Genetically Identical Irises Have Texture Similarity That Is Not Detected By Iris Biometrics

Karen Hollingsworth, Kevin W. Bowyer, Stephen Lagree, Samuel P. Fenker, and Patrick J. Flynn

Computer Science and Engineering Department, University of Notre Dame, Notre Dame, IN 46556

Abstract

As the standard iris biometric algorithm “sees” them, the left and right irises of the same person are as different as irises of unrelated people. Similarly, in terms of iris biometric matching, the eyes of identical twins are as different as irises of unrelated people. The left and right eyes of an individual or the eyes of identical twins are examples of genetically identical irises. In experiments with human observers viewing pairs of iris images acquired using an iris biometric system, we have found that there is recognizable similarity in the left and right irises of an individual and in the irises of identical twins. This result suggests that iris texture analysis different from that performed in the standard iris biometric algorithm may be able to answer questions that iris biometrics cannot answer.

Keywords: ocular biometrics, iris recognition, periocular recognition, twins, genetically identical irises, texture analysis.

1. Introduction

Biometric systems aim to differentiate between individuals using physical or behavioral characteristics. Physical appearance is affected by genes and by environment. Identical twins share the same or nearly the same genetic code. Consequently, physical traits determined primarily by genotype are not effective discriminators between identical twins. Body type, voice, and face are strongly affected by genes and are very similar for identical twins.

Some physical traits are affected by environment around the fetus during gestation. In the case of fingerprints, genes determine the general pattern, but the position in the uterus and the flow of amniotic fluid around the fetus are different for each finger. The fine details of fingerprints are affected by this changing microenvironment.

Jain et al. [3] tested how minutiae information in identical twins was related in the context of an automatic fingerprint-based authentication system. They found that for a minutiae-based fingerprint matcher, twins’ fingerprints are generally more similar than nontwins’ fingerprints. However, automatic fingerprint verification can still be used to distinguish between identical twins without drastic degradation in performance. Jain et al. found that the similarity in twin fingerprint comparisons is related to the general fingerprint pattern determined by the twins’ genotype. Fingerprints are divided into five classes: whorl, right loop, left loop, arch, and tented arch. Jain et al. manually classified 94 pairs of identical twin fingerprints. The fraction of identical twin pairs in their data whose index fingers had the same class label was 0.775, compared with 0.2718 for two index finger prints randomly selected from a large database. The degradation in performance for identical twin comparisons was the same order of magnitude as the degradation in performance for unrelated people who had the same fingerprint class label. The authors concluded that similarity observed in identical twins is due to the high class correlation in their fingerprint types.

Irises follow a developmental pattern similar to fingerprints [3]. It has been claimed that eye color and general appearance are determined by genotype while small texture details are epigenetic [4]. Daugman and Downing [4] tested how iris texture

This is an extended and revised version of two papers: “Human versus biometric detection of similarity in left and right irises” [1] © 2010 IEEE and “Similarity of iris texture between identical twins” [2] © 2010 IEEE.

information in identical twins was related in the context of a canonical iris-based authentication system. The canonical approach to iris biometrics applies a set of Gabor filters at predefined locations on the iris image, creates a binary iris code by taking only the phase of the Gabor filter results, and measures the difference between two iris codes as the fractional Hamming distance. In a test with six pairs of monozygotic twins, Daugman and Downing found that for this type of iris biometrics system, similarity in twins' irises was statistically indistinguishable from similarity in nontwins' irises. Another source of genetically identical irises is found in left and right irises of the same person. Daugman and Downing found that left and right irises are different from each other.

It is generally accepted that the left and right irises of the same person have uncorrelated iris codes, and also that the irises of identical twins have uncorrelated iris codes [5, 6, 7]. In this work, we reproduce the result that according to an iris biometrics authentication system, comparisons between left and right irises and twins' irises are statistically indistinguishable from comparisons of unrelated people's irises. We also investigate how human observers view the overall iris texture pattern. While an iris biometrics system sees no similarities in genetically identical irises based on the fine epigentic texture details, we find that human observers detect similarities in genetically identical irises.

One prior paper has investigated how humans perceive similarities in irises; Stark et al. [8] studied how humans group iris textures into categories. To our knowledge, there is no other related work on the abilities of humans to perceive similarities or differences in iris texture patterns. This is a relatively new and unexplored area of research.

2. Iris Codes of Genetically Identical Irises are Uncorrelated

This section presents results of two experiments. The first experiment looks at biometric matching of the left and right iris images of the same person. The second experiment looks at matching iris images of identical twins. In both cases, comparisons between genetically identical irises yield similar scores as comparisons between unrelated irises.

2.1. Biometric performance between left and right irises

Our experiment with matching left and right irises of the same person uses images selected from the ND-IRIS-0405 iris image dataset [9]. This dataset contains the images used in the Iris Challenge Evaluation program [10]. These images were acquired at the University of Notre Dame using an LG 2200 EOU enrollment camera [11]. The images are 640x480 in size, and contain the eye and the immediately surrounding area. They are 8-bit intensity images, taken under near-infrared illumination.

Images for each subject in the dataset were manually reviewed to select images that have the iris in good focus, that do not have too much iris occlusion, and in which the subject is not wearing cosmetic or gas-permeable contact lenses [12]. Some subjects may be wearing normal soft contact lenses. One left and one right iris image were selected for each of 322 subjects for use in this experiment. The two images selected for a given subject were not necessarily acquired on the same day, and do not necessarily have the same pupil dilation.

The selected left and right iris images were matched against each other using our enhanced version of the IrisBEE iris biometrics system made available through the ICE program [10]. This software segments the iris region using circles for the pupil-iris and iris-sclera boundaries, and employs active contours to detect the eyelid occlusion boundaries. The software unwraps the iris region to a rectangular normalized image then uses log-Gabor wavelets to generate an iris code. The result of a match between two iris codes is reported as a fractional Hamming distance. The scores are then normalized using the score normalization technique proposed by Daugman [13].

The two impostor distribution histograms appear in Figure 1. The same-person impostor histogram contains the normalized Hamming distances for 322 matches between left and right irises of the same person. The different-person impostor histogram contains the normalized Hamming distances for $\binom{322}{2}$ matches between left and right irises of different people. The two histograms are clearly very similar, with the peaks occurring at essentially the same Hamming distance. A Kolmogorov-Smirnov [14] test to compare the histograms finds no evidence to reject the null hypothesis that the two sets of scores represent the same distribution (p-value 0.6873).

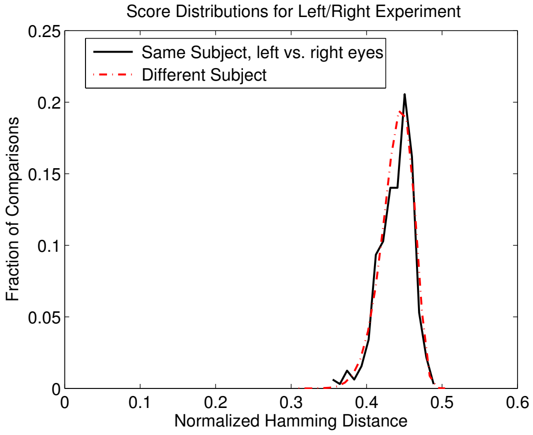

Figure 1: The distribution of iris biometric scores between left and right eyes of the same subject is statistically indistinguishable from the distribution of scores from left and right eyes of different subjects, as noted in [4].

