The Iris Challenge Evaluation 2005

P. Jonathon Phillips, Kevin W. Bowyer, Patrick J. Flynn, Xiaomei Liu, W. Todd Scruggs

Abstract— **This paper describes the Iris Challenge Evaluation (ICE) 2005. The ICE 2005 contains a dataset of 2953 iris images from 132 subjects. The data is organized into two experiments: right and left eye. Iris recognition performance is presented for twelve algorithms from nine groups that participated in the ICE 2005. For the top performers, verification rate on the right iris is above 0.995 at a false accept rate of 0.001. For the left iris, the corresponding verification rates are between 0.990 and 0.995 at a false accept rate of 0.001. The results from the ICE 2005 challenge problem were the first to observe correlations between the right and left irises for match and non-match scores, and quality measures.**

I. INTRODUCTION

The Iris Challenge Evaluation (ICE) 2005 was the first iris recognition challenge problem and was modeled after the Face Recognition Grand Challenge [20]. The goals of the ICE 2005 were to foster the development of iris recognition algorithms and iris processing algorithms and to provide an open benchmark for iris recognition performance. The benchmark is open because performance is reported on a publicly available dataset and researchers can tune their algorithms to the dataset. The ICE 2005 provided a dataset for algorithm development, an experimental protocol for measuring performance, and the irisBEE baseline algorithm. The ICE 2005 data collection protocol was designed to collect iris images of a broader quality range than was encountered in standard existing sensor configurations.

While there are other publicly available iris datasets [5], [9], [10], [16], [23], [29], the ICE 2005 is the first iris challenge problem. The ICE 2005 has the following properties:

- One of the largest publicly-availabe datasets.
- Provides a common protocol for measuring algorithm performance.
- The dataset is organized into experiments that allow for direct comparison among algorithms.
- The irisBEE baseline iris recognition algorithm based on Masek [17].

While the organized portion of the ICE 2005 ended in March 2006, this paper serves as an archival documentation of the ICE 2005 challenge problem and a summary of experimental results through March 2006 (the organized portion of the ICE 2005). The ICE 2005 challenge problem is still available and being distributed to researchers worldwide[1]. The ICE 2005 effort was the first that allowed a direct comparison among iris recognition algorithms and showed correlations between the left and right irises for match and non-match similarity scores, and quality measures. The ICE 2005 is still influencing research because researchers continue to request the dataset and challenge problem; and researchers continue to report results on experiments that use the ICE 2005 dataset in papers [2], [4], [7], [8], [13], [14], [15], [18], [22], [25], [24], [26], [28], [27], [30] (see [3] for a survey of iris recognition).

The ICE 2005 was followed by the ICE 2006 [21]. The ICE 2006 was an independent evaluation of iris recognition algorithms. Participants submitted algorithms to the National Institute of Standards and Technology for testing on sequestered iris images. The ICE 2006 was one of three third party evaluations to measure performance of iris recognition algorithms on sequestered data. To assess the state-of-the-art in iris recognition, Newton and Phillips [19] performed a meta-analysis on these evaluations: the ICE 2006, the Independent Test of Iris Recognition Technology (ITIRT), and the Iris 06 [1],[12]. The meta-analysis found across all three evaluations, reported false reject rate (FRR) at a false accept rate (FAR) of 0.001 ranged from 0.012 to 0.038. At an FAR of 0.001, the range of FRR for the best performers in each test was 0.012 to 0.015, with an average FRR of 0.014. Despite the differences in the testing protocols, sensors, image quality, subject variability and failures to enroll and acquire, the performance results from all three evaluations were comparable.

II. IRIS DATA

The ICE 2005 images were collected with the LG EOU 2200 and intentionally represent a broader range of quality than the sensor would normally acquire. This includes iris images that did not pass the quality control software embedded in the LG EOU 2200. The LG EOU 2200 is a complete acquisition system and has automatic image quality control checks.

The image quality software embedded in the LG EOU 2200 is one of numerous iris quality measures. Flynn and Phillips [11] showed that in the ICE 2006, quality measures are paired with matching algorithms; different quality measures are not correlated; and none of the iris quality measures generalize to all algorithms in the ICE 2006 [21]. This implies that evaluations risk being biased against submissions

P. J. Phillips is with NIST, Gaithersburg, MD 20899 jonathon@nist.gov
K. W. Bowyer is with the University of Notre Dame kwb@cse.nd.edu
P. J. Flynn is with the University of Notre Dame flynn@nd.edu
X. Liu is with the University of Notre Dame
W. T. Scruggs is with SAIC WENDELL.T.SCRUGGS@saic.com

[1]For information on obtaining the ICE 2005 dataset and challenge problem see http://iris.nist.gov/ice.

978-1-4244-2730-7/08/$25.00 ©2008 IEEE

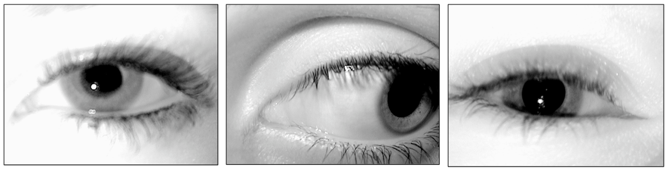


Fig. 2. Examples of "lower" quality iris images in the ICE 2005 dataset.

if the iris images are screened by a quality measure. Prior to the start of the ICE 2005 collection, an arrangement was made to minimize the effect of the LG EOU 2200 quality screening software on the data collection. The subsequent analysis of the effect of quality scores on performance shows that this decision was appropriate.

