Identity Verification Using Iris Images: Performance of Human Examiners

Kevin McGinn, Samuel Tarin and Kevin W. Bowyer
Department of Computer Science and Engineering
University of Notre Dame
kmcginn3, starin and kwb @nd.edu

Abstract—**We are not aware of any previous systematic investigation of how well human examiners perform at identity verification using the same type of images as acquired for automated iris recognition. This paper presents results of an experiment in which examiners consider a pair of iris images to decide if they are either (a) two images of the same eye of the same person, or (b) images of two different eyes, with the two different individuals having the same gender, ethnicity and approximate age. Results suggest that novice examiners can readily achieve accuracy exceeding 90% and can exceed 96% when they judge their decision as "certain". Results also suggest that examiners may be able to improve their accuracy with experience.**

Keywords—iris recognition; forensic examination.

I. INTRODUCTION

The only previous work that we aware of on the topic of human evaluation of the similarity of iris images is that of Hollingsworth et al [1]. They investigated whether or not human examiners can detect similarity in the texture of "monozygotic" irises. In one experiment, they presented examiners with pairs of left and right iris images, and found that the examiners performed well at classifying the pairs as belonging to the same person or to different persons. In another experiment, they presented examiners with a pair of iris images that were either from identical twins or from unrelated persons. Again, they found that examiners performed well at classifying the pairs of images.

Our work is focused on the question of how accurately a human examiner can determine if two iris images come from the same iris. Although automated iris recognition technology is already very accurate and continues to improve [2,3,4], there will always be some small rate of false matches and false non-matches. In the case of disputed results, human examiners may be called upon to make a final decision. In addition, the American justice system would likely require some level of human expert verification of a match in order to use biometric information in a courtroom setting. Just as fingerprint database searches still require a human to make a final decision from a set of potential matches, so too might iris recognition leave the final decision to a human being in some instances. Additionally, if human examiners make different sorts of errors than automated matching, hybrid matching may achieve greater accuracy than either alone. Thus, for various reasons, the ability of human examiners to determine if two iris images belong to the same person is an important topic, and one that is largely unexplored. This paper presents results of the first systematic investigation of this question

II. DATASET OF AUTTHENTIC AND IMPOSTOR IMAGE PAIRS

All images used in this experiment were acquired with a common commercial iris recognition sensor, the LG 4000 [11]. This sensor produces near-IR-illuminated images of size 480x640 with the eye approximately centered in the image. See Figure 1 for example images. A total of 95 "authentic" image pairs and 95 "imposter" pairs were selected for the experiment.

We define an authentic pair as two images of the same iris. As an additional restriction, we select images taken approximately two years apart. The substantial time lapse between images should better correspond to conditions of a practical identity verification task than two images acquired in the same day. There is evidence in the automated iris matching literature that using two images of the same iris acquired in the same session could result in unrealistically good matching performance [5]. The demographics for the 93 unique subjects featured in the 95 authentic pairs are shown in Table I. Demographic data was self-reported by persons participating in image acquisition sessions.

TABLE I. DEMOGRAPHICS FOR AUTHENTIC IRIS IMAGE PAIRS

Gender	Male: 42; Female: 51
Ethnicity	Asian: 19; African-American: 2; White: 72
Eye color	Black: 6; Br.: 43; Blue: 16; Gray: 3; Green: 13; Hazel: 12
Year of Birth	1940s: 3; 1950s: 4; 1960s: 8; 1970s: 8; 1980s: 66; 1990s: 4

We define an impostor pair of images as coming from different persons who have the same gender, ethnicity and approximate age as each other. The requirement of same gender, ethnicity and approximate age again complicates the creation of the experimental dataset, but should better correspond to conditions of a forensic identity verification task. If the two images in an impostor pair were not "demographically yoked" in this way, the experiment might find unrealistically high recognition performance. There is evidence in the automated iris recognition literature that there may be similarities in iris texture between gender [6] and ethnicity [7,8]. The impostor pairs of images were selected so that the impostor distribution approximately matches the distribution of (eye color, age, gender, race) in the authentic pairs.

Of the 95 impostor pairs, eight were from identical twins. Based on the results in [1], these eight impostor pairs are expected to be more difficult for viewers to classify accurately than impostor pairs from unrelated persons.

