

RESEARCH

Open Access



Utility-based performance evaluation of biometric sample quality measures

Olaf Henniger^{1,2*} , Biying Fu¹ and Alexander Kurz^{1,2}

*Correspondence:
olaf.henniger@igd.fraunhofer.de

¹ Fraunhofer Institute
for Computer Graphics Research
IGD, Darmstadt, Germany

² Department of Computer
Science, Technical University
of Darmstadt, Darmstadt,
Germany

Abstract

The quality score of a biometric sample is intended to predict the sample's degree of utility for biometric recognition. Different authors proposed different definitions for utility. A harmonized definition of utility would be useful to facilitate the comparison of biometric sample quality assessment algorithms. In this article, we compare different definitions of utility and apply them to both face image and fingerprint image data sets containing multiple samples per biometric instance and covering a wide range of potential quality issues. The results differ only slightly. We show that discarding samples with low utility scores results in rapidly declining false non-match rates. The obtained utility scores can be used as target labels for training biometric sample quality assessment algorithms and as baseline when summarizing utility-prediction performance in a single plot or even in a single figure of merit.

Keywords: Biometrics, Quality estimation, Performance evaluation

1 Introduction

The quality score of a biometric sample is expected to express the utility of this sample for automated recognition of the individual [1]. It can be used, e.g., for deciding about re-acquisition of a biometric sample, for weighting partial results in multi-biometric systems, or for selecting the best from a series of samples. The utility of a biometric sample depends on the faithfulness of the sample to its source (i.e. fidelity) and the distinctiveness of the biometric features (referred to as character [1]). It also depends on the biometric references the sample is compared with: for instance, a face image quality score can reflect utility when comparing with full-frontal reference face images compliant to the requirements on passport photographs [2] or with reference images captured in the wild. A biometric sample's utility for automated recognition also depends on the number of years passed since capturing the mated reference (ageing). However, since the time elapsed cannot be determined from one sample alone, it does not contribute to the calculation of quality scores.

Based on NIST's ground-breaking work on fingerprint image quality assessment [3], the international standard setting the framework for objective biometric sample quality assessment [1] defines utility as degree to which a biometric sample supports biometric recognition performance but leaves open how to quantify the utility of a particular

sample. As all biometric sample quality assessment algorithms try to predict utility, a harmonized approach for calculating target utility labels would be useful to enable the comparison of results. The contributions of this article are:

1. We compare two controversial approaches to measuring and calibrating the utility of biometric samples with and without consideration of non-mated comparison scores (see Sect. 2).
2. Using open-source face image and fingerprint image quality assessment algorithms, we demonstrate how to summarize the utility-prediction performance of biometric sample quality assessment algorithms in a single informative plot and a single figure of merit based on the chosen utility definition (see Sect. 3).
3. Using publicly available face image and fingerprint image data sets and open-source comparison software, we show that the proposed definitions of utility are remarkably effective in finding the samples responsible for false non-matches (see Sects. 4 and 5).

This article focuses on assessing the utility of biometric samples a posteriori, i.e. after comparing with other samples in a given data set, and on simplifying the performance evaluation of biometric sample quality assessment algorithms. The challenge how to predict the utility of a sample a priori, i.e. without comparing it with other samples in a data set, is out of scope of this article. This article extends related conference papers [4, 5] by studying the contribution of non-mated comparisons and by contrasting face images (with ICAO-compliant canonical reference images [2]) with fingerprint images (without canonical references). The approach can be applied to other types of biometric characteristics as well.

2 Utility assessment methods

2.1 Related work

In contrast to general image quality assessment algorithms, which try to predict the image quality that an average human observer would perceive, starting with version 1 of the NIST Fingerprint Image Quality (NFIQ) assessment software [3], biometric sample quality assessment algorithms try to predict the utility of the given sample for computer-based recognition. The concept of utility is independent of the considered type of biometric characteristics and has been studied for diverse types of characteristics.

