An Assessment of the Human Performance of Iris Identification

Richard M Guest, Hongmei He
School of Engineering and Digital Arts
University of Kent
Canterbury, Kent, UK
r.m.guest@kent.ac.uk

Sarah Stevenage, Greg J Neil
Department of Psychology
University of Southampton
Southampton, UK
S.V.Stevenage@soton.ac.uk

***Abstract*—Biometric iris recognition systems are widely used for a range of identity recognition applications and have been shown to perform with high accuracy. For large-scale deployments, however, system enhancements leading to a reduction in error rates are continually sought. In this paper we investigate the performance of human verification of iris images and compare against a standard computer-based method. Our results suggest that performance using the computer-based system is no better than performance of the human participants. Additionally and importantly, however, performance can be improved through incorporation of the human as a 'second decision maker'. This fusion system yields a false acceptance rate of just 9% when disagreements are resolved in line with strengths of each 'decision-maker'. The results are presented as an illustration of the benefits that can be gained when combining human and automated systems in biometric processing.**

Keywords-biometrics; iris verification; human assessment; data fusion

I. INTRODUCTION

Within the commercial domain, iris recognition has proven itself as being a valuable and reliable means of biometric verification and identification. The automated analysis of iris images are now used within systems that control access to secured services and sites, and even within the suite of biometric checks performed at border control [1]. This use of the iris as a biometric typically entails robust implementation of four key tasks: iris acquisition, iris segmentation, texture analysis, and the matching of texture representations [2]. Although all tasks are important, this latter task is critical, and underpins the determination of whether the iris code presented by an individual matches a stored representation or not. Accordingly, it is this latter task which represents the focus for the current paper, and we present what we believe to be the first attempt to improve this matching stage through the fusion of human and automated decision making.

A. Automated IrisCode Assessment

Although many techniques exist for assessing iris texture, IrisCode [3] is one of the best known and most widely used automated methods. Within this method, having captured an iris image, the iris zone is unravelled and size-normalised, following which an IrisCode – a binary code which describes the iris texture in a way that is said to be unique to that iris – is extracted. One of the most widely used distance metrics used to establish provenance between pairs of images (for example an enrolment image and a probe image) is the Hamming distance. This metric measures the number of substitutions that need to be made to convert one string to the second string. When normalised, this Hamming distance is zero for two identical iris images, but it increases towards 1 as the two images become more different. Assessing this distance with respect to a threshold enables a verification to be established.

The IrisCode system now represents a universally recognized approach to iris texture analysis, and the results of this system have been variously combined with additional metrics such as the local binary pattern (LBP) [4] and Euler codes [5] each of which captures global characteristics of the iris texture. More recently, the Hamming distance scores have been combined with a measure related to 'fragile bits' – those elements of the IrisCode which are not consistent from one image of an iris to another [6]. Rather than mask these fragile bits, Hollinsgworth, Bowyer and Flynn [1] used the coincidence of these bits across images to improve the iris matching decisions. Consequently, there is a precedent of improvement to iris processing through a fusion approach. The work presented in this paper has much in common with this approach and explores the potential gains that are made possible when decisions based on Hamming distance scores are supplemented by the judgments of human observers.

B. Human Perception

On the basis of a distance score (Hamming or otherwise), the iris textures of two genetically identical eyes (right and left eye of a single individual, or of identical twin individuals) are as different to one another as the irises of two unrelated people. However, recent exploration of iris processing amongst human perceivers suggests that humans may detect similarities between the left and right irises of an individual [7], or between the irises of identical twins [8], that an algorithm misses. Indeed, Bowyer and colleagues provide evidence to suggest that human perceivers show 86% accuracy when determining whether two irises come from the same person or not, and that this performance increases to 93% accuracy when only their confident answers are taken into consideration.

This work has been supported by the UK Engineering and Physical Sciences Research Council.

These results carry considerable importance because they suggest that there are aspects of iris texture similarity that the human is capable of perceiving but which are missed by computer-based techniques. Within a real world context, these results suggest that the human perceiver may make fewer false-rejection errors because they are capable of detecting similarities between two iris images where the machine fails.