By agreement between U. of Notre Dame and Iridian, a modified version of the acquisition software was provided. The modified software allowed all images from the sensor to be saved under certain conditions, as explained below.

The iris images are 480x640 in resolution. For most "good" iris images, the diameter of the iris in the image exceeds 200 pixels. The images are stored with 8 bits of intensity, but every third intensity level is unused. This is the result of a contrast stretching automatically applied within the LG EOU 2200 system.

In our acquisitions, the subject was seated in front of the system. The system provides recorded voice prompts to aid the subject to position their eye at the appropriate distance from the sensor. The system takes images in "shots" of three, with each image corresponding to illumination of one of the three near-infrared light-emitting diodes (LEDs) used to illuminate the iris.

For a given subject at a given iris acquistion session, two "shots" of three images each are taken for each eye, for a total of 12 images, see Figure 1 for an example set of images from an acquisition session. The system provides a feedback sound when an acceptable shot of images is taken. An acceptable shot has one or more images that pass the LG EOU 2200's built-in quality checks, but all three images are saved. If none of the three images pass the built-in quality checks, then none of the three images are saved. At least one third of the iris images do pass the Iridian quality control checks, and up to two thirds do not pass; see Figure 2 for "lower" quality iris images.

A manual quality control step was performed at Notre Dame to remove images in which, for example, the eye was not visible at all due to the subject having turned their head.

The data was collected at Notre Dame in January and February 2004. Subject 240596 wore cosmetic contacts for some of the image acquisitions.

III. EXPERIMENTS

The ICE 2005 consisted of two experiments. Experiment 1 measured performance of the right eye and Experiment 2 measured performance of the left eye. Table I reports the number of subjects, irises, and match and non-match scores for each experiment (see Section VI-A for definition of match and non-match score). The total number of subjects is 132 with 112 subjects overlapping between Experiments 1 and 2. The total number of iris images is 2953.

TABLE I
NUMBER OF IRIS IMAGES, SUBJECTS, AND MATCH AND NON-MATCH SCORES PER EXPERIMENT.

Exp.	No. iris images	No. Subjects	No. match scores	No. non-match scores
1 (right eye)	1425	124	8376	659,365
2 (left eye)	1528	120	10,438	758,952

There are two components to the ICE 2005 experiments. The first is recognition. For each experiment there is one signature set. In the ICE 2005, a signature set consists of a list of iris images. For the recognition experiment, an algorithm computes a similarity score between all pairs of iris images in a signature set. In the ICE 2005, performance was only computed from similarity scores where the target and query images were taken on different days.

The second component of the ICE 2005 is measuring iris image quality. In the quality component, a quality score was reported for each iris image in a signature set.

In the original ICE 2005 distribution, image 246260.tiff of subject 289824 was mistakenly included in Exp. 2 (left eye). Image 246260.tiff is in fact an image of a right iris. The mask matrices for Exp. 2 were modified so that image 246260.tiff was not included in calculating performance for Exp. 2. The meta-data distributed with the ICE 2005 lists image 246260.tiff as a right iris of subject 289824.

IV. PROTOCOL

The complete ICE 2005 data and challenge problem were made available to participants on 30 August 2005. For a ICE 2005 participant's results to be included in the initial analysis in this paper, complete similarity matrices or a complete set of quality scores needed to be submitted to the first author by 3 March 2006. The initial analysis was presented at the Second Iris Challenge Evaluation Organizational Workshop held on 23 March 2006, Participants could submit results for either of Exp. 1, Exp. 2, or both. Participants could submit results for multiple algorithms for an experiment. All groups that submitted results agreed in advance that their performance results would be attributed.

V. PARTICIPANTS

The ICE 2005 was open to academic institutions, research laboratories and to companies worldwide. Results were submitted from nine groups from six countries. The full list of participants is given in Table II. The rightmost column lists the symbol for each algorithm used in the legends of the figures in this paper. If more than one symbol is listed, then a group submitted results for more than one algorithm.