Any bit in an iris code is equally likely to be a 1 or a 0, so the expected fraction of agreeing bits between two unrelated iris codes is 0.5. In practice, the average impostor Hamming distance is less than 0.5 because of the way that we account for image rotation. Two iris images may be rotated within the image plane but the iris code is computed only at a single orientation. During the matching step, one iris code is shifted to represent rotation of the iris image, then compared with a second iris code. The best match of n rotations is taken to be the appropriate rotation. The "best of n" test skews the Hamming distance distribution so that the mean impostor Hamming distance is less than 0.5. Figure 1 shows that comparisons between left and right eyes of the same subject yield similar Hamming distances to comparisons between unrelated eyes. The impostor distributions in Figure 1 are consistent with the observation by Wang, Tan and Jain [15] that "iris images of left and right eyes are known to be different." Daugman had earlier presented results showing that the left and right irises of the same person do not match any more closely than do the irises of different people [4].

2.2. Biometric performance between twins' irises

Our experiment with matching iris images of twins uses data that was collected at the Twins Days Festival in Twinsburg, Ohio in August of 2009. This festival is billed as "the largest annual gathering of twins in the world" [16]. Using an LG 2200 system, we acquired video clips pf each iris for each person in 76 pairs of self-reported identical twins (152 people), plus an additional 44 people. The self-reported identical twins did not necessarily have any clinical test for verification, and so it is possible that a small fraction are actually very similar-looking fraternal twins.

The analog signal from the LG 2200 camera was digitized using a high bit rate (effectively lossless) compressed MP4 format. The frames from each video clip were automatically processed to reject all frames with average intensity value less than 115, to reject frames with too much high-frequency noise and then to select the ten most in-focus frames. These frames were manually reviewed and frames with any significant artifacts were rejected.

The two impostor distribution histograms appear in Figure 2. The twin impostor histogram contains the normalized Hamming distances for matches between the same (left or right) iris of each person in a pair of identical twins. The different-person impostor histogram contains the scores for matches between eyes of different people. The peaks of the two histograms occur at the same Hamming distance. In this case, a Komogorov-Smirnov test comparing the two histograms finds small but statistically significant evidence to reject the null hypothesis that the two sets of scores are from the same distribution (p-value $< 10^{-4}$). However, we examined the twin comparisons yielding the slightly lower scores and found that they were comparisons with imperfect segmentation. Specifically, some images had undetected specular highlights, and others had elliptical pupils which our software is not designed to handle. If we exclude those images from the comparison, the two distributions are statistically indistinguishable. Thus when segmentation is accurate, we find no evidence that twins' iris codes are more similar than nontwins' iris codes.

According to our current evidence, iris codes are different than fingerprint minutiae in that twins' fingerprint minutiae are correlated [3] but we have yet to find statistically significant evidence of correlation between twins' iris codes. Sun et al. [7] failed to find significant evidence of correlation in a biometrics experiment on iris data from 51 pairs of identical twins. And these results are consistent with observations or suggestions by Flom and Safir [17], Daugman [18], and Wildes [5].

The results of our iris biometrics experiments on matching left and right irises of the same person, and on matching irises of identical twins, are consistent with previous results in the literature [7].

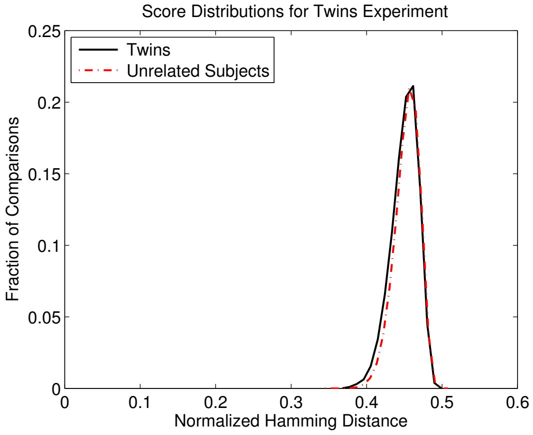

Figure 2: The distribution of iris biometric scores between twins' eyes peaks at the same point as the distribution of scores from non-twins' eyes.

The basic conclusion is that, to current iris biometrics technology, genetically identical iris codes are as different as unrelated people's iris codes. Companies selling biometrics products appropriately focus on the differences in iris texture without considering similarities in iris appearance. L-1 Identity Solutions, a company that licenses the algorithms developed by Dr. John Daugman, advertises that "No two irises are alike. There is no detailed correlation between the iris patterns of even identical twins, or the right and left eye of an individual" [19]. However, this result has sometimes been over-simplified, in effect, to assume that what is seen by current iris biometrics technology is all that there is to see. For example, Wikipedia states that, "Even genetically identical individuals have *completely independent iris textures*" [20] (emphasis added). Our experiments with humans observing pairs of iris images are aimed at determining what similarities exist in the texture patterns of genetically identical irises. Our results show that genetically identical individuals decidedly do **not** have completely independent iris textures.

3. Human Perception of Texture Similarity in Left and Right Irises

We investigated how well humans could verify whether two images in a left/right pair were from the same person. For this study, we used the same left and right iris images as we used in the left/right biometrics test in the previous section.

3.1. Experimental Set-up

This experiment involved displaying pairs of eye images to human participants. We used images from 322 subjects in the ND-IRIS-0405 data set, and wrote custom software to display the images. For a given trial, a left eye image and a right eye image were presented side-by-side on the computer display for three seconds. After three seconds, the display changed and the volunteer was asked to select one of five options to represent their degree of certainty about whether the two images were correctly paired together as representing the same person:

1. Certain that they are a matched left-right pair.
2. Likely that they are a matched left-right pair.
3. Can't tell.
4. Likely that they are NOT a matched left-right pair.
5. Certain that they are NOT a matched left-right pair.

After responding, the program provided feedback on whether the response was correct.

We randomly chose 105 people from the pool of subjects in our iris image data to use in 105 matched-pair trials. We randomly chose another 105 people for the left iris images and an additional 105 people for the right iris images in non-matched-pair trials. No person had an iris image appear in more than one trial. Therefore, we used a total of 315 people to create 210 trials, where half of the trials were matched-pairs and half were non-matched-pairs. Each volunteer in our experiment viewed the same 210 trials, but the order of the presentation of the trials was determined randomly for each participant. Images from additional people were used at the start of the experiment to familiarize the volunteer with the task, by presenting three example matched pairs and two example non-matched pairs. These example trials at the beginning of the experiment were the extent of each participant's training in how to distinguish between pairs of iris images with matching versus non-matching iris texture.