With this in mind, the goal of the current paper is to explore the levels of performance displayed by both computer-based and human perceivers for an iris-matching task. However, we go further through exploration of the capacity to improve iris matching by utilising a fusion approach in which human and computer-based performances are combined. Rather than generate a fusion score through use of a cascaded classifier (in which the decisions of one system are refined by another) or a weighted average (in which inputs from both systems contribute to a weighted fusion score), we use here a method more analogous to a committee model in which two independent classifiers (the IrisCode/Hamming System and the human perceiver) operate, and disagreements are resolved through weighting the human input. To our knowledge, this work represents the first attempt to improve iris texture matching in this way and, if successful, the results will hold real world significance in a number of domains including issues of admissibility in court settings.

The remainder of this paper is organised as follows: Section II details the iris images together with the design adopted for all experimentation. Section III provides the methodological details both for the application of the IrisCode/Hamming distance system and the testing of human participants. Section IV presents the results of all experimentation, with particular emphasis on the exploration of a fusion approach. Finally, Section V discusses the implications of these results before describing ideas for future work.

II. Data Acquisition and Experimental Design

Two right-eye infrared images from each of 104 database subjects were used within all the experimentation in this study. The images were collected using a VistaEY2 iris scanner at a standard distance of 50cm with a resolution of 640 x 480 pixels. They were selected from a larger database [9] on the basis of human inspection of image clarity and lack of distinguishing ocular features such as stys or heavy makeup. All individuals were Caucasians and ages ranged from18-35 years with a mean of 21.8 and SD 3.06. Iris images were cropped within Gimp 2.0 to extract a rectilinear portion of the image that completely contained the iris itself plus some extent of eyelash or lid as required.

For each of the 104 subjects in our test dataset, matching and nonmatching trials were created to yield 208 trials in total.

For *Matching trials* two separate right eye images for the same person were presented. The correct response would be to indicate that the individuals in each image were the 'same' whilst an error would amount to a 'false rejection'. In using two separate images of the same eye this minimised the possibility that performance could be based on image cues rather than iris cues.

Nonmatching trials were created by pairing a 'genuine' image from each database subject with a foil or distractor iris belonging to another subject within the database. The correct response would be to indicate that the individuals in each image were 'different' and an error would amount to a 'false acceptance'. The foil was always carefully selected based on the judgements of 3 human experimenters so that the target and the foil together displayed high similarity (taking account of image quality, apparent iris color, makeup, extent of pupil dilation, apparent iris texture and pattern). Consequently, the nonmatching trials were not trivially easy to complete.

Fig I shows an example of a trial pair from a Matching and Non-matching trial.

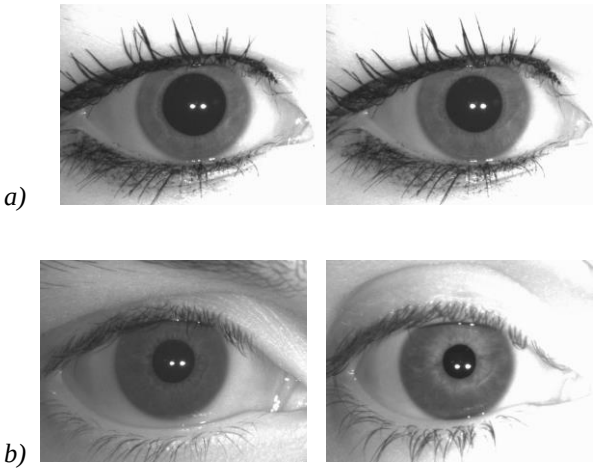

FIG. I. Example Iris Pairs a) Matching Trial b) Non-Matching Trial

The computer-based system was presented with all 208 trials. However, human participants were presented with 52 trials only, ensuring that each iris was seen only once to prevent learning, fatigue, or mere exposure effects. Across all human participants in the study, each iris was presented as target and as test, in both matching and nonmatching trials yielding the 208 combinations described above.

III. Experimental Method

IrisCode/Hamming Distance System: The computer-based analysis was applied to all 208 trials, and a normalised Hamming distance score was extracted for each of the 104 'matching' trials and the 104 'nonmatching' trials. Through application of an ROC analysis to optimise the proportion of false rejection and false acceptance errors, a threshold Hamming distance of 0.409 was calculated. Scores above this threshold were judged to indicate a 'different' decision and scores below this threshold were judged to indicate a 'same' decision. In this way, the continuous Hamming distance scores were converted to provide a binary same/different outcome. Finally, percentage accuracy of these outcomes for matching and nonmatching trials was determined through comparison with ground truth for each trial.