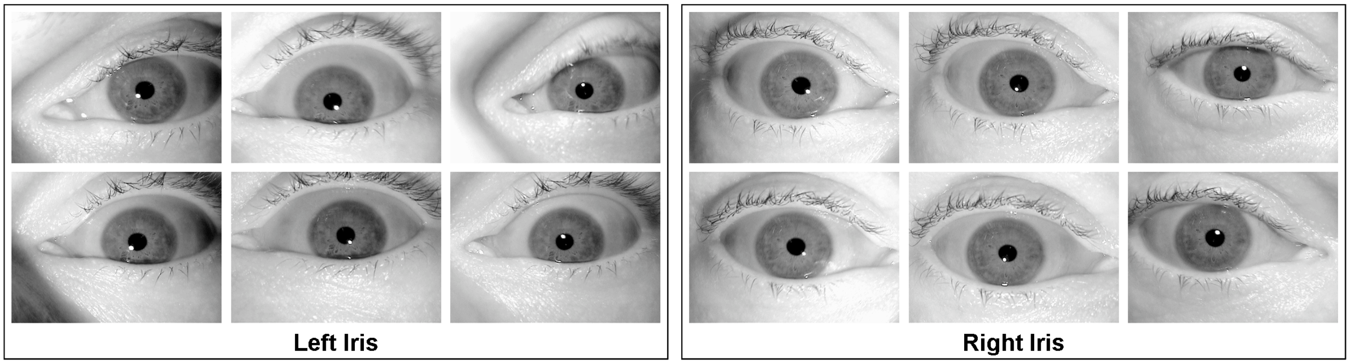


Fig. 1. Example of iris images from an acquisition session.

TABLE II
LIST OF PARTICIPANTS IN ICE 2005 AND SYMBOLS USED IN FIGURE LEGENDS.

Group	Legend symbols
Cambridge University	Cam 1, Cam 2
Carnegie Mellon University [27]	CMU
Chinese Academy of Sciences, Center for Information Science	CAS 1, CAS 2, CAS 3
Indiana University, Purdue University, Indianapolis	IUPUI
Iritech	IritchD
PELCO	Pelco
SAGEM - Iridian	SAGEM
West Virginia University [13][30]	WVU
Yamataki Corp / Tohoku University [18]	Tohoku
National Institute of Standards and Technology	irisBEE

VI. ANALYSIS

The ICE 2005 reported algorithm performance and analyzed the correlation between the right and left irises for match and non-match scores, and quality measures.

A. Receiver Operating Characteristics

For algorithm performance, the ICE 2005 reported receiver operating characteristics (ROCs) for a verification task. In the verification task, an algorithm compares a query image q_j to a target image t_i and produces a similarity score s_{ij}. A similarity score is a measure of the sameness of identity of the individuals appearing in two iris images. A large similarity score implies that the identifies are more likely to be the same. Algorithms could report either a similarity score or distance measure. Distance measures, where a small value indicates sameness of identity, have their values negated before any processing. The verification task models the situation were a person presents a biometric sample q_j to a system with a claimed identity. The system either accepts or rejects the claim. If t_i is the enrolled biometric sample of the person with the claimed identity, then the claim is accepted if the similarity score s_{ij} comparing the samples q_j and t_i is greater than or equal to a threshold τ. The threshold τ is the system's operating point. Verification performance is quantified by two performance measures. The first is the false accept rate (FAR). A false accept occurs when an imposter claims an identity and is matched by the system above threshold. The second is the verification rate (VR). A successful verification occurs when the system correctly matches two iris images of an individual above threshold.

The ROC is computed to quantify verification performance. It shows the tradeoff between the verification performance measures by plotting estimates of the VR against the FAR as a parametric function of an operating threshold, τ. The VR is the fraction of match similarity scores greater than or equal to a threshold value τ:

$$\mathrm{VR}(\tau) = \frac{|\{s_{ij} \geq \tau, \quad \text{where } s_{ij} \in M\}|}{|M|}, \tag{1}$$

where M is the set of match similarity scores. In a match similarity score s_{ij}, the two images q_j and t_i are of the same individual. The FAR is the fraction of non-match similarity scores greater than or equal to a threshold value τ:

$$\mathrm{FAR}(\tau) = \frac{|\{s_{ij} \geq \tau, \quad \text{where } s_{ij} \in N\}|}{|N|}, \tag{2}$$

where N is the set of non-match similarity scores. In a non-match similarity score s_{ij}, the two images q_j and t_i are of different individuals.

Verification performance is reported in Figures 3 and 4. Figure 3 reports ROCs for Exp. 1 and 2. For clarity, results are reported at two different scales for the verification rate (the vertical axis). Figure 4 reports the verification rate at a false accept rate of 0.001. For the top performers in Experiment 1 (right eye) the verification rate at a FAR of 0.001 is above 0.995 and for Experiment 2 (left eye) the verification rate is between 0.990 and 0.995. For the algorithms with a verification rate above 0.95 at a FAR of 0.001, all except CAS 1 and CAS 3, the observed recognition rate from the right iris is higher than the left iris.

