Large Scale Iris Image Quality Evaluation

Elham Tabassi

Information Access Division - Image Group
National Institute of Standards and Technology
100 Bureau Drive MS 89403
Gaithersburg, MD 20899
elham.tabassi@nist.gov

Abstract:
Several recent studies have shown that while iris images captured at near infrared are viable biometrics for verification and identification, similar to other biometrics, its performance drops when comparing images from imperfect sources (e.g. subject blinking), under imperfect conditions (e.g. out of focus) or non-ideal capture device. The immediate question to ask is what factors and to what degree are most influential on iris recognition performance. Motivated by this need, National Institute of Standards and Technology (NIST) initiated Iris Quality Evaluation and Calibration (IQCE). IQCE aims to define and quantify iris image properties that are influential on performance of iris recognition. This paper gives an overview of the IQCE.

1 Introduction

Automatically and consistently determining the quality of a given biometric sample for identification and/or verification is a problem with far-reaching ramifications. If one can identify low-quality biometric samples, the information can be used to improve the acquisition of new data. The same quality measure can be used to selectively improve an archival biometric database by replacing low quality biometric samples with high quality samples. Weights for multimodal biometric fusion can be selected to allow high quality biometric samples to dominate the fusion. All of these applications require that the quality of the biometric sample be determined prior to identification or verification. Most of these applications also require that the quality of the biometric sample be computed in real time during data acquisition.

Current state-of-the-art biometric recognition systems perform at reasonably low error rates. However, the performance degrades substantially as the quality of the input samples drop. Although only a small fraction of input data are of poor quality, the bulk of recognition errors can be attributed to poor quality samples. Poor quality samples decrease the likelihood of a correct verification and/or identification, while extremely poor-quality samples might be impossible to verify and/or identify. If quality can be improved, either by sensor design, by user interface design, or by standards compliance, better performance can be realized. For those aspects of quality that cannot be designed in, an ability to analyze the quality of a live sample is needed. This is useful primarily in initiating the reacquisition from a user, but also for the real-time selection of the best sample, the selective invocation of different processing methods, or fusion. Accordingly, biometric quality measurement algorithms are increasingly deployed in operational systems [GT06]. US-VISIT, PIV, EU VIS, and India's UID each mandate the measurement and reporting of quality scores of captured images. With the increase in deployment of quality algorithms, the need to standardize an interoperable way to store and exchange biometric quality scores and methods for evaluating the effectiveness of quality algorithms, increases.

2 Iris Quality Evaluation and Calibration

Motivated by this need, National Institute of Standards and Technology (NIST) initiated Iris Quality Evaluation and Calibration (IQCE). IQCE aims to define and quantify iris image properties that are influential on performance of iris recognition. Iris is rapidly gaining acceptance and support as a viable biometric. US-VISIT, PIV and India's UID are using iris as their secondary or primary biometric for verification. While there are several academic publications on iris image quality, IQCE is the first public challenge in iris image quality aiming at identifying iris image quality components that are algorithm or camera agnostic.

IQCE is the second activity under IREX. NIST's iris interoperability program, IREX, was initiated to support an expanded marketplace of iris recognition applications in identity management deployments.

IREX I was conducted to give quantitative support to the recently completed ISO/IEC 19794-6 standard which regulates cross-party interchange of iris imagery. The IREX I test engaged ten iris recognition vendors in implementing the standard and testing various proposed formats. The test established safe size limits for storage of iris data on credentials (e.g., PIV), and for transmission on networks.

While IREX I showed that iris images captured at near infrared are viable biometrics for verification and identification, it also confirmed findings in related studies [VDF05, Dau07, NZSC10, ZZB09] that similar to other biometrics, its performance drops when comparing images from imperfect sources (e.g., subject blinking) or under imperfect conditions (e.g., out of focus). IQCE is motivated by a need to quantitatively define iris image quality and seeks to identify image properties that are influential on recognition accuracy.

