HUMAN VS. AUTOMATIC MEASUREMENT OF BIOMETRIC SAMPLE QUALITY

Andy Adler
School of Information Technology
University of Ottawa, Canada
adler@site.uottawa.ca

Tatyana Dembinsky
School of Information Technology
University of Ottawa, Canada
tdembins@site.uottawa.ca

Abstract

Biometric systems are designed to identify a person based on physiological or behavioral characteristics. In order to predict the utility of a particular image for identification, there is an interest in measures to calculate the biometric image quality. Such measures often assume (implicitly or explicitly) that human image quality evaluations are a gold standard. In order to test this assumption, we measured biometric image quality for face and iris recognition by 8 human volunteers and from 6 face recognition and 1 iris recognition algorithm. Algorithm quality measures were based on a log-linear fit of quality to genuine score values. Results indicate that human quality scores correlate strongly with each other ($r=0.723$ (iris), $r=0.613$ (face), $p<0.001$). Algorithm scores also correlate strongly with each other ($r=0.534$, $p<0.001$ (face)). However, human quality scores do not correlate with those from algorithms ($r=0.234$ (face), $r=0.175$ (iris)).

***Keywords**: Biometric sample quality; human assessment; automatic measurement; face recognition; iris recognition.*

1. Introduction

Biometric systems are designed to identify a person based on physiological or behavioral characteristics [7]. Examples of such systems are automatic fingerprint, iris, and face recognition. Currently, such systems are seeing an increasing level of interest and applications for a wide variety of applications, from national identification applications, criminal searches, and for civil systems for employee identification and access control.

One recent development is the significant level of interest in standards for measurement of biometric quality. For example, ISO has recently established a biometric sample quality draft standard [2].

According to [2], biometric sample quality may be considered from the point of view of character (inherent features), fidelity (accuracy of features), or utility (predicted biometrics performance). A general consensus has developed that the most important measure of a quality metric is its utility – images evaluated as higher quality must be those that result in better identification of individuals, as measured by an increased separation of genuine and impostor match score distributions.

One common assumption is that human evaluations of biometric quality are an appropriate gold standard against which biometric sample quality may be measured. This is reflected in the use of the design of biometric quality standards (eg. [6]), and in the use of manual image quality verification in the workflow of many government immigration and passport agencies.

In this paper, we seek to test the relevance of human evaluations of biometric sample quality. We ask the following questions: 1) are human evaluators of image quality consistent with each other? 2) are quality metrics from biometric algorithms consistent with each other?, and 3) are human evaluators consistent with measures from biometric algorithms?

2. Methods

We compare biometric quality measures from human participants and from biometric algorithms for face and iris recognition. We choose not to evaluate fingerprint image quality, because fingerprints have a well defined forensic expert community, which would indicate that our experimental design would need to take into account the level of expertise of each examiner. Face and iris images, on the other hand, do not have a well defined expert community, and are common images which are seen every day by all examiners.

2.1. Image quality evaluation by human participants

Biometric image quality of each sample was assessed by human participants using a web based evaluation form shown in figure 1. In order to help ensure a uniform evaluation across samples and participants, each participant viewed a sample of the images beforehand. Instructions to participants were “Evaluate the quality of each image for a biometric identification application”

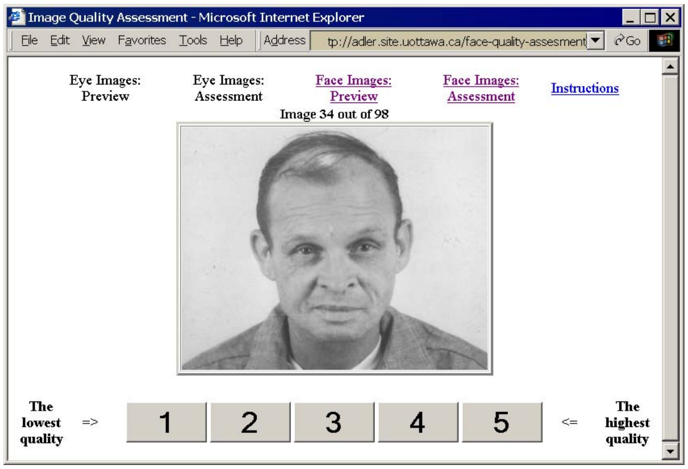

Figure 1. Web based evaluation form for human evaluation of biometric image quality. Participants assigned each image a quality level from 1 to 5.