III. Experimental Method

A total of 22 volunteers participated in the experiment. They were recruited from the undergraduate and graduate student population at the University of Notre Dame. Some may have been familiar with general concepts of biometrics and computer vision, but none were working on research in iris recognition or had any previous experience in examining iris images of the type used in the experiment.

Once the experiment software is started, the user is first presented with a brief description of the experiment, as well as instructions for operating the software. The user is then presented with several example pairs of images, illustrating authentic pairs, impostor pairs, and special cases such as specular highlights and contact lenses. The user is then given the option to review the example pairs again, or to begin the actual experiment.

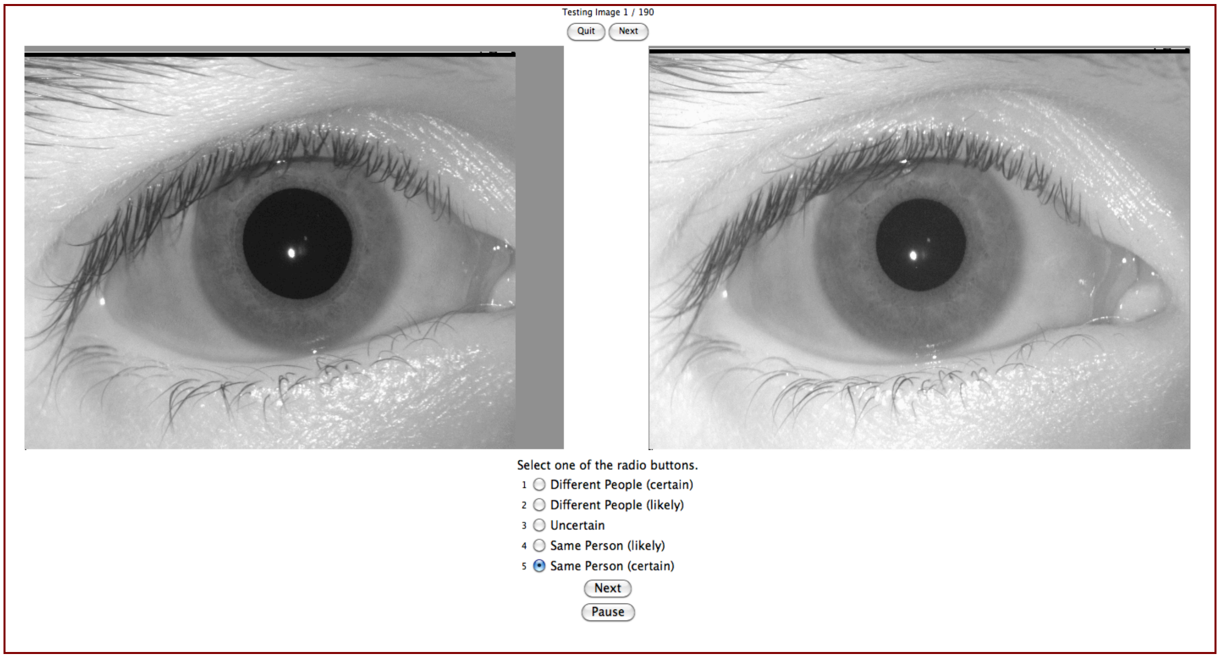
A screenshot of an example trial in the experiment is shown in Figure 1. In each trial, the user is presented with a pair of 640x480 iris images, and asked to select one of five possible responses:

1 – Different People (certain)
2 – Different People (likely)
3 – Uncertain
4 – Same Person (likely)
5 – Same Person (certain)

In each run of the experiment, image pairs are randomly selected from the list of unused iris pairs, until the user has seen all 190 pairs. Therefore, all examiners see the same set of 190 image pairs, but see them in a different randomly-selected order.

Assuming that the user selects an option other than "uncertain", they are then presented with text below the iris pair indicating whether or not their selected categorization was correct. At any point in the process, the user can either quit or pause the experiment, which hides the current images and pauses the timer that is recording the time spent on that pair.

Note that the whole 640x480 image as acquired by the LG 4000 sensor is displayed for the examiner. There is nothing to prevent the examiner from using image content beyond the iris texture in order to classify an image pair. For example, the examiner might also consider similarity in eyelashes, eyebrows, eye shape, or other factors. Traditional automated iris recognition algorithms use the texture of the iris region as the primary or only source of information, although they of course also find eyelid and eyelash shape for purposes of detecting occlusion of the iris region. Thus, the human examiner is likely using more of the image content to reach a decision than current automated iris recognition algorithms use. However, we feel that it is unreasonable to artificially limit the data available to the human examiner to only that used by current automated iris recognition algorithms. This is especially the case since we hope to get a decision from the human examiner that is in some sense independent of that of the automated algorithm.