NIST invited commercial providers, universities, and non-profit research laboratories with capabilities in producing quality score, either overall scalar quality or specific aspects of quality (e.g., sharpness) to participate. Furthermore, organizations who implemented biometric verification softwares using iris images were invited to participate in IQCE. The comparison scores generated by such submissions were used to quantify the goodness of quality scores. Subsequently, NIST received fourteen IQAAs from nine organizations. Additionally, six out of the nine participants, submitted iris verification SDKs (comparators). Comparison scores of these submissions were used to quantify the predictive power of quality components generated by submitted IQAAs.

This paper gives an overview of IQCE evaluation of the scalar quality scores using only one of the three IQCE test data sets. Due to space limitation, results on evaluation of the quality components or other datasets is not included here.

The rest of the paper is organized as follows: Section 3 presents the metrics employed for evaluation of the IQAAS. It is followed by a brief introduction of the imagery used. Section 5 summarizes IQCE evaluation of the scalar quality scores. Relevance and relation to the development of the international iris image quality standard is discussed in section 6.

3 Quantitative evaluation of quality scores

The quality measurement algorithm, whether generating a scalar summary of a biometric sample's quality or a measurement of a specific aspect of quality (quality component), is regarded as a black box that takes an input iris image and outputs a scalar quality and/or a vector of quality components.

Evaluations are done by quantifying the association between quality scores of each quality component and the observed matching results.

IQCE deployed several metrics to assess the predictive power of the quality scores. In this paper, we overview the two most prominent ones, namely, error (false non-match) vs reject curves and ranked order DET . Results are presented in section 5.

3.1 Error vs reject curve

One metric for comparative evaluation of IQAAS is the error versus reject curves that is introduced in [GT07]. The goal is to demonstate how efficiently rejection of low-quality samples results in improved performance. This models the operational case in which quality is maintained by reacquisition after a low-quality sample is detected. Consider that a pair of samples (from the same eye), with qualities $q_i^{(1)}$ and $q_i^{(2)}$, are compared to produce a geuine score, and this is repeated for N such pairs.

We introduce thresholds u and v that define levels of acceptable quality and define the set of low-quality entries as

$$R(u,v) = \left\{ j \;:\; q_j^{(1)} < u, \quad q_j^{(2)} < v \right\} \tag{1}$$

We compute FNMR as the fraction of genuine scores above threshold computed for those samples *not* in this set

$$\text{FNMR}(\tau) = \frac{M_\psi(\tau)}{M_\psi(-\infty)} \tag{2}$$

$$M_\psi(\tau,u,v) = \sum_{s \in \mathcal{G} \cap R^C} H(s-\tau) \tag{3}$$

where R^C is the complement of R.

If the quality values are perfectly correlated with the genuine comparison scores, setting threshold τ to give an overall FNMR of x and then rejecting x percent with the lowest qualities should result in FNMR of zero after recomputing FNMR.

For an effective IQAA FNMR should decrease quickly with the fraction rejected. An almost flat curve suggests that the quality algorithm is not effective in prediction of performance. The IQAA with the largest negative derivative at the low rejection rate is the most effective, hence the best performer.

The most operationally relevant part of the error vs. reject curves is usually on the left side where a small fraction, x, of low-quality rejections would be tolerable from the perspective of forcing a second enrollment attempt. However, for the ICE2006 data sets, the appropriate fraction is probably larger because the camera's own quality measurement apparatus was suppressed.

Error vs reject curve allows for quantifying the generalizability of an IQAA to other comparators than its mated one. It is a common contention that the efficacy of a quality algorithm is necessarily tied to a particular comparator. We observe that this one-comparator case is commonplace and useful in a limited fashion and should therefore be subject to evaluation. However, we also observe that it is possible and perhaps desirable for an IQAA to be capable of generalizing across all (or a class of) matchers, and this too should be evaluated. Generality to multiple comparators can be thought of as an interoperability issue: can IQAA As quality measure be used with comparator Bs matcher? Such a capability will exist to the extent that pathological samples do present problems to both A and Bs matching algorithms.

3.2 Ranked-ordered detection error trade-off (DET) curves

DET characteristic curves are the primary performance metric for offline testing of biometrics recognition algorithms [MDK+97], [505]. Each point on a DET curve exhibits the false match and false non-match rates associated with a certain threshold value. The DET curve spans the whole range of possible threshold values, which is normally the range of the comparison scores. An IQAA is useful if it can at least give an ordered indication of an eventual performance. For example, for L distinct quality levels, there should notionally be L DET characteristics that do not cross.