2.2. Automatic Quality Measurement

Quality measures were derived from the biometric match score distributions. We wanted to model exclusively the *utility* aspect of quality measures, and to specifically not consider other aspects of biometrics images. Under this model, a quality measure defined to be the best predictor of identification performance, which is related to the separation of the genuine (comparison of images of the same person) and impostor (comparison of images of the different persons) match score distributions. The effect of biometric image quality on impostor distributions is not well understood; most reports seem to suggest little absolute change.

We model the impostor distribution as constant and independent of biometric image quality, while the genuine distribution increases with quality. Thus, we expect

$$E\left[MS_{i,j}\right] = E\left[Q_i Q_j\right] \qquad (1)$$

where MS is the match score ($0 \le MS \le 1$) calculated by an algorithm *A* between images *i* and *j*, and Q_i is the biometric image quality ($0 \le Q \le 1$) of image *i*. We calculate values of *Q* to satisfy (1), by minimizing

$$\sum_i \sum_{j, i \neq j} \left(MS_{i,j} - Q_i Q_j\right)^2 \qquad (2)$$

for all genuine comparisons *i* and *j*. Comparisons of identical images occur when $i = j$, and are excluded from this model, because most algorithms will give such a comparison a MS of 1.0, independent of the biometric image quality. Equation (2) is equivalent to minimizing the linear equation

$$\sum_{i,j} \left(\log MS_{i,j} - \log Q_i - \log Q_j\right)^2 \qquad (3)$$

To illustrate this calculation, we consider a similarity matrix between for images *i* of the same person. Equation (3) is minimized by solving the least square matrix equation:

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ \vdots & & & \vdots \end{bmatrix} \begin{bmatrix} \log Q_1 \\ \log Q_2 \\ \log Q_3 \\ \log Q_4 \end{bmatrix} = \begin{bmatrix} \log MS_{1,2} \\ \log MS_{1,3} \\ \log MS_{2,3} \\ \vdots \end{bmatrix} \qquad (4)$$

Thus,

$$\log \mathbf{Q} = \left(\mathbf{H}^t \mathbf{H}\right)^{-1} \mathbf{H}^t \left(\log \mathbf{MS}\right) \qquad (5)$$

where **Q** is a N×1 vector of biometric image quality measures, and **MS** is a M×1 vector of biometric match score comparisons for an algorithm. N is the number of biometric images, and M is the number of available match score comparisons. Matrix H (size M×N) has an element of value 1 at each position $\mathbf{H}_{k,j}$ or $\mathbf{H}_{i,j}$ when $\mathbf{MS}_k$ is between images *i* and *j*.

Using equation (5), we are able to calculate an estimate of the biometric image quality for each image for each algorithm. One limitation to this approach is the requirement of having more than two images of each individual. With only two images, equation (5) will be singular.

In order to calculate another measure of image quality, we use the IQM algorithm of [3]. The IQM software calculates a quality measure is based on the modulation transfer function (MTF) of an image, and is thus a close measure of image sharpness. We used version 7.1 of the IQM software to score each image face and iris image.

3. Results

Using each of the approaches discussed in the previous section a set of face and iris images was evaluated. Face images were selected from the Mugshot Identification Database [5]. Images from 29 different people using 3-4 samples per person were used, for a total of 98 images. All selected face images were of frontal pose, and samples from each person were selected in with approximately the same age. Iris images were selected from an internal image database captured using the L.G. camera. Iris images from 7 people with 6 samples from each eye were used, for a total of 84 images.

Eight participants took part in the tests; all were graduate students or researchers in electrical engineering of age 20-40; seven were male and one was a female.

Six different face recognition algorithms were used. All algorithms are proprietary vendor face recognition software products from three different vendors released during the period 1999–2005. Match score values were normalized to be in the range 0–1. The iris recognition algorithm used was from the software developed by Masek [4]. A match score was calculated from this algorithm as *1-HD*, where *HD* is the Hamming Distance parameter of by this algorithm.

The calculated biometric image quality values for each human participant, algorithm, and from the IQM software were compared by calculating the Pearson *r* correlation co-efficient. These correlation results allow us to address the questions which motivated the study.

Are human evaluators of image quality consistent with each other? Yes. Human evaluators are consistent, with an average correlation co-efficient of r=0.723 (iris), r=0.613 (face), p<0.001.

Are quality metrics from biometric algorithms consistent with each other? Yes. Face recognition algorithm scores also correlate strongly with each other (r=0.534, p<0.001). It was not possible to study correlation of iris recognition algorithms because only one algorithm was used. The largest correlation co-efficient (0.773) was between two recent software versions from the same vendor, but software from different vendors also had large correlations (0.672).