Regarding fingerprint images, Tabassi et al. defined the utility of a biometric sample as the normalized difference between the sample's mated similarity score and the mean of the sample's non-mated similarity scores in a data set that contains two samples per finger instance [3]. More than two samples per instance were not available for a large enough number of finger instances. Predicting a real-valued scalar, like utility, is a regression problem. However, as regression methods failed to give adequate predictions, for NFIQ 1, the machine-learning problem was restated in terms of classification into five levels of utility: excellent, very good, good, fair and poor [3]. The boundaries between the levels of utility were defined based on comparison score quantiles, i.e. the worst samples in the data set were labelled poor, etc. The improved version 2 of NFIQ [6, 7] provides more granular quality scores in the range from 0 to 100 (lowest to highest quality). For NFIQ 2, Tabassi

et al. trained a random decision forest for binary classification into two utility classes (high and low utility), which outputs the probability that an image belongs to the high-utility class [6]. Images whose mated similarity score was in the top 10% were selected as high-utility training images from a data set that again contained just two samples per instance. Images whose mated similarity score was smaller than a threshold value corresponding to a false match rate of 0.01% were selected as low-utility training images.

Regarding iris images, the mated similarity scores in a data set with two samples per instance were used as target utility labels [8].

Regarding face images, target utility labels are also required for training supervised machine learning-based quality assessment models. Related approaches can be clustered based on the labelling methods:

- The training face images can be labelled with human assessments of perceived quality [9].
- Training face images can be labelled with utility scores automatically derived from comparison scores [10, 11]. For instance, FaceQnet [10] quantifies utility as similarity score between the face image and a high-quality mated reference image, ignoring non-mated similarity scores, whereas SDD-FIQA [11] quantifies utility as the Wasserstein distance between mated and non-mated comparison score sets.
- Face images of high and low quality can be selected for training [12–14] like what is done for NFIQ.

Regarding handwritten signatures/signs, the normalized difference between the arithmetic means of mated similarity scores and of similarity scores for skilled forgeries was used to classify samples into four levels of utility [15].

2.2 Measuring utility

The utility of a biometric sample can be assessed by comparing this sample with other samples in a biometric data set containing multiple encounters of the included biometric instances and covering a wide range of potential quality issues [3]. When comparing with other samples, the biometric references can be selected in different ways depending on the intended use case: canonical samples fulfilling certain requirements (e.g., face images compliant to the requirements on passport photographs [2]) can be selected as references, or any available samples can be used as references (e.g., if no canonical representation is defined).

No matter how the references are selected, we measure the utility of a biometric sample i for a comparison algorithm that outputs similarity scores (the more similar, the higher) first as normalized difference between the mean of i 's mated similarity scores and the mean of i 's non-mated similarity scores:

$$u_i = \frac{\mu_{mi} - \mu_{ni}}{\sqrt{\sigma_m^2 + \sigma_n^2}}, \quad (1)$$

where μ_{mi} is the arithmetic mean of the similarity scores for i and mated references; μ_{ni} is the arithmetic mean of the similarity scores for i and non-mated references; σ_m is the standard probe deviation of all mated similarity scores; σ_n is the standard probe

deviation of all non-mated similarity scores. If only one mated comparison score is available per sample, Eq. 1 specializes to the normalized difference between the sample's mated score and the mean of its non-mated scores as in [3]. In the unfavourable case that only one mated similarity score is available per sample, and all compared samples are of unknown quality, a low or high comparison score cannot be attributed to either of the compared samples.

Note that the normalized difference between the mean of all mated similarity scores μ_m and the mean of all non-mated similarity scores μ_n was established in [17, 18] as d' (d-prime), a measure of the separability of the mated and non-mated comparison score distributions:

$$d' = \frac{\mu_m - \mu_n}{\sqrt{\sigma_m^2 + \sigma_n^2}}. \quad (2)$$

d' squared is also known as Fisher's discriminant ratio [19].