Using the geometric mean of the two samples ($\sqrt{q_1 * q_2}$) as their pairwise quality, we divide each comparator's comparison scores into three groups based on the pairwise quality of the images being compared. The set of the lowest quality contains comparison scores with pairwise qualities in the lowest 15 percentile. Comparisons with pairwise quality in the middle 70 percent comprise the second or medium quality set. Finally, comparison scores of images whose pairwise quality are in the highest 15 percentile make up the third or best quality set. Three DET characteristic curves, one for each set above, are generated, as shown in figure 1. Each cell in figure 1 shows three DET curves where quality scores of the identified IQAA are used to partition the comparison scores generated by comparator Hz1. To reveal the dependance of FNMR and FMR on quality at a fixed threshold, τ, the DET curves of each cell are connected at false non-match and false match rates that are observed at the "same threshold" values.

The proper behavior is to observe lower FNMR and FMR as quality improves. An IQAA is effective if the DET curves are separated, with the DET curve corresponding to the lowest quality images appearing at the top (i.e. higher FNMR), and the DET curve of highest quality images at the bottom (i.e. lower FNMR). Overlapping DET curves indicate poor IQAA performance. A higher separation among these three curves indicates a more effective IQAA .

The ranking and the separation of the DET curves, as explained above, will reveal the effect of quality on FNMR . Effect of quality on FMR is demonstrated by the lines connecting the DET curves (the brown lines of figure 1). Assuming the correct ranking, a positive slope is expected meaning high quality images produce low FMR. A negative slope means that high quality images produce higher FMR than the low quality images, which is not desired.

Another observation to make is which IQAA is the best predictor of the comparator whose comparison score was used to generated the graph, in case of figure 1, comparators Hz1 and E2a. It is rightly assumed that an IQAA would be the best predictor of its mated comparator, but it is not always the case. As we can see in figure 1, IQAA F1 provide a better separation of the DET curves of the comparator E2a than E2a's mated IQAA.

4 Data

IQCE deployed three different datasets. For the sake of space and time, only results on one of them, ICE2006 is presented in this paper. Specifically, we used a subset of ICE2006 dataset provided to NIST by the MBGC program[ea08]. The subset we used contains 56871 images of right and left irises from 193 subjects collected from a university population over six semesters within 2004 - 2006 time frame. The images were acquired using an LG EOU 2200 iris scanner.

All images are $640 * 480$ pixels with a median iris diameter of 240 pixels. The variation in iris sizes is mostly due to the subjects iris sizes and their distance from camera at the time of capture.

Because of the collection protocol used for ICE2006 dataset, it contains images of various quality, per description in [PBF07]:

> 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 32 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 infrared (IR) light emitting diodes (LED)s used to illuminate the iris.
>
> For a given subject at a given iris acquisition session, two "shots" of three images each are taken for each eye, for a total of 12 images. 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.
>
> A manual quality control step at Notre Dame was performed to remove images in which, for example, the eye was not visible at all due to the subject having turned their head.

The ICE2006 images are useful for comparative analysis of IQAAS. Its range and diversity of image impairments makes it suitable for investigating the causes of failure and viability of algorithms on the core iris feature extraction and matching problem.

The test data set was divided into two sets, namely enrollment and verification. IQCE received ten submissions from six organizations that compare two iris images and generate a dissimilarity score. These comparators were used to compare all images in the enrollment set with all the images in the verification set. As a result, using the ICE2006 images, slightly over 4 million genuine and more than 7 million impostor comparison scores were generated for each of the comparators.

5 Results

DET curves of two comparators for images with low, medium and high SCALAR QUALITY quality scores of nine IQAAS are shown in figure 1.

The first observation is that the performance of either comparator is significantly affected by the quality of the images. High false non-match and false match rates are observed when the quality of the images being compared are low. In other words, the SCALAR QUALITY scores of IQAAS A2a, C4x, E2a, F1 and Hx are reasonable predictors of performance.

The second observation is on the dependance of performance evaluation of IQAAS on the comparators used for the evaluation.