Human subjects participated in these trials under a protocol approved by the Human Subjects Institutional Review Board at the University of Notre Dame. Volunteers for our experiments were recruited from the students and staff at the University of Notre Dame. Volunteers were offered a ten-dollar payment for participating in the experiment, plus an additional ten dollars if they categorized more than 80% of the experimental trials correctly. Each selected an appointment time to perform the experiment. To control the quality of the image display, all participants performed the experiment on the same computer workstation. None of the participants were experienced in iris biometrics.

Our initial experiment showed the whole 640x480 eye image acquired by the LG 2200 camera. From a subjective visual evaluation of the results, it seemed possible that similarity in the appearance of the periocular region could affect volunteers' responses. In order to investigate the relative contribution of the iris texture versus the periocular region, and because periocular biometrics is a developing research area of its own [21, 22, 23], a follow-up experiment was designed to separately evaluate the contributions of the iris region and the periocular region. The follow-up experiment was similar to the initial whole-eye experiment, but with two different image viewing conditions. One condition was images with the periocular area masked out, so that only the iris region was visible. The second condition was images with the iris and pupil region masked out, so that only the periocular region was visible.

We initially intended to use altered versions of exactly those images used in the initial experiment in our follow-up experiment. However, about 10% of those images had an eye corner or other part of the eye that fell outside of the image. These images seemed likely to bias the follow-up experiment toward poor performance for the periocular region, and so we replaced such images with different images of the same eye. Replacement images were selected to have as much periocular region as possible visible in the image.

The iris region and the eyelid occlusion boundaries were manually marked for each image. Two different versions of each image were then produced. One version has the image area outside of the visible iris region blacked out. The other version has only the iris and pupil region blacked out.

Twenty-seven volunteers participated in the whole-eye experiment, and twenty-nine other volunteers participated in the iris-only and periocular-only experiment. The results of all three viewing conditions are presented in this section.

3.2. Can humans determine whether left and right iris images come from the same person?

To find an overall accuracy score we made no distinction based on the tester's confidence level, only on whether they believed a pair to be a matched left/right pair or not. We divided the number of correct responses by the total number of queries to yield an accuracy score. When viewing whole eyes, the average percent correct was 85.7% (standard deviation 4.2%). The minimum score was 160 correct out of 210, or 76.2% and the maximum score was 191 out of 210, or 91.0%.

When less information was shown, accuracy dropped slightly. The average performances on both the iris-only and the periocular-only tests were both about 83%. For the iris test, average performance was 82.5% (standard deviation 4.0%), and for the periocular test, average performance was 82.7% (standard deviation 4.1%).

We used a t-test to evaluate the null hypothesis that humans did not perform differently than random guessing. The resulting p-values for all three conditions were less that 10^{-4}. Thus, we have statistically significant evidence that our testers were correctly matching left and right irises at a rate higher than random guessing. We conclude that humans are able to detect similarities between left and right eye images, even when given only the iris, or only the periocular region.

3.3. How confident were volunteers of their responses?

As mentioned above, volunteers had five different options to select, based on their judgment of whether the images were correctly paired, and how certain they were of their response. Some volunteers responded more confidently than others. On the whole-eye experiment, two volunteers never selected a "certain" response. Similarly, two volunteers in the iris/periocular experiment never selected a certain response. At the other extreme, one volunteer in the whole-eye experiment and two in the iris/periocular experiment always selected certain responses.

Overall, on the whole-eye experiment, volunteers were "certain" of their response 47.9% of the time. They only selected "can't tell" 1.0% of the time. When volunteers saw less information, their confidence was slightly lower on average. Volunteers viewing the iris-only trials were "certain" of their responses 34.1% of the time, and when viewing periocular-only trials, they were "certain" 40.0% of the time. A histogram showing the distribution of responses for all three conditions is shown in Figure 3. The relatively infrequent use of the "can't tell" option could be related to the fact that we of-

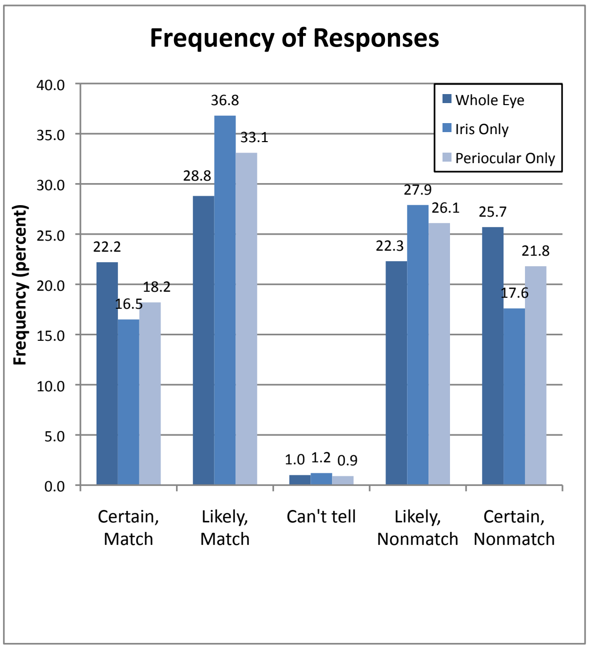

Figure 3: This graph shows the frequency of responses for the five options in our left/right human viewer experiments. Viewers expressed slightly higher confidence in their responses on the whole-eye experiment than on the iris-only or periocular-only experiments.

fered a bonus to participants who got more than 80% correct.

3.4. *Did volunteers score higher when they felt more certain?*

Figure 4 shows the accuracy broken down by response. Viewers scored higher on the subset of trials where they marked that they were certain of their answers. Based on these results, we infer that volunteers correctly judged their relative confidence in their responses.

3.5. *Did volunteers learn as the test progressed?*

For each of the three types of viewing conditions, we computed the difference between accuracy on the second half of the test and the first half of the test. We found this difference for each volunteer, and then computed the average across all participants. On the whole-eye test, 21 of the 27 participants performed better on the second half of the test, one stayed the same, and five performed worse. The average score for the second part was 3.4% higher than the first part (standard deviation 4.2%). The minimum difference was -5.7% and the maximum difference was 9.5%.

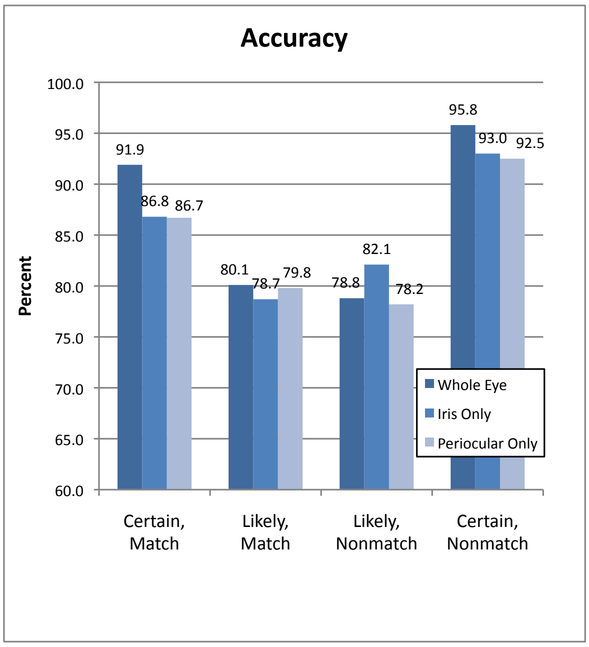

Figure 4: This graph shows the accuracy on the left/right human viewer experiment, broken down by response. Accuracy was higher on the subset of queries where viewers expressed certainty in their responses. Overall, performance was higher in the whole-eye case than on the iris-only or periocular-only cases.