If the sample-specific non-mated comparison score distributions are the same for all samples, utility depends only on mated comparison scores as hypothesized in [16]. To examine the contribution of non-mated comparisons in Sects. 4 and 5, we introduce a slimmed-down utility formula:

$$\hat{u}_i = \frac{\mu_{mi}}{\sigma_m}, \quad (3)$$

by removing all terms related to non-mated comparisons from Eq. 1.

For comparison algorithms that output dissimilarity scores (the more similar, the lower) rather than similarity scores, we use $-u_i$ and $-\hat{u}_i$, respectively, so that the values increase with utility as required by [1].

Using only canonical samples known to be of high utility as references ensures that low similarity scores are clearly attributable to the biometric probe [16]. If other than high-quality canonical samples are used as references, comparing with as many as possible other samples ensures that u_i is dominated by the influence of the sample to be assessed and not by individual other samples. To avoid potential bias, there should be the same number of samples for each biometric instance.

2.3 Utility score calibration

Score calibration enables comparability of scores across quality assessment algorithms and across data sets. Higher quality score values are intended to imply higher utility [1]. The u_i and \hat{u}_i values meet this expectation. A biometric sample is the more useful for biometric recognition the more similar it is to its mated samples and the larger the distance between the sample-specific distributions of mated and non-mated comparison scores is. Furthermore, the standard requires quality score values to be in the range from 0 to 100 [1]. The u_i values do not meet this requirement. Hence, the u_i values must be mapped to the range from 0 to 100 to serve as target quality score. To fit into the standardized biometric data interchange formats [20, 21], quality scores also need to be quantized to integers, which can be achieved by rounding.

A sigmoid function having an S-shaped curve can be used to map u_i (Eq. 1) and \hat{u}_i (Eq. 3) to the range from 0 to 100:

$$S(u_i) = \frac{100}{1 + e^{\frac{u_0 - u_i}{w}}}, \quad (4)$$

where u_0 indicates the inflection point and w the slope of the sigmoid. The non-linear mapping allows score calibration. For calibrating u_i , we chose the parameters u_0 and w in Eq. 4 such that $S(0) = 25$ and $S(u_A) = 50$, where

$$u_A = \min(\{u_i | i \in A\}) \quad (5)$$

is the lowest of the u_i values of the samples in the set A of unobjectionable samples. A sample belongs into the set of unobjectionable samples if and only if all its mated similarity scores are greater than any sample's non-mated similarity scores (or all its mated dissimilarity scores are less than any sample's non-mated dissimilarity scores). This results in $u_0 = u_A$ and $w = \frac{u_A}{\ln 3}$. Hence, i 's utility score u_i^* in the range from 0 to 100 is:

$$u_i^* = \frac{100}{1 + 3^{1 - \frac{u_i}{u_A}}}. \quad (6)$$

If the subset A of unobjectionable samples is empty in the biometric data set, u_A can be approximated by $\max(\{u_i | i \notin A\})$ as in [4]. In this case, no sample in the data set reaches a utility score higher than 50.

Contrary to bin boundaries based on score quantiles as proposed in [3, 22], the resulting bin boundaries are robust against data set shift. $S(0) = 25$ means that clearly deficient samples with $u_i < 0$ (which appear more similar to non-mated samples than to mated samples) are assigned utility scores less than 25. $S(u_A) = 50$ means that samples whose utility u_i is at least as high as that of any sample in the set A of unobjectionable samples are assigned utility scores greater than or equal to 50. This is in line with [23], which recommends that quality scores from 0 to 25 indicate deficient (for many use cases unacceptable) quality, from 26 to 50 marginal, from 51 to 75 adequate, and from 76 to 100 excellent quality.