Testing Image 1 / 190

Select one of the radio buttons.

1 ○ Different People (certain)
2 ○ Different People (likely)
3 ○ Uncertain
4 ○ Same Person (likely)
5 ● Same Person (certain)

Figure 1 – Screen Shot of Sample Trial from the Experiment.

IV. Results

The average time taken to complete the 190-trial experiment was approximately 26 minutes. The maximum was approximately 42 minutes, and the minimum was approximately 15 minutes. The distribution by subject of the time taken for the experiment and overall accuracy in classifying image pairs is shown in Figure 2. Note that the subject with the lowest overall accuracy also took the least amount of time. As shown in Figure 2, this subject's accuracy is considerably lower than that of the other subjects, and can be considered an outlier.

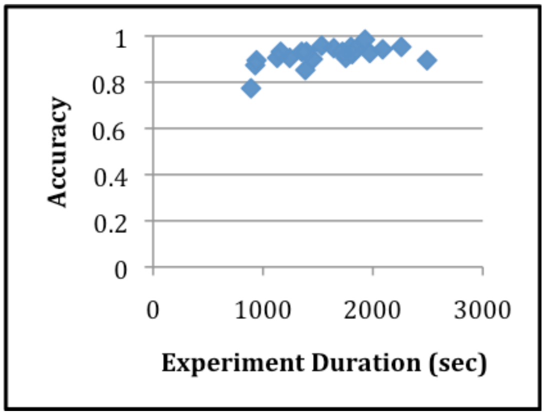


Figure 2 – Accuracy vs. Experiment Time.

The subjects correctly classified an average of just over 174 pairs out of 190, for an overall average accuracy of almost 92%. Across the subjects, the minimum correct was 147 pairs (just over 77%), and the maximum correct was 187 pairs (just over 98%).

Of the 95 authentic pairs, the subjects classified an average of just over 89 pairs correctly (almost 94%). The minimum number of authentic pairs that any subject classified correctly was 76 pairs (80%), and the maximum was 95 pairs (100%). The standard deviation was 4.6. Of the 95 impostor pairs, the subjects classified an average of almost 85 pairs correctly (just under 90%). The minimum was 71 impostor pairs correct (almost 75%), and the maximum correct was 93 pairs (almost 98%). The standard deviation was 5.7.

The average difference between the number of authentic pairs classified correctly and the number of impostor pairs classified correctly was 4.4 pairs. The standard deviation of the differences was 6.2. A matched pairs t-test gave a p-value of 0.003, so for most accepted levels of significance, it can be concluded that there is a statistically significant difference between the average accuracy for authentic pairs and impostor pairs.

The overall accuracy broken down by authentic trials and imposter trials is shown in Table 2. Based on [1], the twin impostor pairs are expected to be significantly harder than normal impostor pairs. This skews the authentic-versus-impostor results in favor of the authentic results being higher. If we remove the results for the twin pairs from the impostor pairs, the average percentage of impostor pairs classified correctly increases to 93.2% with a minimum of 79.3% and a maximum of 100%. The average difference in percent correctly classified between authentic pairs and non-twin impostor pairs was just 0.7%, with a standard deviation of 6.1. This difference is obviously not statistically significant.

Across the 22 subjects, the option "Same Person (certain)" was selected for about 35% of the pairs, "Same Person (likely)" was selected for about 17% of the pairs, "Unsure" was selected for less than 1% of the pairs, "Different people (likely)" was selected for about 19% of the pairs, and "Different people (certain)" was selected for about 28% of the pairs.

TABLE II. Accuracy by Type of Image Pair in Trial

	Authentic Pairs	Impostor Pairs	Non-Twin Impostors
Average	93.9%	89.3%	93.2%
Maximum	100%	97.9%	100%
Minimum	80%	74.7%	79.3%

The average difference in number of pairs marked as "Same Person (certain)" and marked as "Same Person (likely)" was 34.6, with a standard deviation of 37.5. A matched pairs t-test returned a p-value of 0.0003, so it can be concluded that there is a significant difference in the number of times the two options were selected. Thus, subjects seemed to be confident about most of their choices for the authentic pairs.

