRLIN Iris Recognition. Available at https://www.fotonation.com/products/biometrics/iris-recognition/, 2018. Accessed on June 30, 2018. 3, 6

[9] R. Guest, H. He, S. Stevenage, and G. Neil. An Assessment of the Human performance of Iris Identification. In *IEEE Intl. Conference on Technologies for Homeland Security (HST)*, pages 623–626, 2013. 1, 2, 3

[10] K. Hollingsworth, K. Bowyer, and P. Flynn. Similarity of Iris Texture between Identical Twins. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 22–29, 2010. 1, 2, 3

[11] K. Hollingsworth, K. Bowyer, S. Lagree, S. Fenker, and P. Flynn. Genetically identical irises have texture similarity that is not detected by iris biometrics. *Elsevier Computer Vision and Image Understanding*, 115(11):1493–1502, 2011. 1, 3, 4, 5

[12] Iritech, Inc. IriCore. Available at http://www.iritech.com/products/software/iricore-eye-recognition-software, 2018. Accessed on June 30, 2018. 3, 6

[13] K. McGinn, S. Tarin, and K. Bowyer. Identity Verification Using Iris Images: Performance of Human Examiners. In *IEEE Intl. Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, 2013. 1, 2, 3, 7

[14] H. Nissenbaum. Computing and Accountability. *ACM Communications*, 37(1):72–80, 1994. 1

[15] N. Othman, B. Dorizzi, and S. Garcia-Salicetti. OSIRIS: An open source iris recognition software. *Elsevier Pattern Recognition Letters*, 82(2):124–131, 2016. 3, 4, 6

[16] S. Pizer, P. Amburn, J. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. Zimmerman, and K. Zuiderveld. Adaptive histogram equalization and its variations. *Elsevier Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987. 5

[17] F. Shen and P. Flynn. Are Iris Crypts Useful in Identity Recognition? In *IEEE Intl. Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, 2013. 1, 2, 3

[18] L. Stark, K. Bowyer, and S. Siena. Human Perceptual Categorization of Iris Texture Patterns. In *IEEE Intl. Conference on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–7, 2010. 1, 2

[19] B. Steele. Cornell Tech will help make computers 'accountable'. Available at http://news.cornell.edu/stories/2017/09/cornell-tech-will-help-make-computers-accountable, 2017. Accessed on February 13, 2018. 1

[20] The European Parliament and The Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L(119):1–88, 2016. 1

[21] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Database of iris images acquired in the presence of ocular pathologies and assessment of iris recognition reliability for disease-affected eyes. In *IEEE Intl. Conference on Cybernetics (CYBCONF)*, pages 495–500, 2015. 5

[22] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Human iris recognition in post-mortem subjects: Study and database. In *IEEE Intl. Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, 2016. 5

[23] S. Wachter, B. Mittelstadt, and L. Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *Oxford University Press International Data Privacy Law*, 7(2):76–99, 2017. 1