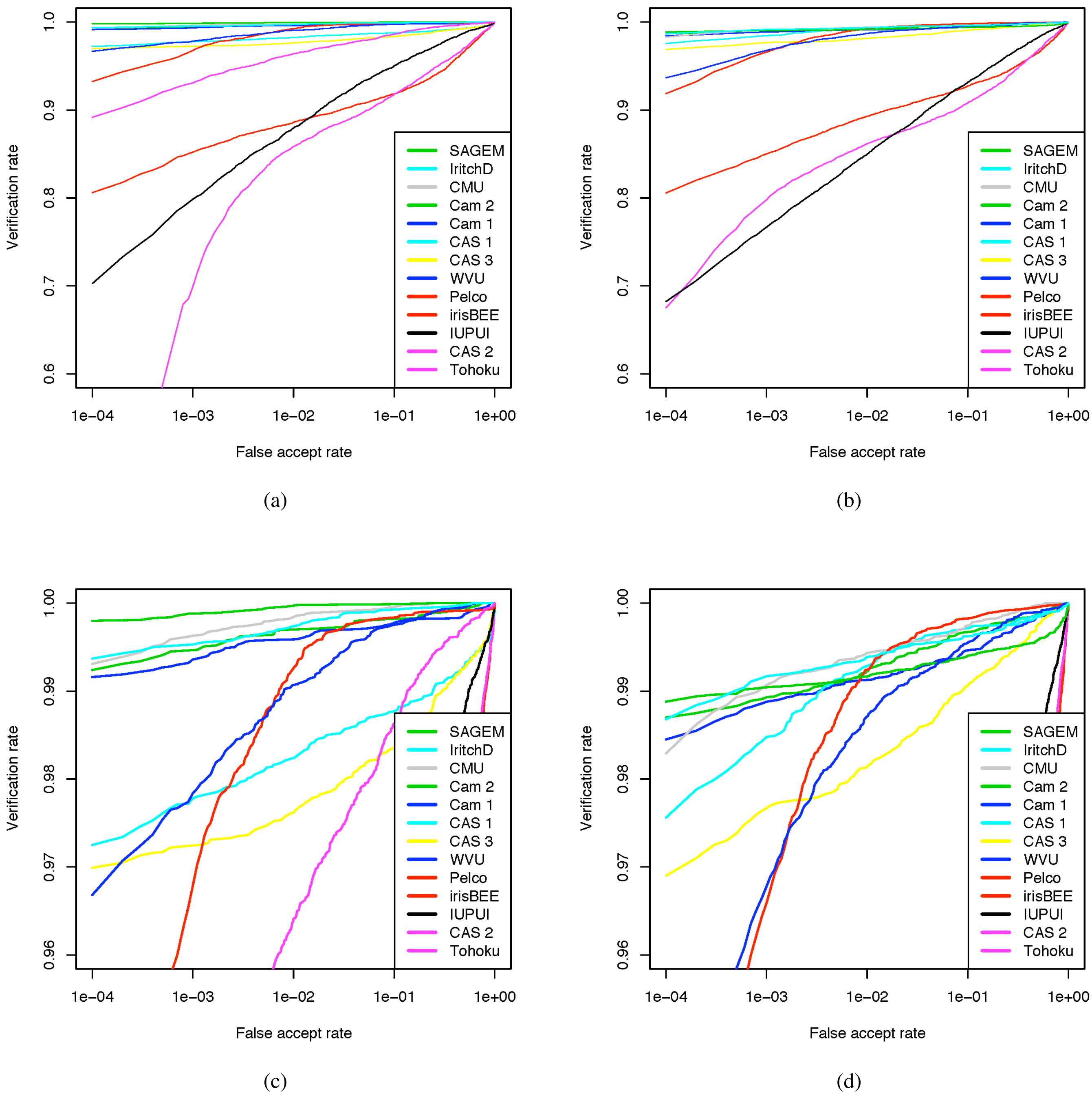


Fig. 3. ROC for ICE 2005 Experiments 1 (right eye) and 2 (left eye). Graphs (a) and (b) report results for Experiments 1 and 2 with the vertical axes scaled between 0.6 and 1.0 for the verification rate. Graphs (c) and (d) report results for Experiments 1 and 2 with the vertical axes scaled between 0.96 and 1.00 for the verification rate.