The average difference in number of pairs marked as "Different Person (certain)" and marked as "Different Person (likely)" was 17.2, with a standard deviation of 38.3. A matched pairs t-test returned a p-value of 0.047. For $\alpha = 0.05$, it can be concluded that there is a difference in the number of times the two options were selected. Thus, subjects seemed to be fairly confident about most of their choices for the impostor pairs, although this confidence does not seem to be as high as the authentic pairs. This is likely due to some of the pairs of impostor images coming from identical twins.

Since both types of "certain" responses were chosen more often than the others, it can be inferred that the subjects generally had high confidence in their responses. These selection patterns are skewed more toward the "certain" responses than the responses in the previous left-right iris matching experiment by Hollingsworth et. al [1]. This could indicate that matching two irises from the same side of the face is an easier task than matching them across the face.

One examiner marked all their responses as "certain", never using the "likely" or "uncertain" options. As a result, this subject's information has been excluded from the following calculations concerning the subjects' ability to judge their confidence accurately.

The percent of correct responses across subjects for the different response types is shown in Figure 3. For all of the times that the "Same Person (certain)" option was selected, the classification was correct 96.3% of the time. For "Same

Person (likely)”, the classification was correct 76.7% of the time. For “Different Person (likely)”, the classification was correct 91.3% of the time. For “Different Person (certain)”, the classification was correct 96.5% of the time.

The average difference in percentage of pairs correct for “Same Person (certain)” and “Same Person (likely)” was 0.201, with a standard deviation of 0.135. A matched pairs t-test resulted in a p-value of 0.000001. Thus it can be concluded that there is a statistically significant difference in the percentage of pairs correct between the “certain” and “likely” responses for authentic pairs.

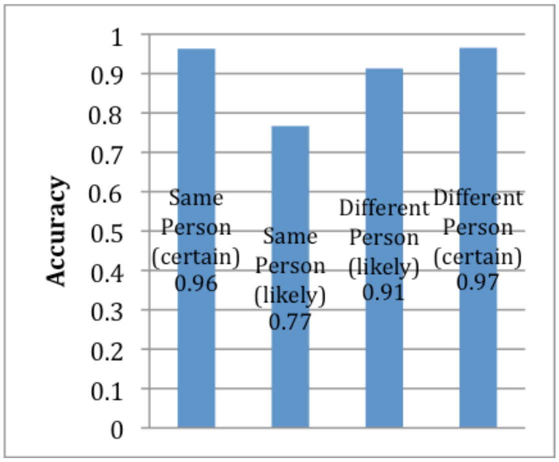


Figure 3 – Accuracy By Category of Response.

The average difference in percentage of pairs correct for “Different Person (certain)” and “Difference Person (likely)” was 0.054, with a standard deviation of 0.053. A matched pairs t-test resulted in a p-value of .0001. Thus, again, it can be concluded that there is a statistically significant difference in the percentage of pairs correct between the “certain” and “likely” options for impostor pairs.

Since both of the “certain” options had higher accuracy than their respective “likely” options for both types of pairs, it can be concluded that the subjects were appropriately judging their confidence when selecting their response.

Recall that examiners were given feedback about the correctness of their response after each trial, and that the image pairs were presented in a randomized order. In order to determine if the subjects learned how to improve their classifications of the pairs over the course of the experiment, the results were split into approximate thirds: the first 65 trials, the middle 60 trials, and the last 65 trials. The middle third was ignored for the purposes of this question. For the first 65 trials of the experiment, the subjects correctly classified an average of 58.36 pairs (89.8%). For the last 65 trials of the experiment, the subjects correctly classified an average of 60.36 pairs (92.9%). The average difference between the first third and last third of the experiment was 2 pairs. A matched pairs t-test on the difference between the thirds resulted in a p-value of 0.0038. Thus, for most accepted levels of significance, there does appear to be a small but statistically significant difference in subject accuracy between the beginning and end of the test. It appears that examiners learned to perform the classification better during the experiment.

V. Conclusions and Discussion

To our knowledge, this is the first work to systematically explore the performance of human classification of iris image pairs of the type acquired for automated iris recognition. In our experiment, examiners performed well at iris matching, with an average accuracy across subjects and types of trials of approximately 91%. The level of accuracy increases to 96% when examiners express confidence in their decision.