For the iris queries, 19 of the participants improved their score, three stayed the same, and seven did worse. The average improvement was 2.7% (standard deviation 5.5%) with a minimum of -7.6% and a maximum of 19.0%.

For the periocular queries, 17 participants improved, one stayed the same, and 11 did worse. The average improvement was 0.7% (standard deviation 5.8%), with a minimum improvement of -10.5% and a maximum improvement of 15.2%.

The average difference for all three conditions is positive, suggesting that some learning is taking place. We computed one-tailed t-tests to check whether the scores on the second half were statistically significantly higher than the scores on the first test. The results demonstrated that the learning on the whole-eye and the iris queries was significant, but the learning on the periocular queries was not (p-values 1.34×10^{-4}, 6.08×10^{-3}, and 2.54×10^{-1} respectively).

3.6. Is it easier to label a matched left-right pair as a match, than it is to label a nonmatch pair as unmatched?

In all three tests, viewers responded that the images pairs matched slightly more often than they responded that the images came from different people. For the whole-eye and the periocular tests, they responded "matched-pair" about 51% of the time. For the iris test, they responded "matched-pair" about 53% of the time. When queries are broken down by subjects responses, as in Figure 4, performance on the queries where subjects responded match is slightly lower, simply because they used that response more often.

When accuracy is computed for actual match pairs and actual nonmatch pairs, viewers performed slightly better on the actual match pairs for all three viewing conditions.

3.7. Which image pairs were most frequently classified correctly, and which were most frequently classified incorrectly?

Some image pairs were easier to classify than others. Of the 105 matched left-right pairs presented in the whole-eye experiment, 28 of them were correctly classified as a match by all volunteers who participated. An example of one of these easy match pairs is given in Figure 5.

The matching left-right pair that was most frequently misclassified is shown in Figure 6. Only six of the 27 participants correctly determined that this pair was a matching left-right pair. One of the distractors in this image pair is that the eye centers are not aligned vertically. We were able to remove this distractor in the subsequent iris-only test, by cropping iris images to a 300x300 image centered around the center of the pupil. The nonmatching image pair most frequently classified incorrectly in the whole-eye experiment is shown in Figure 7. This image pair shows images from two different Asian subjects that have similar eye shape. Viewing image pairs like this one after the whole-eye experiment suggested the idea that the periocular region might distract participants from focusing on iris texture patterns. This idea led to the iris-only and periocular-only experiments.

Figure 8 shows a difficult match pair from the iris-only viewing condition. The pair was incorrectly classified by 16 of the 29 testers. Figure 9 shows a difficult nonmatch iris pair. Examples of difficult image pairs from the periocular viewing condition are shown in Figures 10 and 11.

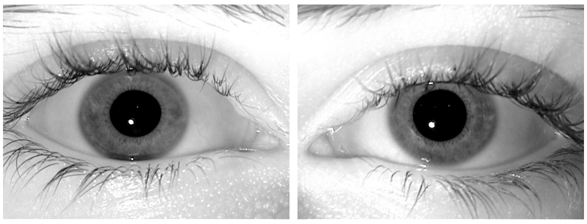

Figure 5: This match pair was an easy query in the left/right human viewer experiment. All volunteers correctly classified this pair as a matching left-right pair. (These are images 04460d351 and 04460d345 from the ND-IRIS-0405 dataset.)

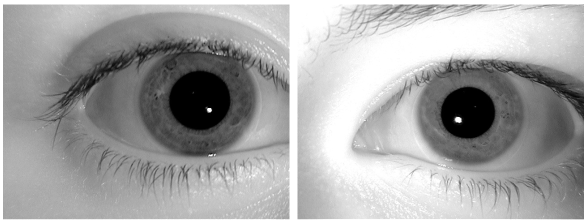

Figure 6: This match pair was a challenging query in the left/right human viewer experiment. 21 of 27 participants incorrectly responded that this pair was from unrelated eyes. (These are images 04408d476 and 04408d457 from the ND-IRIS-0405 dataset.)

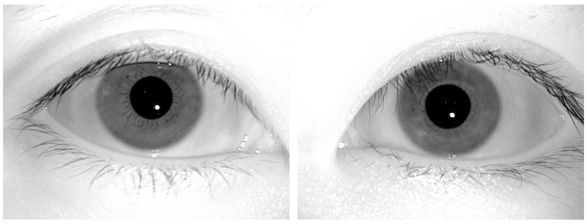

Figure 7: This nonmatch pair was a challenging query in the left/right human viewer experiment. 23 of 27 participants incorrectly responded that this image pair was from the same person, making this the pair that most frequently generated an incorrect response. (These are images 04633d493 and 04851d803 from the ND-IRIS-0405 dataset.) Figure reprinted from Bowyer et al. [1] ©2010 IEEE.

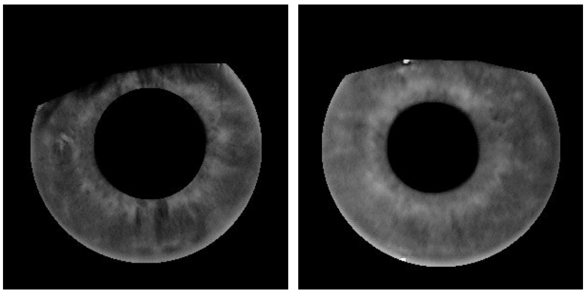

Figure 8: This example match pair from the iris-only experiment shows left and right irises of the same person. This pair of images was incorrectly classified by 16 of 29 participants who judged it to represent different people. The difference in pupil dilation may have been a misleading factor in this pair of images. (These irises were taken from images 04273d348 and 04273d296 in the ND-IRIS-0405 dataset.) Figure reprinted from Bowyer et al. [1] ©2010 IEEE.

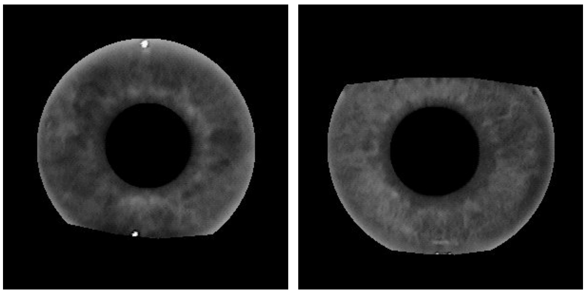

Figure 9: This example nonmatch pair from the iris-only experiment shows the left and right irises of different people. This pair of images was incorrectly classified by 25 of 29 participants who judged it to represent the same person. (These irises were taken from images 04609d236 and 04839d206 in the ND-IRIS-0405 dataset.) Figure reprinted from Bowyer et al. [1] ©2010 IEEE.