For calibrating \hat{u}_i , the parameters should be set regarding \hat{u}_i instead of u_i . We chose u_0 and w such that $S(\hat{u}_B) = 25$ and $S(\hat{u}_A) = 50$, where $\hat{u}_B = \frac{\mu_n}{\sigma_n}$, μ_n is the arithmetic mean of all non-mated similarity scores, $\hat{u}_A = \frac{\mu_n + 3\sigma_n}{\sigma_n}$ (if the non-mated comparison scores have a normal/Gaussian distribution, then almost all of them except for outliers are less than $\mu_n + 3\sigma_n$), and σ_n is the standard probe deviation of all non-mated similarity scores. This results in $u_0 = \hat{u}_A$ and $w = \frac{\hat{u}_A - \hat{u}_B}{\ln 3}$. After insertion and simplification, i 's slimmed-down utility score \hat{u}_i^* is:

$$\hat{u}_i^* = \frac{100}{1 + 3^{1 - \frac{\mu_{mi} - \mu_n}{3\sigma_n}}}. \quad (7)$$

3 Utility-prediction performance testing methods

3.1 Related work

Utility-prediction models should neither underestimate nor overestimate the utility of individual biometric samples to avoid mistakenly discarding good samples or retaining bad samples. Their performance can be measured in several ways: whereas the widely

used error vs. discard characteristics (Sect. 3.2) and the d' vs. discard characteristics (Sect. 3.3) are summary statistics showing how efficiently discarding samples with a low score results in improved biometric recognition performance, prediction errors (Sect. 3.4) aggregate the sample-specific distances between predicted and observed utility.

Grother and Tabassi introduced error vs. discard characteristics (Sect. 3.2) as error vs. reject characteristics with respect to false non-match errors [22]. ISO/IEC 29794-1 [1] standardized and extended them to false match errors. The change of name is to avoid confusion with the reject/accept outcome of biometric verification systems. Details of this method for utility-prediction performance testing are studied in [24]. Some related work recalculates the decision threshold at each sampled discard ratio to fix the FMR rather than the decision threshold when calculating error vs. discard characteristics [11, 25–27]. This approach renders false match error vs. discard characteristics superfluous (same FMR value for any supported discard ratio) but increases the complexity of calculating error vs. discard characteristics and prevents the results from being comparable with the standard approach. To support the comparability of results, ISO/IEC 29794-1 [1] includes a Python script for calculating error vs. discard characteristics, which is available from the ISO Standards Maintenance Portal.¹

d' vs. discard characteristics were introduced in [4] to simplify benchmarking by summarizing the utility-prediction performance in a single plot.

Quality-assessment error rates (named incorrect sample rejection rate and incorrect sample acceptance rate) were introduced in [16] and are included in [1]; however, taking only a single mated comparison score per sample into account. These metrics were generalized in [5] for the case that multiple comparison scores are available per sample (see Sect. 3.5).

The use of statistical prediction errors like the root mean square error is widespread practice, e.g., in telephone-transmission quality assessment [28] and in general no-reference image quality assessment [29].

3.2 Error vs. discard characteristics

An error vs. discard characteristic (EDC) shows the dependence of an error rate (such as the false non-match rate FNMR) at a fixed decision threshold on the percentage of reference and probe samples discarded based on lowest quality scores (discard ratio) [22]. The steeper the FNMR decreases with increasing discard ratio and without significantly increasing the false match rate FMR, the better. Values at high discard ratios can be omitted for lack of statistical reliability in case of few remaining error cases (remember the Rule of 30 [30]: to be 90% confident that the true error rate is within $\pm 30\%$ of the observed error rate, at least 30 errors must be observed). The percentage of samples that the operator of a biometric system is willing to discard, if any, depends on the use case.

As not only false non-match but also false match errors happen, a false non-match EDC should be considered together with the corresponding false match EDC. A decrease in FNMR may inadvertently come along with an increase in FMR.

¹ <https://standards.iso.org/iso-iec/29794/-1/ed-3/en/>.

The EDCs depend on the decision threshold value. Therefore, it is common to plot EDCs for several decision thresholds. The drawback of EDCs is that they do not provide a single plot: to compare the utility-prediction performances of biometric sample quality assessment algorithms, you should look at both false non-match and false match EDCs and that for various decision thresholds.