A majority of the image pairs that were most frequently categorized incorrectly were impostor pairs, especially pairs from twins. An example of one of these pairs is shown in Figure 4a. In general, the twin pairs had very similar overall iris texture and periocular features. It is worth noting that while human image interpreters may see substantial similarity in the iris texture of identical twins, automated iris recognition algorithms “see” no greater similarity in iris codes of identical twins that in unrelated persons [1].

To improve the ability for human interpreters to correctly distinguish iris images from twins, it might be useful for “points of interest” to be highlighted independently on each of the iris images. This could serve to make local differences in the overall similar texture pattern more readily apparent. However, this would require significant preprocessing of the images, and likely also the ability for the examiner to control overlays that indicate points of interest.

The most frequently misclassified authentic pairs all exhibited significant differences in pupil dilation between the two images. The larger pupil in these images tended to “warp” the texture of the iris relative to that in the smaller dilation, and thus cause some of the examiners to mistakenly believe that the textures belonged to different people. Examples of these pairs are shown in Figure 4b and 4c.

In cases where there is substantial difference in pupil dilation between a pair of images, it might be useful for a transformation to be applied to one of the images to equalize the pupil dilation and stretch the iris pattern appropriately. However, it is also known that large differences in pupil dilation cause problems for the standard “rubber sheet” model used in automated iris recognition [9,10]. This suggests that it may be important to develop more sophisticated and realistic models for iris texture change with dilation.

In the context of pupil dilation difference between images, it is worth noting that the choice of iris sensor may be an important consideration. The iris sensor used to acquire the images for this experiment, the LG 4000, does not use a visible light source to try to control pupil dilation. Other iris sensors, such as the Iris Guard AD 100 [12], incorporate a visible light source in addition to the near-IR light source, and use the visible light source to avoid highly-dilated pupils. This may effectively reduce the number of classification errors caused by difference in pupil dilation.

The hardest of the most frequently misclassified impostor pairs belonged to identical twins. This accords the observations in [1]. The only twin iris pair that did not have as high a rate of misclassification exhibited a significant pupil dilation difference. This pupil dilation difference likely exaggerated the differences that existed between the two irises, or distorted what should have been very similar textures enough to mislead the subjects into believing that the patterns were more different than they actually were. The most frequently missed non-twin pairs all shared the same characteristics: the two irises had very similar textures, the iris textures did not have many distinguishing points of interest, and the periocular regions were very similar.

There appears to be a small increase in overall accuracy in the last third of the experiment, compared to the first third. This increase in accuracy was observed over the course of one session involving 190 pairs. This suggests that human image interpreters may be able to improve their accuracy with experience and training. This issue deserves further examination, to confirm that a learning effect is seen in larger experiments, and to determine if the learning effect results in improvement on the more-difficult-to-classify image pairs.

References

[1] K. Hollingsworth, et al, Genetically Identical Irises Have Texture Similarity That Is Not Detected By Iris Biometrics, *Computer Vision and Image Understanding* 115 (2011), 1493-1502.

[2] J. Daugman, New Methods In Iris Recognition, *IEEE Trans. on Systems Man and Cybernetics B: Cybernetics* 37 (5), 1167-1175, October 2007.

[3] P. J. Phillips, et al. FRVT 2006 and ICE 2006 large-scale experimental results, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32 (5), May 2010, 831-846.

[4] M. J. Burge and K. W. Bowyer, Handbook of Iris Recognition, Springer, 2013.

[5] P. Tome-Gonzalez, F. Alonso-Fernandez and J. Ortega-Garcia, On the effects of time variability in iris recognition, *Int. Conf. on Biometrics (ICB)*, 2008.

[6] V. Thomas, et al, Learning to Predict Gender from Irises, *IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems (BTAS)*, Sept. 2007, Washington, DC.

[7] X. Qiu, et al, Learning Appearance Primitives of Iris Images for Ethnic Classification, *IEEE Int. Conf. on Image Processing (ICIP)*, 2007.