Are human evaluators consistent with measures from biometric algorithms? No. On average, human assessments of quality do not correlate or show very low correlation, with those from algorithms (r=0.234, p<0.05 (face), and r=0.175, not significant (iris)).

Correlations between humans, algorithms, and IQM values are shown in table 1 (for face) and table 2 (for iris). Quality values from IQM did not correlate with either human or algorithm assessments, except for human vs. IQM for iris images (r=0.458, p<0.001).

In order to further explore the features which motivate the various quality selections, the images with the highest and lowest rated quality are shown in figure 2 (face) and figure 3 (iris). We note that the same face is rated lowest by both software and humans. For iris images, there is a clear preference of humans for in-focus and sharp images, while the iris measures do not show any clear preference. This preference for sharp images appears to explain this correlation of humans and IQM. On the other hand, iris recognition algorithms are not very sensitive to image sharpness [1].

4. Discussion

This paper set out to assess the relationship between human and algorithm based measures of biometric image quality. Biometric image quality was evaluated from six face recognition and one iris recognition algorithm, by fitting a least squares linear model to normalized match score data from genuine comparisons, and human quality was assessed from eight volunteers, using a web-based tool. Results show that, in general, both algorithms and humans are consistent with others of the same group. On the other hand, correlations between different evaluator groups (humans, algorithms, IQM) showed low significance or no correlations. The only exception is the correlation between human and IQM evaluations of irises, which appear to be due to a preference for sharp images.

These results indicate that biometric image quality measurement is more subtle and difficult than may be expected. Naïve notions of techniques to evaluate images may not in fact be relevant to biometric algorithms. This effect is perhaps important to consider for face recognition applications, in which passport and national identification issuance authorities wish to maintain databases with the maximum value for face recognition applications. This paper suggests that human evaluation of submitted photo images may not be the best way to achieve that goal.

	Mean Algorithm	**IQM**
Mean Human	0.234	0.159
Mean Algorithm		0.003

Table 1. Correlation of biometric image quality measures for face images.

	Mean Algorithm	**IQM**
Mean Human	0.175	0.458
Mean Algorithm		-0.036

Table 2. Correlation of biometric image quality measures for iris images.

References

[1] J.G. Daugman, "How iris recognition works", *IEEE Trans. Circuits Syst. Video Technol.*, vol.14, no.1, pp.21–30, 2004.

[2] InterNational Committee for Information Technology Standards, “Biometric Sample Quality Standard Draft (Revision 4)”, *document number M1/06-0003*, February 7, 2005.

[3] The MITRE Corporation, Image Quality Measure (IQM)© Software, http://www.mitre.org/tech/mtf/.

[4] L.Masek, “Recognition of Human Iris Patterns for Biometric Identification”, BE Dissertation, The University of Western Australia, 2003, www.csse.uwa.edu.au/~pk/studentprojects/libor/index.htm

[5] NIST, *NIST Special Database 18: Mugshot Identification Database* (MID), http://www.nist.gov/srd/nistsd18.htm.

[6] E.Tabassi, C.L. Wilson, C.I. Watson, *Fingerprint Image Quality,* NISTIR 7151, NIST, August 2004.

[7] J.L. Wayman, “A Definition of Biometrics”, 2001, http://www.engr.sjsu.edu/biometrics/nbtccw.pdf.