Human Procedure: A total of 32 participants (9 males, 23 females) took part on a volunteer basis or in return for course credit. Ages ranged between 19 and 55 years (Mean = 26.1, SD = 10.3), and all participants had normal, or corrected-to-normal, vision. None of the participants had medical expertise and they were unlikely to have had experience in discriminating iris texture patterns before.

Participants were tested individually within a small and quiet experimental cubicle. Following instruction, and three practice trials, participants were presented with 52 experimental trials (26 matching and 26 non-matching trials) in which each iris was presented only once.

With the iris images displayed on a computer screen, the format of each trial was identical and consisted of the presentation of a 'ready' prompt for 1000ms to orient attention, and then the presentation of the first iris image of a pair for 3 seconds on the left hand side of the screen. Participants were asked to study this iris pattern. This was replaced by the second iris image on the right hand side of the screen, and participants were asked whether this second iris was the 'same as' or 'different to' the first iris image previously shown. Participants responded as quickly but as accurately as possible by pressing either 'S' (same) or 'D' (different) on the computer keyboard. The second iris image remained onscreen until response, and accuracy of response was recorded. Finally, participants were asked to indicate their confidence in each decision by pressing a numbered key between 1 (not at all confident) and 7 (certain). The entire experiment lasted 10 minutes and no signs of fatigue were apparent.

IV. RESULTS

Performance was evaluated for 'matching' and 'nonmatching' trials for the IrisCode/Hamming system and separately for the human participants. In this respect, the data should be regarded as representing the outputs of two independent decision-makers, and comparison of the two decision makers across a common set of trials is possible when the community of human participants is taken as a whole. This requires that the human participant data are analysed 'by items' rather than 'by participant' and this draws a distinction between the examination of performance for an iris trial across all individuals (by items) rather than performance for an individual across all iris trials (by participants).

In order to calculate human performance levels 'by items', the average accuracy was calculated score across the 32 participants for each iris trial. This was converted to a binary correct/incorrect decision through application of a threshold accuracy level, and that threshold (61%) was determined based on the number of observed errors across a population which would be required in order to differ significantly from chance. Where this threshold was met or exceeded, the community of participants demonstrated accuracy for that trial. Conversely, where this threshold was not met, the community of participants demonstrated an error for that trial. Treatment of the data in this way enabled exploration of the number of trials (out of 208) for which i) the computer-based system made an error, ii) the human made an error, iii) both made an error, or iv) neither made an error. These data are summarised in Table I below. The DET curve for the computer-based Hamming distance system is show in Fig II.

TABLE I. ACCURACY OF COMPUTER-BASED AND HUMAN EVALUATION OF IRIS IMAGES

	Number of errors by system			
	Computer-based System	***Human Inspection***	***Both***	***Neither***
False Rejection Error	15	1	0	192
False Acceptance Error	9	25	9	168
Overall Accuracy	84.1%	83.2%	-	-

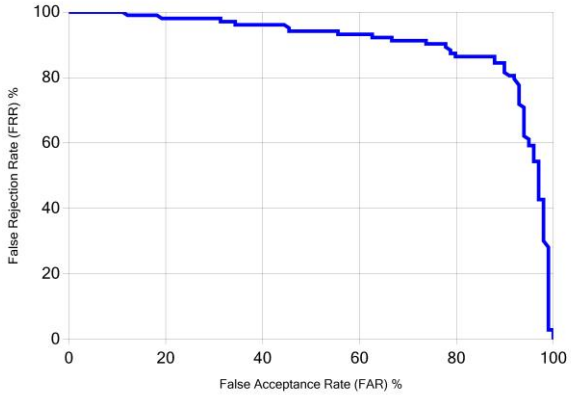

FIG. II. DET CURVE OF COMPUTER-BASED IRIS PERFORMANCE

The results suggested an overall accuracy level of 84.1% for the computer-based IrisCode/Hamming system alone, and an overall accuracy level of 83.2% for the human participants. Statistical analysis by means of a chi-squared test confirmed no significant difference in performance ($\chi^2 = 0.07$, $DF = 1$, $p = 0.79$) suggesting that the human perceiver was as good as the automated system. These data correspond well to the published level of human performance (86%) noted by Bowyer, Lagree and Fenker [7].