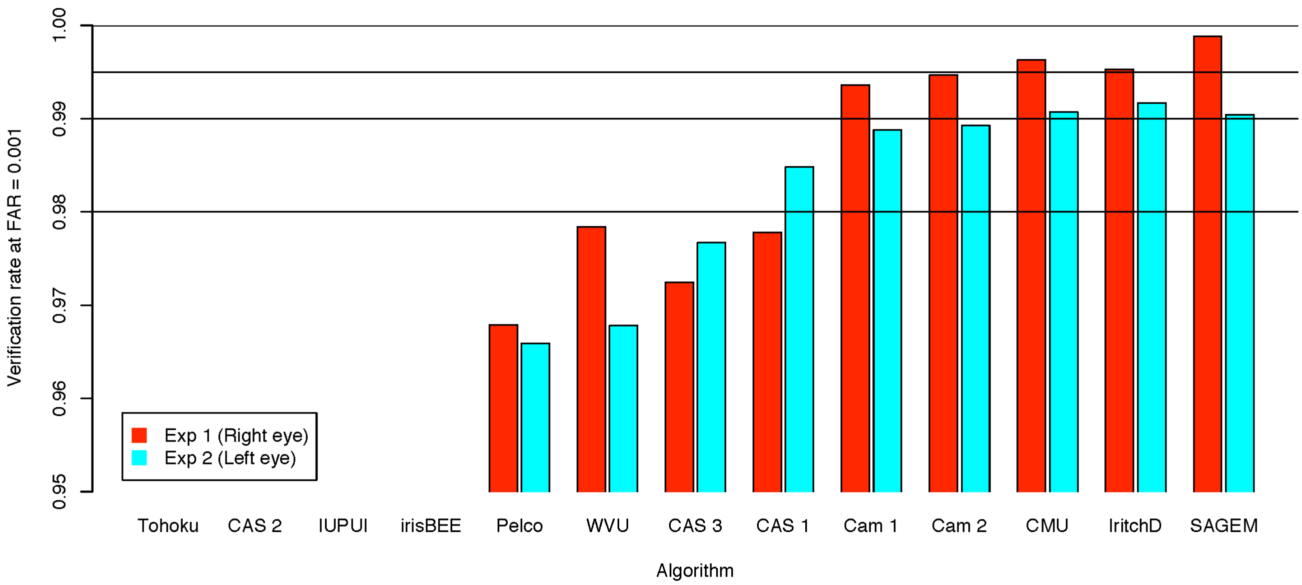


Fig. 4. Barplot showing verification rate at a FAR of 0.001. Performance is broken out by Experiment 1 (right eye) and 2 (left eye).

B. Match and Non-match Score Analysis

The majority of people have two irises, and prior to the ICE 2005 it was accepted wisdom in the iris community that the recognition rates for the two irises of a person are independent. In the section we look at the correlation between a person's left and right eye for match, non-match, and quality scores.

Correlations between right and left eyes is made by computing average match, non-match, and quality scores for each subjects' left and right eye. Then Pearson's correlation coefficient is computed between the average right and left match scores (resp. non-match and quality scores) over the subjects in the ICE 2005 dataset.

The average match and non-match scores are computed directly from a similarity matrix s_{ij}, where s_{ij} is the similarity score between target image t_i and query image q_j. The average match score for the right eye of subject k is $\widehat{\mu}^r_m(k) = \frac{1}{|\Omega^r_m(k)|}\sum_{\Omega^r_m(k)} s^r_{ij}$, where $\Omega^r_m(k)$ is the set of match scores for between right iris images of subject k and s^r_{ij} is a similarity matrix for the right eye (Exp. 1). The average match score for the left eye $\widehat{\mu}^l_m(k)$ of subject k is defined in an analogous manner.

The average non-match score for the right eye of subject k is $\widehat{\mu}^r_n(k) = \frac{1}{|\Omega^r_n(k)|}\sum_{\Omega^r_n(k)} s^r_{ij}$, where $\Omega^r_n(k)$ is the set of non-match scores where either target image t_i or query image q_j is an iris image of subject k. The average non-match score for the left eye $\widehat{\mu}^l_n(k)$ of subject k is defined in an analogous manner.

The average quality score for the right eye of subject k is $\widehat{\mu}^r_q(k) = \frac{1}{|\Omega^r_q(k)|}\sum_{\Omega^r_q(k)} q(i)$, where $\Omega^r_n(k)$ is the set of right iris images of subject k and $q(i)$ is the quality score for iris image i. The average quality score for the left eye $\widehat{\mu}^l_q(k)$ of subject k is defined in an analogous manner.

Figures 5 and 6 examine the correlation between the average subject match and non-match scores for the right and left irises. Figures 5 shows scatter plots for two algorithms, Iritech D and CAS 3. Figure 5(a) and 5(c) show the correlation between average match scores for each subject's right and left iris by plotting $\widehat{\mu}^r_m(k)$ versus $\widehat{\mu}^l_m(k)$. To illustrate the correlation, the regression line between $\widehat{\mu}^r_m(k)$ and $\widehat{\mu}^l_m(k)$ has been included (in the scatter plots, each dot corresponds to a subject). Figure 5(b) and 5(d) show the correlation between average non-match scores for each subject's right and left iris by plotting $\widehat{\mu}^r_n(k)$ versus $\widehat{\mu}^l_n(k)$. The regression lines have been added. Figures 5(b) and 5(d) show the range of correlation for the average subject non-match scores.

Figure 6 summarizes the correlation between the average subject match and non-match scores for the right and left irises. Pearson's correlation coefficient is plotted for each algorithm for both the match and non-match scores.

Two iris image quality measures were submitted by the West Virginia University (WVU) [13], [30]. Figure 7 plots the correlation between the average quality score for both the right and left irises for each subject. Formally this is a scatterplot of $\widehat{\mu}^r_q(k)$ versus $\widehat{\mu}^l_q(k)$. A regression has been added to show the correlation.

The ICE 2005 results show that average match scores between left and right eye were correlated for all algorithms. With the exception of the algorithms from the Chinese Academy of Sciences, Center for Information Science and the WVU, average non-match scores between left and right eye were correlated. For the two quality measures from the WVU, the average quality scores between left and right eye were correlated. These results support the hypothesis that