The DET curves of the mid and high quality images cross for IQAAS B3, C4x, G1 and E2a for comparator Hz1 but not for E2a.

All the IQAAS give a better performance ranking for comparator E2a than Hz1, suggesting that comparator E2a is more sensitive to image quality than Hz1. However, the rankings of the IQAAS remain the same. IQAAS A2a, C4x, E2a, F1 and Hx are the best performers for both comparators. IQAA I1 gives the worst performance ranking, for both comparators.

Lastly, note that, except for IQAA D3, the DET curves of the high quality images are flat.

The ranked-ordered DET curves are useful for comparative analysis of IQAAS. It demonstrates the error rates achieved when comparing low quality images, or high quality images. However, a finer quantification of the effect of quality on performance is not possible at least because sufficient quantity of genuine and impostor comparison scores of certain quality levels are needed to generate a DET curve.

IQCE employed error vs reject curves to investigate the dependance of false non-match rate on quality. Specifically, we examined how quickly FNMR is improved when the lowest quality images are rejected, which is ultimately the most operationally relevant use of quality scores.

Figure 2 shows the error vs reject curves of all fourteen IQCE IQAAS for six different comparators. IQAAS are identified by different line type and color. The threshold is set to give initial FNMR = 0.1. The gray dotted line shows the ideal case where the rejection of the comparisons with the lowest ten percent quality results in zero FNMR.

IQAAs E2a, F1, and Hx are generally the best performer for their mated comparators as well as other comparators (i.e. they are generalizable). E2a slightly outperforms the other two, and performs close to the ideal case (gray dotted line) for its mated comparator. IQAAS A2a and B3 are the best predictor of performance for their mated comparators, but not for other comparators (i.e. they are not generalizable).

The least effective IQAAS are I1, and G2.