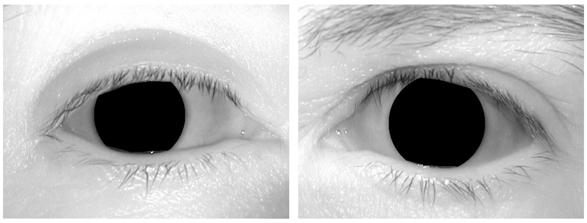

Figure 10: This example match pair shows left and right periocular regions of the same person. This pair of images was incorrectly classified by 22 of 29 participants who judged it to represent different people. (These images show the periocular information from images 04394d600 and 04394d567 in the ND-IRIS-0405 dataset.)

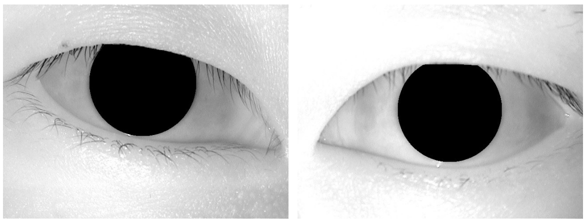

Figure 11: This example nonmatch pair shows left and right periocular regions of different people. This pair of images was incorrectly classified by 25 of 29 participants who judged it to represent the same person. (These images show the periocular information from images 04632d417 and 04850d197 in the ND-IRIS-0405 dataset.) Figure reprinted from Bowyer et al. [1] ©2010 IEEE.

4. Human Perception of Texture Similarity In Identical Twins Irises

From our experiments with humans viewing left and right eye image pairs, we ascertained that humans are detecting similarities in iris texture between left and right eyes. We next conducted an experiment to determine how well humans could detect similarities between irises of identical twins.

4.1. Experimental Set-up

Our experiment on human perception of similarity in identical twins' irises followed the same strategy as our investigations of left and right irises. We used the same display software for the twin experiment as we had used previously in the left/right experiments. Pairs of images – either eyes of twins or eyes of unrelated people – were presented side-by-side on the computer monitor for three seconds. For this experiment, each query showed either both left eyes, or both right eyes; there were no queries that showed a left and a right eye together. After three seconds, the display changed and the volunteer was asked whether the two images were a matched twin pair or not. Participants were asked to select one of five options for each query:

1. Certain these images were from identical twins.
2. Likely they were from identical twins.
3. Can't tell.
4. Likely they were NOT from identical twins.
5. Certain they were NOT from identical twins.

Next, the program gave feedback on whether the user was correct or incorrect.

We used a subset of the same images for the human viewer experiment as we had for the biometrics experiment described in Section 2. We had a total of 196 subjects to use in our queries. Forty-nine twin pairs (98 people) were randomly selected to be used in 49 “twin” queries, and the remaining 98 people were used for 49 unrelated-person queries. No subject appeared more than once in all of the iris image pairs. Next, we constructed 98 queries in a similar manner showing only the periocular region. Unlike the research described in Section 3, we did not conduct a whole-eye experiment with the twin data.

We used the same type of incentives to get volunteers to participate in our experiment. Again, our human subjects participated in these trials under a protocol approved by the Human Subjects Institutional Review Board. Volunteers were recruited from students and staff at the University of Notre Dame. Volunteers received $10 for participating and an additional $10 if they scored above 80%. We had 28 participants, none of whom had participated in the left/right experiments.

4.2. Can humans determine whether a pair of eye images is from identical twins?

We calculated overall accuracy scores for our experiments in the same manner as described in Section 3. The average percent correct on the iris portion of the experiment was 81.3% (standard deviation 5.2%). The minimum score was 68.4%, and the maximum score was 89.8%. The average percent correct on the periocular queries was 76.5% (standard deviation 5.1%). The minimum score on the periocular portion was 63.3% and the maximum score was 86.7%. T-tests showed that on both viewing conditions, users were performing statistically better than random guessing (p-value less than 10^{-4} for both tests).

Accuracy on the twins experiment was lower than the accuracy on the left/right experiment. It may be possible that determining whether images were from twins is a harder problem. However, a more likely explanation is that the image pairs selected for the left/right experiment were of slightly higher quality. The images from the first experiment were taken in a controlled setting at the University of Notre Dame, and were preprocessed by the LG software before saving. The twins images were taken in less controlled conditions at the Twins Days festival.

For the left/right experiment, accuracy was about the same on both iris and periocular conditions. In contrast, the twin experiment showed a drop in accuracy for the periocular condition when compared with the iris condition. We suspect that this result is also due to the quality of the data. For the left/right periocular experiment, we tried to avoid images where a corner of the eye was missing. In contrast, we had less data of twins available, so some image pairs in the twin experiment showed less of the periocular region.

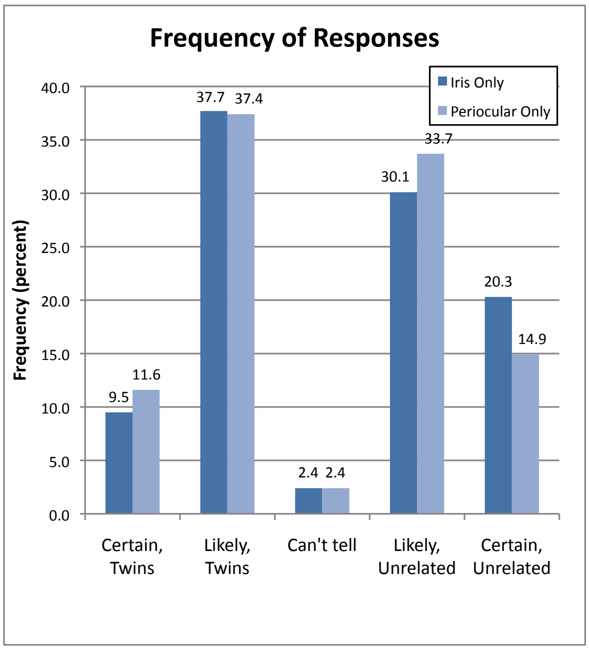

Figure 12: This graph shows the frequency of responses for the five options in our twin human viewer experiments.

4.3. How confident were volunteers of their responses?

As in the earlier test, some of the testers in the twins experiment were more confident than others. One tester responded “certain” for only one of the iris queries and none of the periocular queries. On the other hand, one tester responded “certain” for 64 of the 98 iris queries and 58 of the 98 periocular queries. The average number of “certain” responses on the iris portion of the test was 29.2 out of 98 (standard deviation 17.1). The average number of “certain” responses on the periocular portion of the test was 25.6 out of 98 (standard deviation 17.9). A histogram showing the distribution of responses for both iris and periocular conditions is shown in Figure 12.