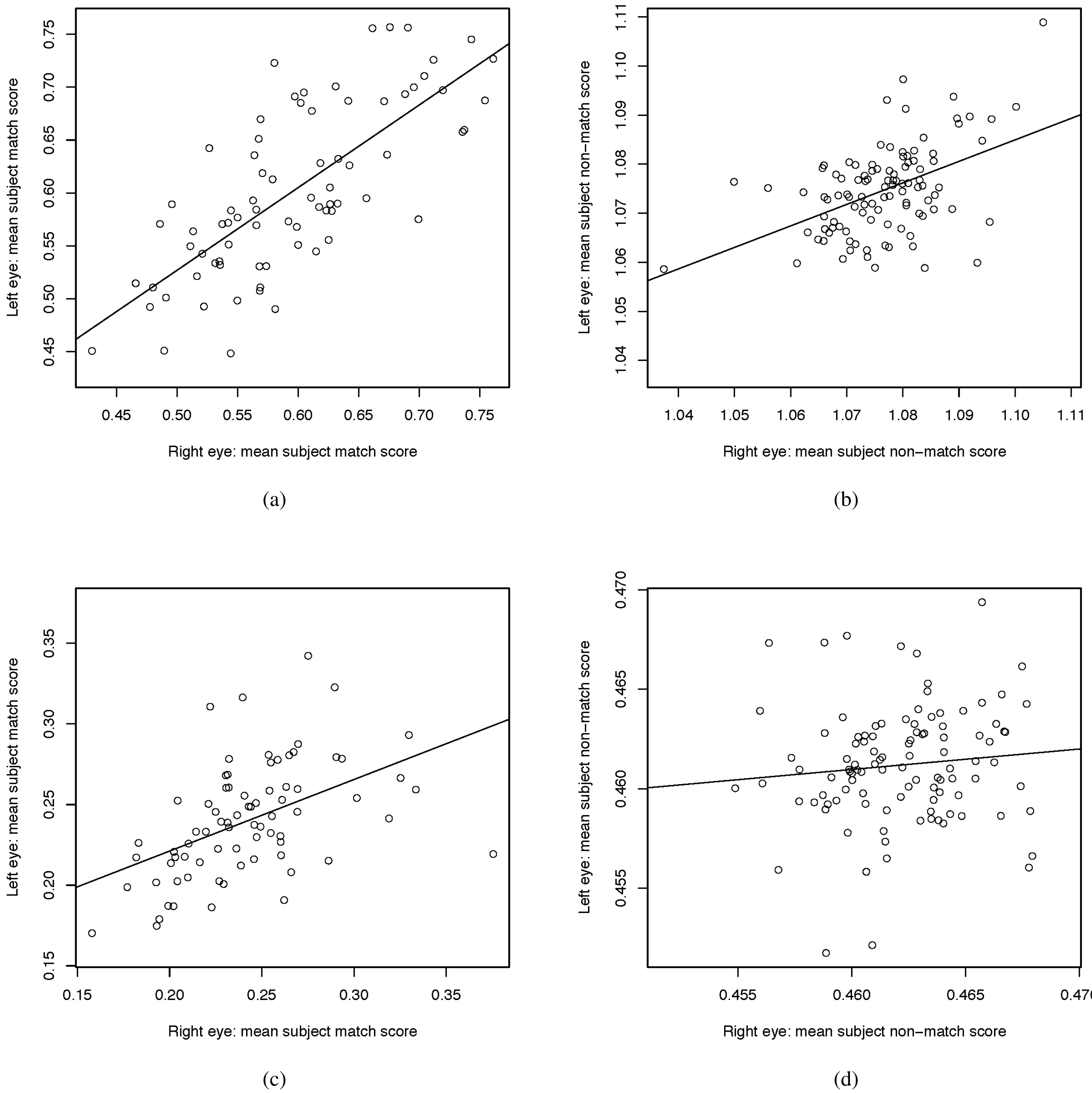


Fig. 5. Subject correlation for Iritech D for right (Exp. 1) and left (Exp. 2) eyes for (a) mean match scores and (b) mean non-match scores. Subject correlation for CASIA 3 for right (Exp. 1) and left (Exp. 2) eyes for (c) mean match scores and (d) mean non-match scores.