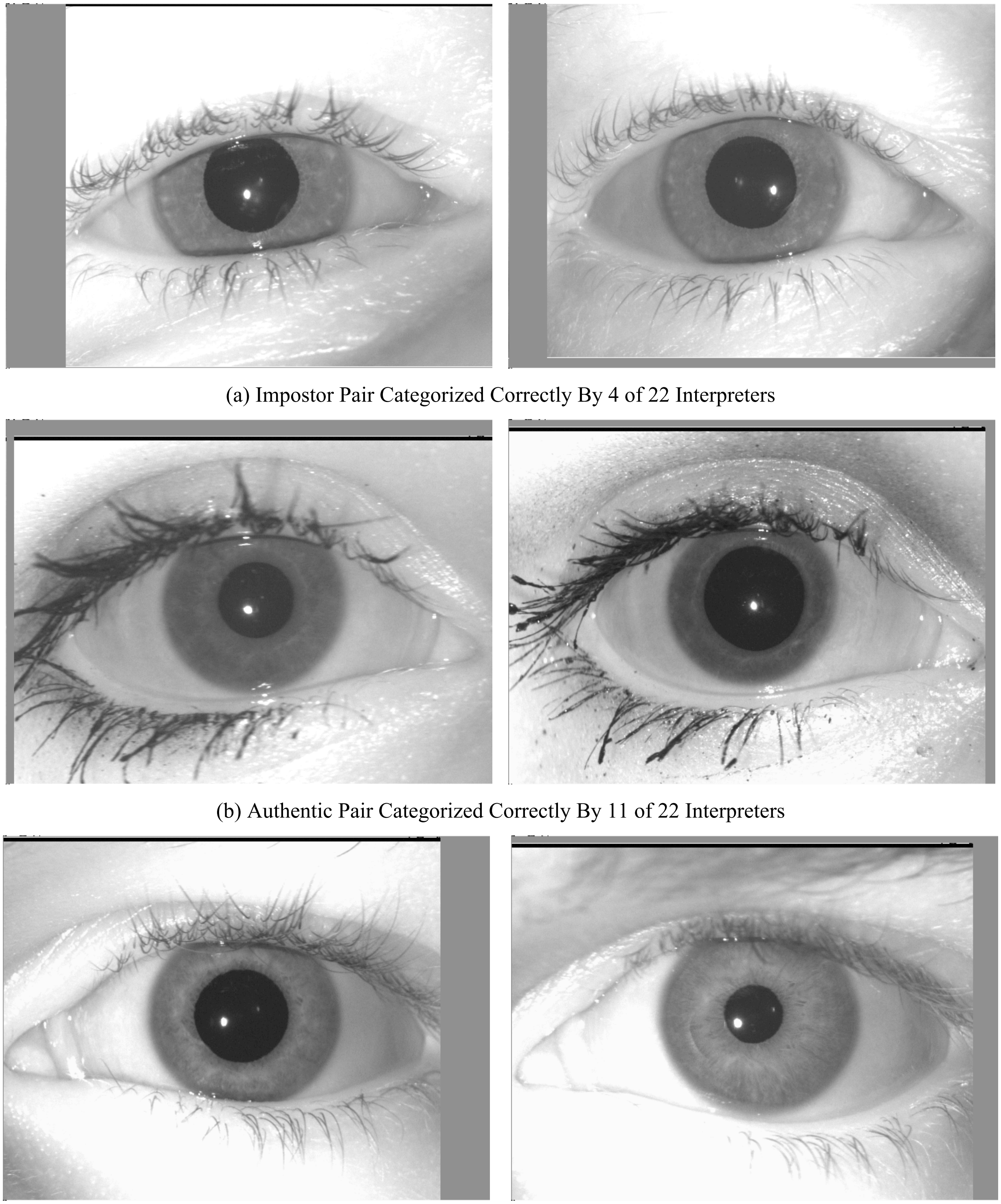
[8] S. Lagree and K. W. Bowyer, Predicting Ethnicity and Gender from Iris Texture, *IEEE Int. Conf. on Technologies for Homeland Security (HST)*, Nov. 2011, Boston, MA.

[9] K. Hollingsworth, et al, Pupil Dilation Degrades Iris Biometric Performance, *Computer Vision and Image Understanding* 113 (1), 150-157, January 2009.

[10] P. Grother, et al, NIST IREX I: Performance of Iris Recognition Algorithms on Standard Images, NIST Interagency Report 7629, Sept 22, 2009.

[11] www.lgiris.com/ps/products/irisaccess4000.htm, accessed July 5, 2013.

[12] www.irisguard.com/pages.php?menu_id=29, accessed July 5, 2013.



(a) Impostor Pair Categorized Correctly By 4 of 22 Interpreters

(b) Authentic Pair Categorized Correctly By 11 of 22 Interpreters

(c) Authentic Pair Categorized Correctly By 12 of 22 Interpreters

Figure 4 – Examples of Image Pairs That Were Difficult to Categorize Correctly.