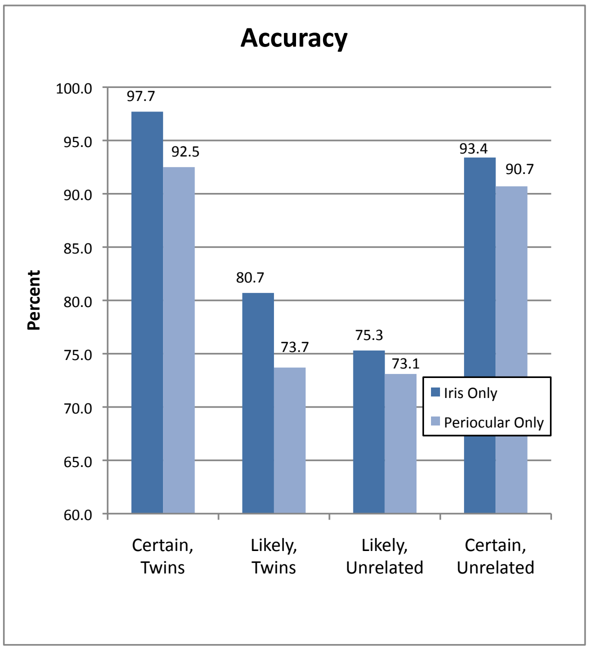

Figure 13: This graph shows accuracy on the twin human viewer experiment, broken down by response. Accuracy was higher on the subset of queries where viewers expressed more certainty in their responses.

4.4. Did volunteers score higher when they felt more certain?

Figure 13 shows the accuracy broken down by response. As in the left/right experiment, viewers scored higher on the subset of trials where they marked that they were certain of their answers.

4.5. Did volunteers learn as the test progressed?

We computed the difference between the accuracy on the second half of the iris queries and the first half of the iris queries. We found this difference for each tester, then computed the average difference across all 28 testers. The average difference was 1.2% (standard deviation 7.4%). The minimum difference was -12.2% and the maximum difference was 18.4%. Thirteen of the 28 participants did better on the second half of the iris queries; eleven did worse, and four stayed the same. Since the average difference is positive, it is possible that some learning is taking place. However, a one-tailed t-test shows that the difference is not statistically significant (p-value 0.2064). The twins experiment had fewer questions than the left/right experiment, so it may be that a longer test is needed in order to see statistically significant evidence of learning.

On the periocular queries, the average performance difference between the first and second halves was 1.7% (standard deviation 9.6%). Fourteen participants did better on the second half of the periocular queries, twelve did worse, and two stayed the same. A one-tailed t-test shows that the small average improvement on the second half of the periocular queries is not statistically significant either (p-value 0.1706).

4.6. Is it easier to label a twin pair as twins than it is to label an unrelated pair as unrelated?

On the twins experiment, participants did not seem to favor answering “twins” nor answering “unrelated”. For the iris section, participants responded “twins” for 47.2% of queries, “can’t tell” for 2.4% of queries, and “unrelated” for 50.4% of queries. For the periocular section, participants responded “twins” for 49.0% of queries, “can’t tell” for 2.4% of queries, and “unrelated” for 48.6% of queries. When queries are divided by subjects’ responses, as in Figure 13, performance on the queries where subjects responded “twins” is slightly higher.

When accuracy is computed for actual twin pairs and actual unrelated pairs, the performance on unrelated pairs is better for the iris condition. For the periocular condition, performance is about the same for twin and unrelated pairs.

4.7. Which image pairs were most frequently classified correctly, and which were most frequently classified incorrectly?

One pair of twins’ irises was classified correctly by all 28 testers. This pair is shown in Figure 14. Six pairs of twins’ periocular images were classified correctly by all 28 testers. An example is shown in Figure 15. There were ten pairs of unrelated iris images that were classified correctly by all 28 testers. An example is shown in Figure 16. There were also three pairs of unrelated periocular images that were classified correctly by all testers. An example is shown in Figure 17.

Figure 18 shows the image pair that was most frequently classified incorrectly. Twenty-five of the 28 participants incorrectly responded that these images were from unrelated people. One of the challenges with this pair of images is the significant difference in pupil radius. Of all unrelated-person pairs, the one most frequently misclassified is shown in Figure 19. The challenge with this pair is that both of the irises have fairly uniform texture.

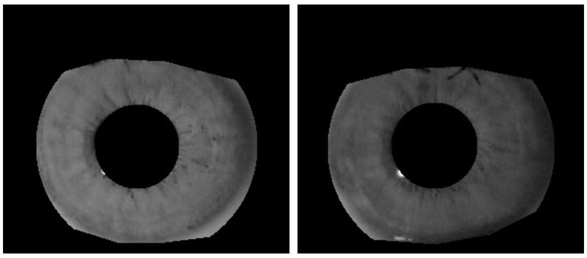

Figure 14: This twin pair was an easy query in the twin human viewer experiment. All 28 testers correctly classified this pair of iris images as being from identical twins. Figure reprinted from Hollingsworth et al. [2] ©2010 IEEE.

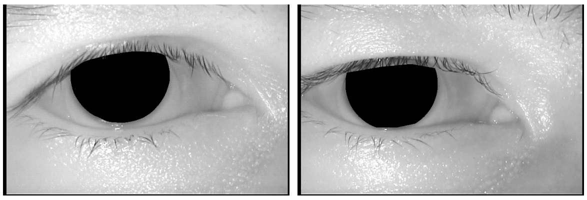

Figure 15: This twin pair was an easy query in the human viewer experiment. All 28 testers correctly classified this pair of periocular images as being from identical twins. Figure reprinted from Hollingsworth et al. [2] ©2010 IEEE.

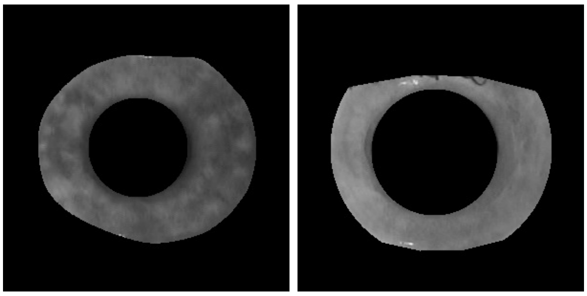

Figure 16: This unrelated pair was an easy query in the twin human viewer experiment. All 28 testers correctly classified this pair of iris images as being from unrelated people. Figure reprinted from Hollingsworth et al. [2] ©2010 IEEE.

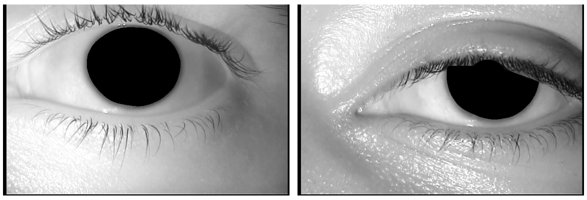

Figure 17: This unrelated pair was an easy query in the twin human viewer experiment. All 28 testers correctly classified this pair of periocular images as being from unrelated people. Figure reprinted from Hollingsworth et al. [2] ©2010 IEEE.