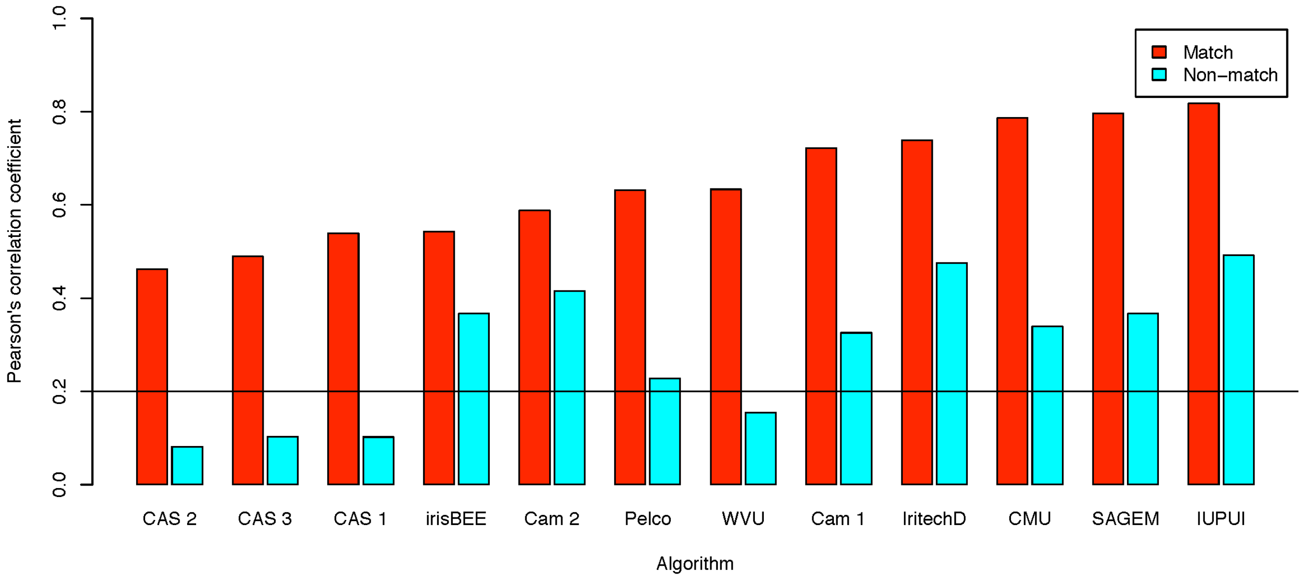
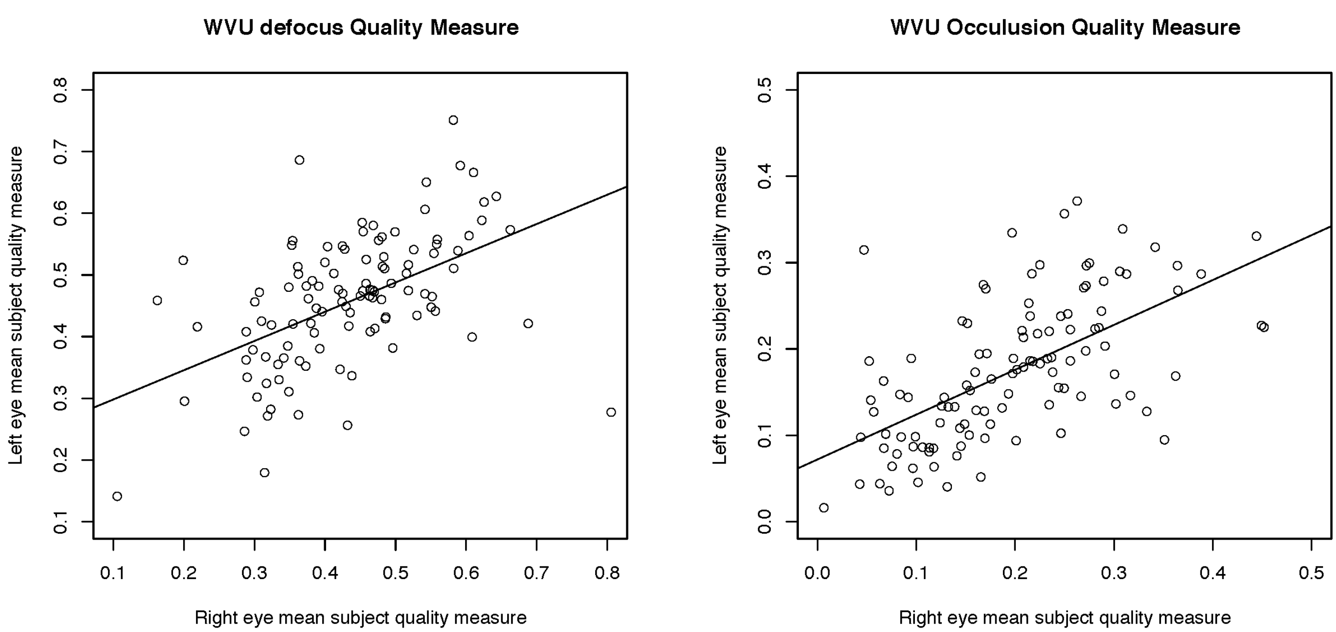


Fig. 6. Barplot Pearson's correlation coefficient for mean subject scores between right (Exp. 1) and left (Exp. 2) eyes. Correlation is reported for both mean match and non-match scores. The horizontal line is at the $p = 0.05$ significance level.



(a)

(b)

Fig. 7. Subject correlation for right and left irises for (a) WVU defocus and (b) WVU occulusion quality scores.

similarity scores between a subject's right and left eyes are correlated. These correlations suggests that fusing right and left eyes for recognition may not be equivalent to fusing two independent irises.

Daugman [6] studied the similarity score distribution of matching iris images of a subject's right to left irises and "found their distribution was statistically indistinguishable from the distribution for unrelated eyes." This result implies that one cannot search a database that contains a subject's left (resp. right) iris and expect to get a match with that subject's right (resp. left) iris. This is different from our results which impact the effectiveness of recognition from fusing the left and right irises of a subject.

VII. CONCLUSIONS

The ICE 2005 was the first iris challenge problem. The challenge problem provided the first open benchmark for iris recognition algorithms. The results from the ICE 2005 challenge problem were the first to observe correlations between the right and left irises for match and non-match scores, and quality measures. The ICE 2005 is still influencing the direction of research in iris recognition and processing as seen by continuing requests for the ICE 2005 challenge problem and number of publications reporting results on either the formal ICE 2005 experiments or on the dataset.