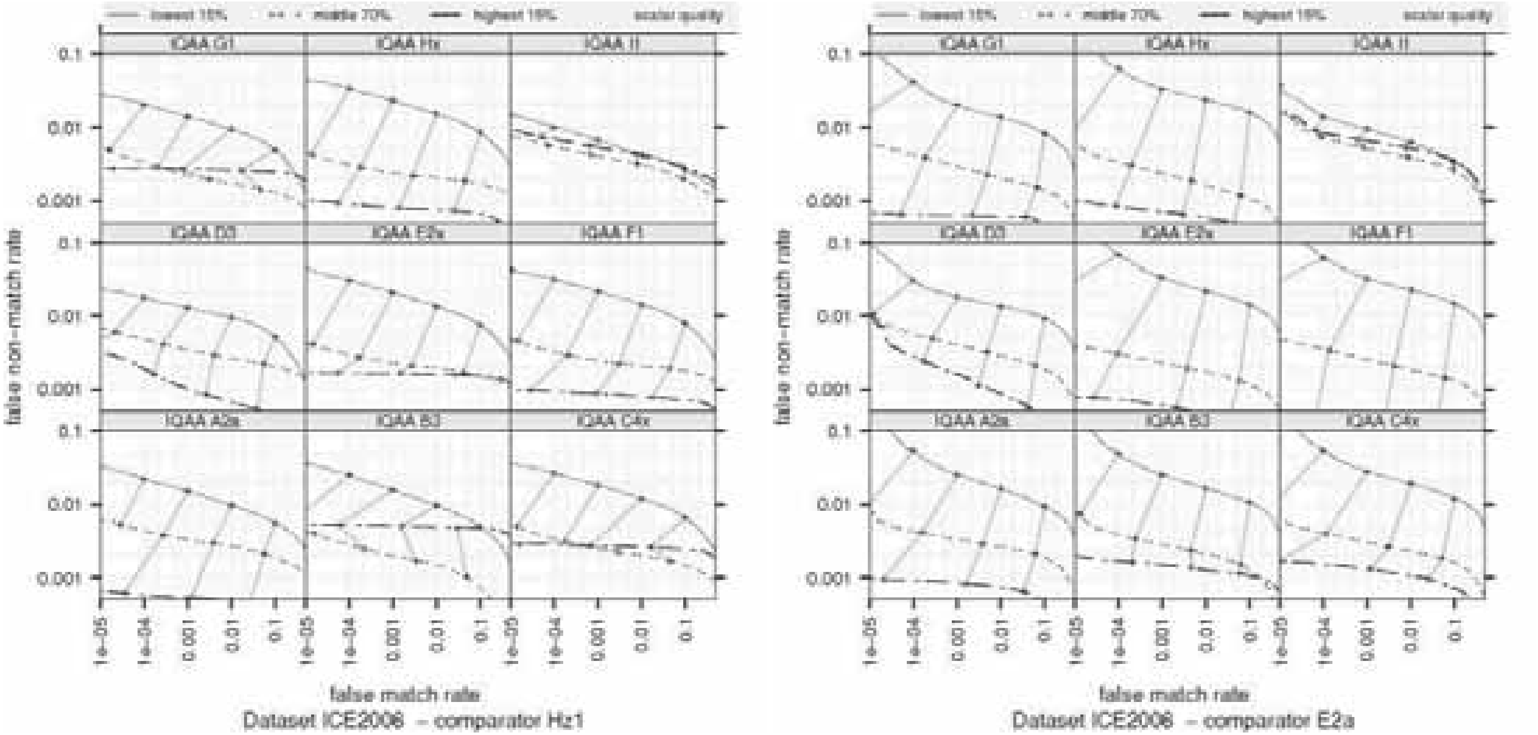

Figure 1: Ranked DET curves for comparator Hz1 on data set ICE2006 The set of all comparisons are partitioned into three groups based on the pair-wise SCALAR QUALITY scores of the images being compared. The lowest quality set contains comparisons with pairwise quality in the lowest 15 percentile. The highest quality set contains comparisons with pairwise quality in the highest 15 percentile. The rest of the comparisons, namely the middle 70%, make up the third set. The DETs are connected at the same score threshold values (brown lines). Lower FNMR and FMR rates are expected for higher quality images. That means well separated curves in each cell, with the DET curve corresponding to the lowest quality appearing above, and the DET curve of highest quality below all the other curves. In addition the lines connecting the DET curves must have a positive slope.