What is of greater interest however, is the pattern of errors shown by the computer-based system and the human participants. In this regard, it was clear from comparison to ground truth that human participants showed a very low level of false acceptance errors, whereas the computer-based system showed both false acceptance and false rejection errors. As a consequence, a Fusion System was explored in which the decisions of the computer-based were refined by the inclusion of a second, human decision-maker. Specifically, in cases of

agreement, no adjustments to a decision were made, but in cases of disagreement, the outcome for a 'matching' trial reflected the human decision, whilst the outcome for a 'nonmatching' trial reflected the IrisCode/Hamming system decision.

This resulted in a correction of 15/33 Hamming System errors and 25/35 human errors, leaving only 19 errors in total (1 false rejection, 18 false acceptances) and an overall accuracy level of 90.1%. Again, statistical analysis by means of a chi-squared test revealed this Fusion Model to represent a significant improvement on both the Hamming system (χ^2 = 4.31, DF = 1, p = 0.038) and the human performance (χ^2 = 5.45, DF = 1, p = 0.02) when taken in isolation.

V. Implications and Future Work

The results presented here suggest that significant improvements can be made in an iris matching task when IrisCode/Hamming distance scores are combined with human decision making. Moreover, the comparability of the current human performance levels with those reported by Bowyer et al. [7] gives cause for confidence in the analysis presented here. Indeed, just as Bowyer et al. demonstrated an increase in human accuracy from 86% to 93% when only 'confident' decisions were taken into account, this pattern was echoed in our own data: Accuracy significantly increased from 79.3% to 82.2% when only decisions attracting a confidence level of 4 or greater (on a 7 point scale) were considered ($t(31) = 4.18$, $p < 0.001$), and this was carried by a significant increase from 90.1% to 93.1% in accuracy for 'matching trials' alone ($t(31) = 3.31$, $p < 0.005$).

The demonstration here of improved performance for our fusion system echoes the work of other researchers in suggesting that a combination of approaches can yield a superior performance than either approach in isolation. More specifically, the combination of man and machine here sits well with more recent work by Stark and colleagues [10] which indicates the potential for human iris texture classifications to support a guided search rather than an exhaustive search of an iris gallery or database.

Where these results perhaps carry most significance is in terms of consideration of admissibility within a legal setting, and with iris matching being relied on more and more in security contexts, it is not difficult to imagine a case where an attempted breach of security results in criminal proceedings. In such a situation, and in common with the consideration of fingerprint evidence, a judge will only deem evidence to be admissible if it is based on human judgement. Consequently, there is a clear emergent need to demonstrate human involvement in the iris matching task. The fusion system we present here reflects this through recognising the role of the human when both man and machine are correct, tolerating the fallibility of the human when both man and machine are wrong, and through relying on the human in the particular category of cases where the machine is demonstrably weaker. As such, the work presented here forms a first but important step in improving iris recognition per se and in establishing its legal admissibility within a forensic setting.

Acknowledgment

The authors thank Emma White at the University of Southampton for assistance with the data collection from human participants.

References

[1] Hollingsworth, K.P., Bowyer, K.W., & Flynn, P.J. Using fragile bit coincidence to improve iris recognition. Third IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 09), September 2009, Washington, DC.

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

[3] Daughman J. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1), 21-30, 2004.

[4] Sun, Z., Tan, T., & Qiu, Z. Graph matching iris image blocks with local binary pattern. In *Proceedings of the International Conference on Biometrics Springer LNCS 3832)*, 366-372, Jan 2006.

[5] Zhang, PF, Li, SH & Wang, Q. A novel iris recognition method based on feature fusion. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, 3661-3665, 2004.

[6] Bolle, R.M., Pankanti, S., Connell, J.H., & Ratha, N. Iris individuality: A partial iris model In *Proceedings of the International Conference on Pattern Recognition*, II: 927-930, 2004.

[7] Bowyer K.W., Lagree, S., & Fenker, S.P. Human versus biometric detection of texture similarity in left and right irises. Security Technology (ICCST), 2010 *IEEE International Carnahan Conference*, San Jose, CA, 5-8th October 2010.

[8] Hollingsworth, K., Bowyer, K.W. Lagree, S., Fenker, S.P., & Flynn, P.J. Genetically identical irises have texture similarity that is not detected by iris biometrics. *Computer Vision and Image Understanding*, 115, 1493-1502, 2011.

[9] Neil, G.J., Stevenage, S.V., Black, S.M., et al. The SuperIdentity Stimulus Dataset – a Multi-Modal Biometric and Cybermetric Database. (in preparation).

[10] Stark L., Bowyer, K.W., & Siena, S. Human perceptual categorization of iris texture patterns. Fourth IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), September 2010.