VIII. ACKNOWLEDGMENTS

We acknowledge the support of Department of Homeland Security's Science and Technology Department and Transportation Security Administration (TSA), the Director of National Intelligence's Information Technology Innovation Center, the Federal Bureau of Investigation (FBI), the National Institute of Justice, and the Technical Support Working Group (TSWG). The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, SAIC, or U. of Notre Dame.

REFERENCES

[1] Authenti-Corp. Iris recognition study 2006 (IRIS06). Technical Report version 0.40, Authenti-Corp, March 2007.

[2] C. Belcher and Y. Du. Selective feature information approach for iris image quality measure. *IEEE Transactions on Information Forensics and Security*, in press.

[3] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn. Image understanding for iris biometrics: a survey. *Computer Vision and Image Understanding*, 110(2):281–307, 2008.

[4] J. Bringer, H. Chabanne, G. Cohen, B. Kindarji, and G. Zémor. Optimal iris fuzzy sketches. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, September 2007.

[5] Chinese Academy of Sciences. Note on CASIA-irisv3. http://www.cbsr.ia.ac.cn/IrisDatabase.htm (accessed March 15, 2007).

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

[7] J. Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(5):1167–1175, October 2007.

[8] J. Daugman and C. Downing. Effect of severe image compression on iris recognition performance. *IEEE Transactions on Information Forensics and Security*, 3(1):52–61, March 2008.

[9] M. Dobeš and L. Machala. UPOL iris database. http://phoenix.inf.upol.cz/iris/ (accessed June 20, 2008).

[10] M. Dobeš, L. Machala, P. Tichavský, and J. Pospíšil. Human eye iris recognition using the mutual information. *Optik*, 115(9):399–405, 2004.

[11] P. J. Flynn and P. J. Phillips. ICE mining: Quality and demographic investigation of ICE 2006 performance results. Technical report, National Institute of Standards and Technology, 2008.

[12] International Biometric Group. Independent testing of iris recognition technology. Technical report, International Biometric Group, May 2005.

[13] N. D. Kalka, J. Zuo, V. Dorairaj, N. A. Schmid, and B. Cukic. Image quality assessment for iris biometric. In *SPIE Conf. on Biometric Technology for Human Identification III*, volume 6202, pages 61020D–1–62020D–11, 2006.

[14] R. Kerekes, B. Narayanaswamy, J. Thornton, M. Savvides, and B. V. K. Vijaya Kumar. Graphical model approach to iris matching under deformation and occlusion. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.

[15] E. Krichen, S. Garcia-Salicetti, and B. Dorizzi. A new probabilistic iris quality measure for comprehensive noise detection. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, September 2007.

[16] L. Ma, T. Tan, Y. Wang, and D. Zhang. Personal identification based on iris texture analysis. *IEEE Trans. Pattern Analysis Machine Intelligence*, 25:1519–1533, 2003.

[17] L. Masek. Recognition of human iris patterns for biometric identification. Technical report, Bachelor of Engineering degree of the School of Computer Science and Software Engineering, The University of Western Australia, 2003.

[18] K. Miyazawa, K. Ito, T. Aoki, K. Kobayashi, and H. Nakajima. An effective approach for iris recognition using phase-based image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 29, 2007.

[19] E. M. Newton and P. J. Phillips. Meta-analysis of third-party evaluations of iris recognition. *IEEE trans. on Systems, Man and Cybernetics Part A*, In press.

[20] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 947–954, 2005.

[21] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. Technical Report NISTIR 7408, National Institute of Standards and Technology, 2007.

[22] H. Proenca and L. Alexandre. Iris recognition: An entropy-based coding strategy robust to noisy imaging environments. In *Lecture Notes in Computer Science 4841: Advances in Visual Computing*. Springer, 2007.

[23] H. Proença and L. A. Alexandre. UBIRIS: a noisy iris image database. In *ICIAP 2005: 13th International Conference on Image Analysis and Processing*, volume 1, pages 970–977, 2005.

[24] K. Roy and P. Bhattacharya. Application of multi-objective genetic algorithm and asymmetrical support vector machine to improve the reliability of an iris recognition system. In *International Conference on Systems, Man and Cybernetics*, pages 1952–1957, October 2007.

[25] K. Roy and P. Bhattacharya. Collarette area localization and asymmetrical support vector machines for efficient iris recognition. In *International Conference on Image Analysis and Processing*, pages 3–8, September 2007.

[26] K. N. Smith, V. P. Pauca, A. Ross, T. Torgersen, and M. C. King. Extended evaluation of simulated wavefront coding technology in iris recognition. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–7, September 2007.

[27] J. Thornton, M. Savvides, and B. V. K. Vijaya Kumar. A bayesian approach to deformed pattern matching of iris images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:596 – 606, 2007.

[28] J. Thornton, M. Savvides, and B. V. K. Vijaya Kumar. An evaluation of iris pattern representations. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–5, September 2007.

[29] University of Bath. University of Bath iris image database. http://www.bath.ac.uk/elec-eng/research/sipg/irisweb/index.htm (accessed June 20, 2008).

[30] J. Zuo, N. Kalka, and N.A.Schmid. A robust iris segmentation procedure for unconstrained subject presentation. In *Biometric Consortium Conf., Special Session on Reserch*, pages 1–6, 2006.