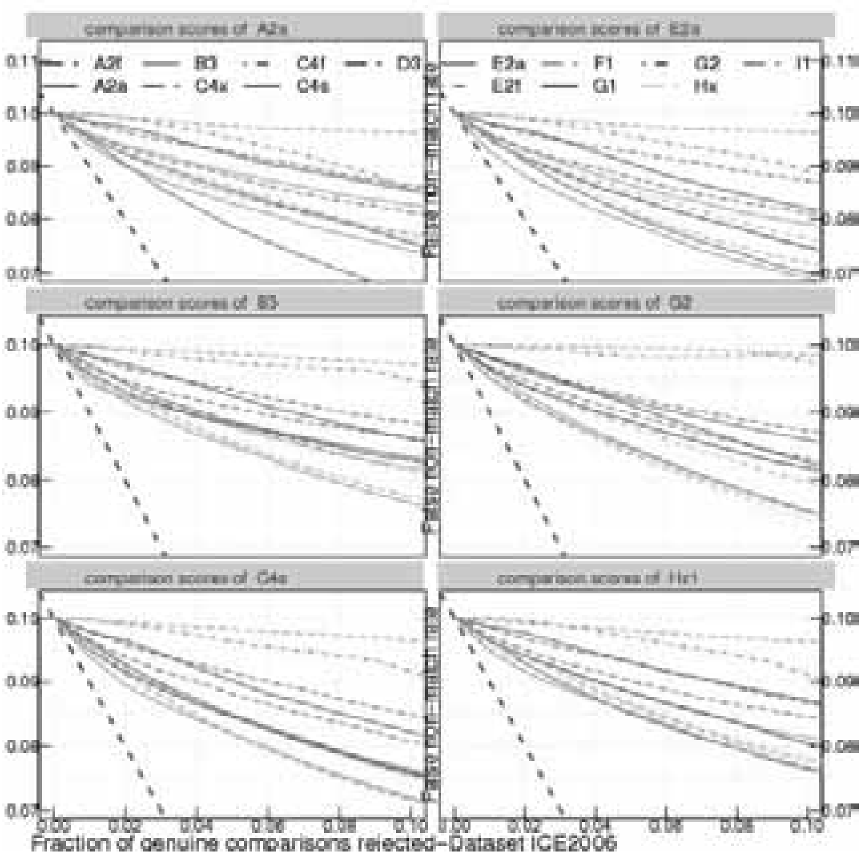

Figure 2: FNMR vs reject curves for SCALAR QUALITY scores on datasets and ICE2006 . The threshold is set to give an initial FNMR = 0.1. The gray dotted line shows the ideal case where the rejection of the comparisons with the lowest ten percent quality results in zero FNMR. IQAA E2a is the best performer, followed by Hx and F1, followed by C4x. IQAAs B3 and A2a perform better for their mated comparators than other comparators. The lowest performance is observed for IQAAs I1 and G2.

6 Support to standard development process

The IQCE activity supports a new, formal, standard addressing iris quality. The standard, ISO/IEC 29794-6 Iris Image Quality [311], was initiated by the Working Group 3 of the ISO SC 37 committee in July 2009. The standard will define a vector of quality components each of which is some quantitative measure of a subject-specific or image-specific covariate. The current working draft (SC 37 N 4302) defines 19 image acquisition or subject covariates and 17 metrics for assessing the utility of an iris image. Furthermore, ISO/IEC 29794-6 aims to establish precise statements of how to compute each of the quality metrics.

A summary of quality components identified in ISO/IEC 29794-6 which IQCE examined, follows.

- **scalar quality** An overall quality score that is an estimate of the matchability of the image. Per ISO/IEC 29794-1 [310] Biometric sample quality – Part 1: Framework, the scalar quality shall an integer between 0 and 100, where 0 indicates the lowest quality and 100 the best quality. Lower recognition error is expected for images with low scalar quality scores.
- **gray level spread** shall measure the overall iris image for evidence of a spread of intensity values in iris data. An "underexposed" image would have too few high intensity pixels, and conversely for "overexposed". An image with a high score (good quality) indicates a properly exposed image, with a wide, well distributed spread of intensity values.
- **iris size** shall be a measure in the image plane, representing half the distance across the iris along the horizontal.
- **pupil_iris ratio** shall represent the degree to which the pupil is dilated or constricted. It is a dimensionless term, being the ratio or pupil radius to iris radius.
- **usable iris** is the percent of the iris portion of the image that is not occluded by eyelids, eyelashes, or saturating specular reflections, expressed as percentage of area of an annulus modelling the iris without such occlusions.
- **iris-sclera contrast** shall represent the image characteristics at the boundary between the iris region and the sclera. Sufficient contrast is needed in many implementations of iris segmentation algorithms. Low or insufficient contrast may result in a failure to process an iris image during feature extraction.
- **iris-pupil contrast** shall represent the image characteristics at the boundary between the iris region and the pupil. Sufficient iris-pupil contrast is needed in many implementations of iris segmentation algorithms. Low or insufficient contrast may result in a failure to process an iris image during feature extraction.
- **iris shape** should be mathematical expression of the iris-sclera boundary. Note that the effect of this component on performance depends on the sensitivity of the segmentation algorithm to the deviation from circularity in iris-sclera and iris-pupil boundary.
- **pupil shape** should be mathematical expression of iris-pupil boundary. Deviation from circularity in the iris-pupil boundary can affect segmentation accuracy. The effect of this metric on performance depends on the sensitivity of the segmentation algorithm to the deviation from circularity in iris-pupil boundaries.
- **margin** shall quantify the degree to which the image achieves positioning of the iris portion of this image relative to the edges of the entire image. The maximum quality value for this metric shall be achieved when the margin requirements of ISO/IEC 19794-6:2011 are satisfied.
- **sharpness** shall measure the degree of defocus present in the image.
- **motion blur** shall measure the degree of distortion in the image due to motion.

- **signal to noise ratio**
- **gaze angle** shall be an estimate of the direction of displacement between the optical axis of the eye and the optical axis of the camera. This measure is inclusive of both head angular orientation and eye-gaze angle relative to the head.