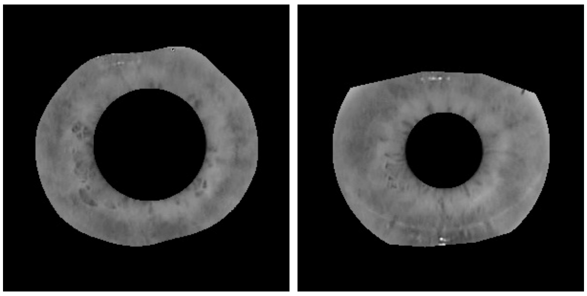

Figure 18: This match pair was a challenging query in the twin human viewer experiment. Twenty-five of 28 participants incorrectly responded that these images were from unrelated people. In fact these irises are from identical twins. The high number of incorrect responses for this pair may be due to the large difference in dilation between the two irises. Figure reprinted from Hollingsworth et al. [2] ©2010 IEEE.

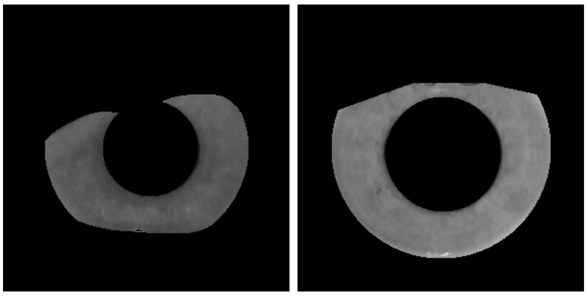

Figure 19: This nonmatch pair was a challenging query in the twin human viewer experiment. Twenty-four of 28 participants incorrectly responded that these images were from twins, when in fact, these irises are from unrelated people. The smoothness of the texture makes this pair difficult to classify correctly. Figure reprinted from Hollingsworth et al. [2] ©2010 IEEE.

5. Variation in Results when using an Unequal Number of Paired and Nonpaired Trials

One limitation of the previous experiments is that they presented equal number of “match” and “nonmatch” queries to the human viewers. In real life, sets of comparisons are not balanced between matches and nonmatches. Therefore, we conducted an additional experiment with pairs of left and right iris images to determine how humans perform when the number of queries showing two different people is larger than the number of queries showing a left/right pair of one individual. For this experiment, we used the same software as before, and allowed participants the same amount of time to view the images. The images used in this experiment were the same as those used in the balanced left/right iris experiment, but this time each volunteer saw only a subset of the original 105 match queries. Each volunteer viewed five times as many nonmatch queries as match queries; that is, he or she saw 105 nonmatch queries and 21 match queries. Different viewers saw different subsets of match queries. In this experiment, as in previous experiments, viewers were not told how many match or nonmatch trials to expect.

We got fifteen volunteers to participate in this experiment. Since we were only displaying one-fifth as many match queries to each participant, each of the original match queries appeared three times at some point during the fifteen tests.

The average accuracy on the unbalanced experiment was 83.2% correct (standard deviation 6.8%). This result is comparable, and in fact an improvement over the accuracy on the balanced iris experiment: 82.6% (standard deviation 4.0%). The increased standard deviation is not surprising, considering that we did not have as many participants for the unbalanced experiment.

Volunteers on this experiment were “certain” of their responses 60.4% of the time. Thus, they selected one of the “certain” options more frequently than the volunteers in the balanced iris experiment, who were confident of their responses on 34.1% of the time. A histogram showing the distribution of responses is given in Figure 20.

Figure 21 shows the accuracy broken down by response. As before, viewers scored higher on the subset of trials where they marked that they were certain of their answers. However, the accuracy on match queries is much lower than before.

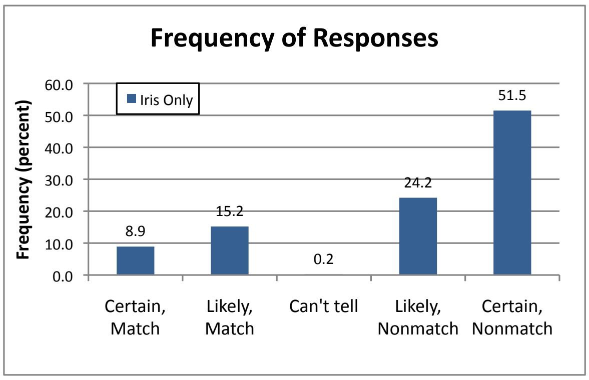

Figure 20: This graph shows the frequency of responses for the five options in the experiment involving unequal numbers of match and nonmatch trials.

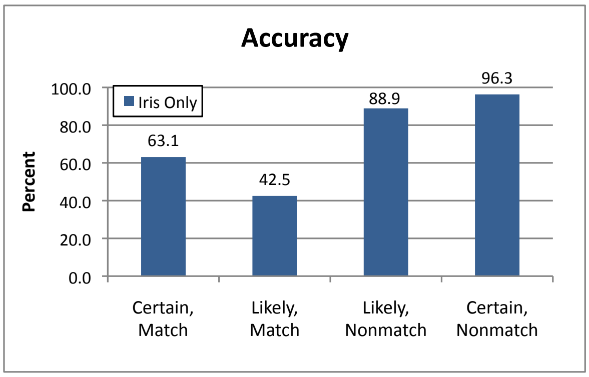

Figure 21: This graph shows accuracy, broken down by response, on the experiment with unequal numbers of match and nonmatch trials. Overall accuracy on this experiment was slightly higher than overall accuracy on the previous iris experiment. The accuracy on the less-frequently presented type of trial (match trials) dropped, and the accuracy on the more-frequently presented type of trial (nonmatch trials) increased.

We conclude from this experiment that participants naturally responded to the uneven distribution of match and nonmatch pairs, selecting the nonmatch response more often. The accuracy on the less-frequently presented type of trial dropped, and the accuracy on the more-frequently presented type of trial increased.

In any real-world scenario with unbalanced data, experimenters must consider the cost of errors when making judgments. They must also be aware of the individual accuracy rates of the different classes, and not simply consider the overall accuracy.

6. Discussion and Conclusion

We have performed a series of experiments on genetically identical irises. We verified previous claims by researchers (e.g. [6]) that iris biometrics systems effectively differentiate between left and right eyes and between irises of identical twins. In terms of the number of twins studied, the largest previously published iris biometrics experiment contained 51 pairs of identical twins [7]. Therefore, our data set contains about fifty percent more twin pairs than any previous study on twin iris biometrics.

Next, we performed a sequence of experiments in which humans viewed a pair of images, one of the left eye and one of the right eye, and evaluated how likely it is that the images are from the two eyes of the same person versus a different person. To our knowledge, our work presents the only research in human perception of similarity between genetically identical irises. We investigated a number of viewing conditions for our human observer experiments. We considered the problem of viewing whole-eye images, and compared that performance to the performance when only the iris, or only the region around the iris was visible. We also conducted experiments in which humans viewed pairs of images and evaluated how likely it is that the images are from identical twins versus unrelated people. The overall accuracy on all experiments was greater than 75% (Table I). On the subset of queries where participants expressed high confidence in their responses, accuracy was greater than 90%. The participants in our experiment were untrained, and were allowed to view the image pairs for only three seconds each. We expect that trained observers with unlimited viewing times could exceed this performance.