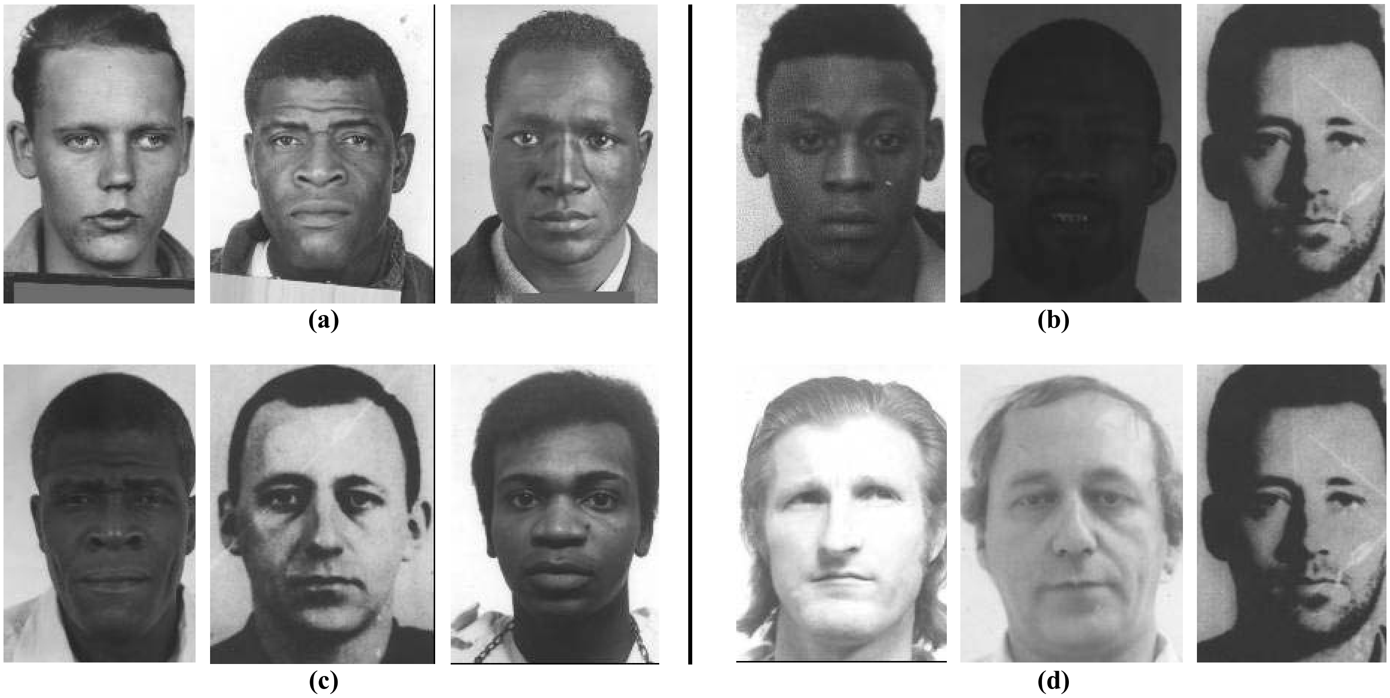

Figure 2. (a) The average highest quality faces selected by humans; (b) the average lowest quality faces selected by humans; (c) the average highest quality faces selected by biometric algorithms; (d) the average lowest quality faces selected by biometric algorithms.

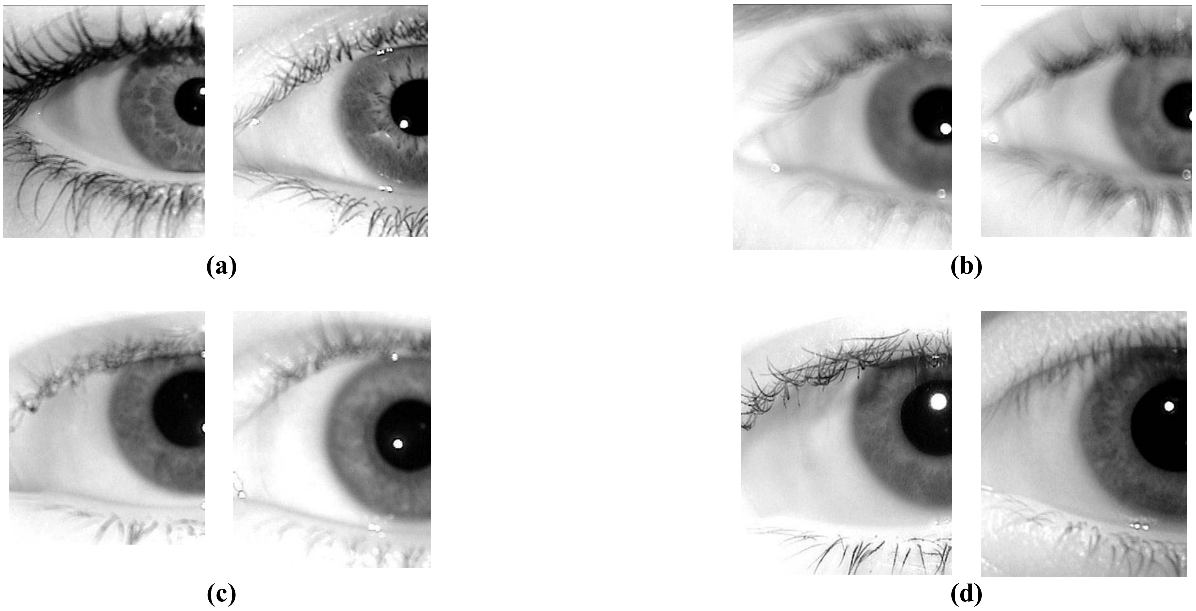

Figure 3. (a) The average highest quality irises selected by humans; (b) the average lowest quality irises selected by humans; (c) the average highest quality irises selected by biometric algorithms; (d) the average lowest quality irises selected by biometric algorithms.