IQCE examined the effectiveness of these quality components in prediction of performance with the goal to produce a refined list of image quality components that significantly affect the iris recognition performance. To ensure that the above list includes all the possible image impairments that could affect performance, IQCE encouraged submission of, and consequently evaluated, other quality components not included above, or any other proprietary component.

The result of the evaluation is not included here due to space limitation, except for comparative analysis of quality components of one of the IQAAS, which follows.

6.1 Predictive power of quality components

Before proceeding to quantifying the predictive power of the quality components of the IQAAS, it is important to emphasize the dependance of this analysis on the imagery used for the evaluation. The effect of a quality component will not be observed if the test data does not represent varying degree of the impairment. For example, ICE2006 data set lacks severely dilated or constricted (probably because lighting condition did change during the capture sessions) or compared to other IQCE data sets, it lacks a wide range of sharpness, as a result sharpness is not a significant factor for ICE2006 images, but its effect on performance is comparable to usable iris area for the IQCE data set that contains images with varying level of sharpness.

Figure 3 shows a variant of the error vs reject curve discussed in 3 that compares the predictive power of different quality components generated by IQAA C4x, and using comparison scores of its mated comparator. The threshold is set to give initial FNMR = 0.1. The gray dotted line shows the ideal case where the rejection of the comparisons with the lowest ten percent quality results in zero FNMR. The most effective IQAA is the one with the biggest negative derivative at the low rejection rate.

For ICE2006 images, with its existing range of defects, the most effective quality components are the scalar quality and the proprietary metric gravitas, followed closely by usable iris area and the proprietary metric auctoritas. Its second most effective tier of quality components consists of proprietary component dignitas, iris-pupil contrasts, proprietary component pietas and iris-sclera contrast.

As mentioned, the ranking of the quality components slightly changes for other data set. Scalar quality, usable iris, iris-pupil contrasts, and iris-sclera contrast, sharpness and gaze angle seem to have the most influence on performance.

7 Summary

An overview of NIST IQCE was presented along with results on the performance evaluation of scalar quality scores. We discuss two metrics for comparative analysis of image quality assessment algorithms.

Predictive power of scalar quality scores generated by different IQAAS vary. For the best performers, the difference between FNMR and FMR of the images in the lowest fifteen percentile and those in the upper fifteen percentile can be as high as an order of magnitude.

References

[310] Working Group 3. *ISO/IEC 29794-1 Information Technology - Biometric Sample Quality - Part 1: Framework*. JTC1 :: SC37, is edition, 2010. http://isotc.iso.org/isotcportal.

[311] Working Group 3. *ISO/IEC 29794-6 Information Technology - Biometric Sample Quality - Part 6: Iris image*. JTC1 :: SC37, working draft 4 edition, 2011. http://isotc.iso.org/isotcportal.

[505] Working Group 5. *ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework*. JTC1 :: SC37, international standard edition, August 2005. http://isotc.iso.org/isotcportal.

[Dau07] John Daugman. New Methods in Iris Recognition. *IEEE Transactions on Systems, Man, and Cybernetics Part B:Cybernetics*, 37(5):1167–1175, 2007.

[ea08] P. J. Phillips et al. Overview of the Multiple Biometrics Grand Challenge. Technical report, National Institute of Standards and Technology, www.nd.edu/ kwb/PhillipsEtAlICB_2009.pdf [on June 24, 2009], 2008.

[GT06] P. Grother and E. Tabassi. In *Proceedings of the NIST Biometric Quality Workshop*, March 2006. http://www.itl.nist.gov/iad/894.03/quality/workshop/presentations.htm.

[GT07] P.J. Grother and E. Tabassi. Performance of Biometric Quality Measures. *IEEE Trans. Pattern Anal. Mach. Intelligence (PAMI)*, 29(4):531–543, April 2007.

[MDK$^+$97] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. *In Proceedings of Eurospeech 97*, pages 1895–1898, 1997.

[NZSC10] N.Kalka, J. Zuo, N. Schmid, and B. Cukic. Estimating and fusing quality factors for iris biometric images. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, 40(3):509–524, 2010.

[PBF07] P. Jonathon Phillips, Kevin W. Bowyer, and Patrick J. Flynn. Comments on the CASIA version 1.0 Iris Data Set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1869–1870, 2007.