Table I: Accuracy of Responses, All Experiments

Viewing Condition	Average Performance	Performance When Expressing High Confidence
Left/right, whole-eye	85.7%	95.5%
Left/right, iris	82.5%	92.0%
Left/right, periocular	82.7%	93.1%
Twins, iris	81.3%	92.1%
Twins, periocular	76.5%	93.3%
Left/right, iris using unbalanced data set	83.2%	93.4%

For both the left/right experiments and the twins experiments, the average performance was slightly higher (between 0.7% and 3.4%) on the second half of the test versus the first half. The improvement was statistically significant for the left/right whole-eye experiment and for the left/right iris experiment. This suggests that participants learned from the feedback given after each query. It also suggests that humans could be trained to evaluate whether two eyes are genetically identical. This experiment tested humans ability to judge genetic relatedness, but we speculate that humans could also be trained to address the biometric technology question of whether two pictures show images of the same eye.

We found that performance was about 3% higher when participants viewed the entire eye image, rather than just the iris, or just the periocular region. Surprisingly, our left/right experiments showed that there was no significant different in performance between humans viewing iris-only and humans viewing periocular-only images. For the twins experiments, the performance was higher on the iris-only images than on the periocular-only images. However, this phenomenon can be attributed to our limited data set and consequential inability to include the whole eye in all the images. We suspect that better performance can be achieved when both eye corners are present in all periocular images. It is worth noting that the LG 2200 iris images used in our experiments are acquired with near-infrared illumination. This means that the images of the iris and of the periocular region are not the same as would be acquired using visible-wavelength illumination. Thus it may be useful in future work to explore similar issues based on using images acquired under visible-wavelength illumination.

We noted some factors that made it more difficult to judge whether two images came from genetically identical eyes. Specifically, it is misleading when two eyes are not aligned vertically in the images. We recommend that eyes be centered and aligned for optimal human performance on any eye comparison tasks. In addition, a difference in pupil size can be misleading in determining whether two eyes are genetically related. This result is parallels a result from biometric research [24] that found that iris biometric performance drops when there is a large difference in pupil dilation. Ideally, for any forensic applications, eye comparisons should be done with images having equal sized pupils.

Our experiments highlight the difference between automated biometric technology and human viewers. Biometric technology makes use of epigenetic texture features of the iris [6]. In contrast, humans can analyze the overall iris texture pattern, which clearly contains some genetic component. Given that humans can perform this one task that current iris biometric technology cannot, it may be interesting to ask what other tasks humans can also perform. For instance, is there a similarity in iris texture between parent and child?

Our work suggests that human examination of pairs of iris images for forensic purposes may be feasible. Our results also suggest that development of different approaches to automated iris image analysis could be used to answer questions that current iris biometric technology cannot answer.

Acknowledgment

This work is supported by the Federal Bureau of Investigation, the Central Intelligence Agency, the Intelligence Advanced Research Projects Activity, the Biometrics Task Force, the Technical Support Working Group under US Army contract W91CRB-08-C-0093, and the Intelligence Community Postdoctoral Fellowship Program. Samuel Fenker is supported as an Ateyeh Undergraduate Research Scholar through the Ateyeh Endowment for Excellence at the University of Notre Dame. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of our sponsors.

References

[1] K. W. Bowyer, S. Lagree, S. Fenker, Human versus biometric detection of similarity in left and right irises, in: IEEE International Carnahan Conference on Security Technology (ICCST), 2010, pp. 1–7.

[2] K. Hollingsworth, K. W. Bowyer, P. J. Flynn, Similarity of iris texture between identical twins, in: IEEE Computer Vision and Pattern Recognition Biometrics Workshop, 2010, pp. 1–8.

[3] A. K. Jain, S. Prabhakar, S. Pankanti, On the similarity of identical twin fingerprints, Pattern Recognition 35 (11) (2002) 2653–2663.

[4] J. Daugman, C. Downing, Epigenetic randomness, complexity and singularity of human iris patterns, The Royal Society 268 (1477) (2001) 1737–1740.

[5] R. P. Wildes, Iris recognition: An emerging biometric technology, Proceedings of the IEEE 85 (9) (1997) 1348–1363.

[6] J. Daugman, How iris recognition works, IEEE Transactions on Circuits and Systems for Video Technology 14 (1) (2004) 21–30.

[7] Z. Sun, A. A. Paulino, J. Feng, Z. Chai, T. Tan, A. K. Jain, A study of multibiometric traits of identical twins, in: SPIE, 2010, pp. 1–12.

[8] L. Stark, K. W. Bowyer, S. Siena, Human perceptual categorization of iris texture patterns, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2010, pp. 1–7.

[9] K. W. Bowyer, P. J. Flynn, The ND-IRIS-0405 iris image dataset, Tech. rep., University of Notre Dame, http://www.nd.edu/~cvrl/papers/ND-IRIS-0405.pdf (2009).

[10] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. Bowyer, C. L. Schott, M. Sharpe, FRVT 2006 and ICE 2006 large-scale experimental results, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (5) (2010) 831–846.

[11] LG, http://www.lgiris.com/ (accessed Jan 2008).

[12] S. Baker, A. Hentz, K. W. Bowyer, P. J. Flynn, Contact lenses: Handle with care for iris recognition, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2009, pp. 1–8.

[13] J. Daugman, New methods in iris recognition, IEEE Transactions on Systems, Man and Cybernetics - B 37 (5) (2007) 1167–1175.

[14] MathWorks, Two-sample kolmogorov-smirnov test, http://www.mathworks.com/help/toolbox/stats/kstest2.html (accessed Jun 2011).

[15] Y. Wang, T. Tan, A. K. Jain, Combining face and iris biometrics for identity verification, in: Int. Conf. on Audio- and Video-Based Biometric Person Authentication, 2003, p. 805.

[16] Twins days festival official website, http://www.twinsdays.org/td_times/td_times_2001.html, accessed January 2011.

[17] L. Flom, A. Safir, Iris recognition system, U.S. Patent 4,641,349 (1987).

[18] J. Daugman, High confidence visual recognition of persons by a test of statistical independence, IEEE Trans-

actions on Pattern Analysis and Machine Intelligence 15 (11) (1993) 1148–1161.

[19] L-1 identity solutions, understanding iris recognition, http://www.l1id.com/pages/383-science-behind-the-technology, accessed March 2010.

[20] Iris recognition, http://en.wikipedia.org/wiki/Iris_recognition, accessed March 2010.

[21] U. Park, A. Ross, A. K. Jain, Periocular biometrics in the visible spectrum: A feasibility study, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2009, pp. 1–6.

[22] P. Miller, A. Rawls, S. Pundlik, D. Woodard, Personal identification using periocular skin texture, in: Proc. ACM 25th Symposium on Applied Computing (SAC2010), 2010, pp. 1496–1500.

[23] D. L. Woodard, S. Pundlik, P. Miller, R. Jillela, A. Ross, On the fusion of periocular and iris biometrics in non-ideal imagery, in: Proc. Int. Conf. on Pattern Recognition, 2010, pp. 201–204.

[24] K. P. Hollingsworth, K. W. Bowyer, P. J. Flynn, Pupil dilation degrades iris biometric performance, Computer Vision and Image Understanding 113 (1) (2009) 150–157.