[VDF05] N. Schmid V. Dorairaj and G. Fahmy. Performance evaluation of non? ideal iris based recognition system implementing global ICA encoding. In *IEEE International Conference on Image Processing ICIP-05*, pages 285–288, Genoa, Italy, September 2005.

[ZZB09] Y. Du Z. Zhou and C. Belcher. Transforming traditional iris recognition systems to work in nonideal situations. In *IEEE Transaction on Industrial Electronics*, volume 56, August 2009.

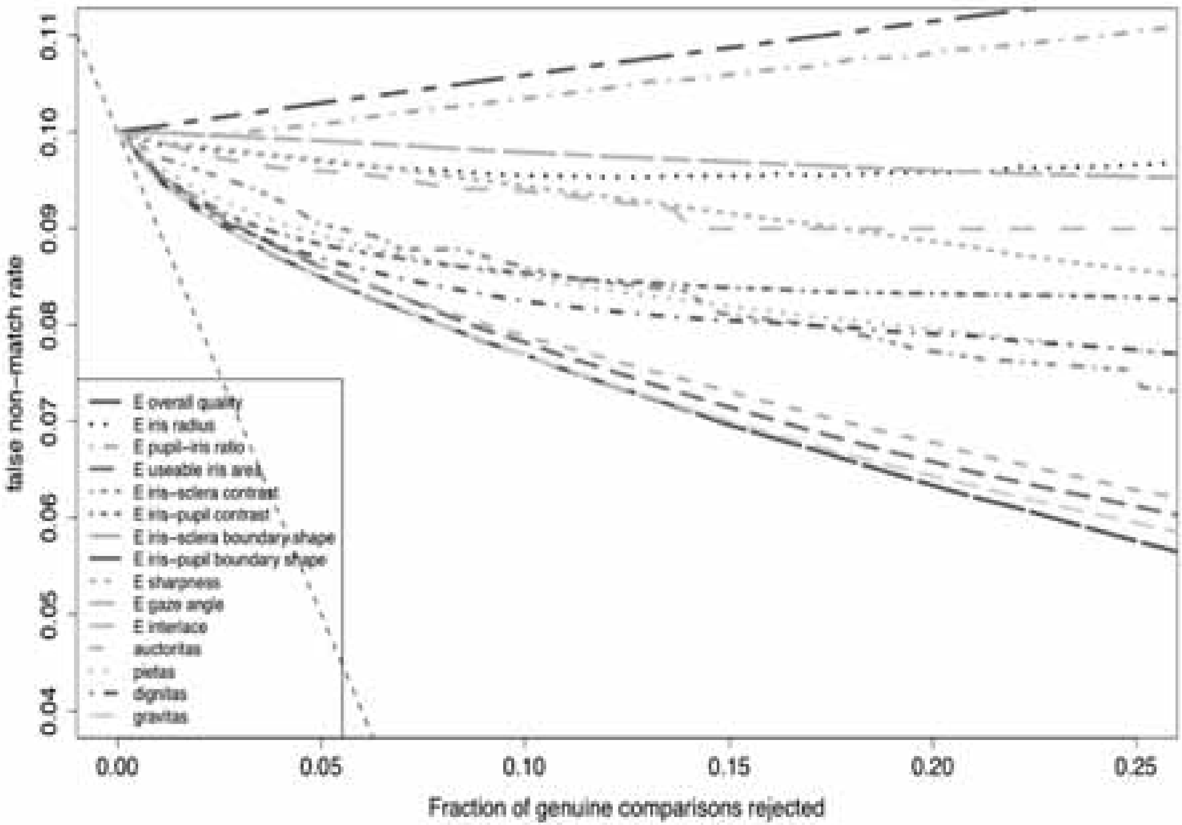

Figure 3: FNMR vs reject curves for quality component scores of IQAA C4x using its mated comparison scores on ICE2006 images. The threshold is set to give an initial FNMR = 0.1. The gray dotted line shows the ideal case where the rejection of the comparisons with the lowest ten percent quality results in zero FNMR. For the ICE2006 images, with its existing range of defects, the most effective quality components are the scalar quality and the proprietary metric gravitas, followed closely by usable iris area and the proprietary metric auctoritas.