3.3 d' vs. discard characteristics

A d' vs. discard characteristic shows the dependence of d' (Eq. 2) on the percentage of reference and probe samples discarded based on lowest quality scores (discard ratio) [4]. The better the utility prediction works, the steeper d' (the distance between the mated and non-mated comparison score distributions) increases with increasing discard ratio.

The d' vs. discard characteristics depend on the underlying biometric data set, comparison algorithm and quality assessment algorithm. However, if the same data set and the same comparison algorithm are used, the utility-prediction performance of different quality assessment algorithms can be compared simply based on their d' vs. discard characteristics.

3.4 Prediction errors

The prediction performance of regression models can be measured taking the sample-specific differences between predicted and observed values into account.

A common method for characterizing a prediction model's performance is to use the root mean square error (RMSE) between observed and predicted values. For quality scores in the same standardized range from 0 to 100 as utility scores and slimmed-down utility scores, RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (q_i - u_i^*)^2}, \quad (8)$$

where u_i^* are observed utility scores or observed slimmed-down utility scores, q_i are the corresponding quality scores considered as the predicted values, and N is the number of samples in the testing data set. The lower the RMSE value, the better the quality scores predict utility scores.

Provided they deliver quality scores in the standardized value range, the utility-prediction performance of different biometric sample quality assessment algorithms can be compared based on their RMSE values calculated in the same data set.

3.5 Quality-assessment error rates

If the quality scores are used to make decisions on whether to discard or to retain biometric samples for further processing, then deviations between predicted and observed utility do not matter as long as mistakenly discarding useful samples or retaining less useful ones is avoided. To express this, the following metrics were defined in [5]:

- Quality-assessment false negative rate (QFNR): proportion of biometric samples with a quality score below a quality-score threshold u but a utility score above 50.

- Quality-assessment false positive rate (QFPR): proportion of biometric samples with a quality score above a quality-score threshold u but a utility score below 50.

The utility score of 50 separates objectionable (marginal or deficient) samples from unobjectionable (adequate or excellent) samples.

QFNR and QFPR both depend on the chosen data set, comparison system and the quality-score threshold u and should be quoted together at the same quality-score threshold. The higher the quality-score threshold, the more samples are discarded, including samples of unobjectionable quality that had better be retained. The lower the quality-score threshold, the more samples are retained, including samples of objectionable quality that had better be discarded.

4 Face image quality assessment experiments

4.1 Experimental setup

4.1.1 Face image data set

We used a subset of the publicly available data set Multi-PIE [31] because this data set contains multiple face images per subject taken under varying conditions (pose, illumination, facial expression), which raises comprehensible quality issues. Face images were taken in up to four sessions over the span of five months from 15 viewpoints (using 13 cameras located at head height and spaced in 15° intervals plus two cameras located above head height, simulating surveillance camera views), under 19 illumination conditions (using 18 flashlights) and displaying several facial expressions (including neutral, smile, surprise, squint, disgust and scream).

As probe images, we used a random subset containing 21,343 face images of 250 subjects. All probe images are of size 640×480 pixels. As high-quality references, we used 250 well-lit frontal face images with neutral facial expression, one per subject. We used only frontal face images as references because many face recognition systems store such images as references (e.g., ePassports, forensic databases, entry/exit systems). All reference images are of size 2048×3072 pixels.

4.1.2 Face comparison

Before comparison, faces must be detected and aligned, and features must be extracted from the aligned face images. For face detection, we used multitask cascaded convolutional networks (MTCNN) [32]. For 16 profile face images, face detection failed. As no faces were detected, feature extraction and face image quality assignment failed for these images.

We performed face alignment by applying an affine transformation based on the five facial landmarks detected by MTCNN (centres of the left and of the right eye, tip of the nose, left and right corners of the mouth). This ensures that the relative positions, angles, and distances between the landmarks are preserved while adjusting for scale, rotation, and translation variations in the images. The feature extraction algorithms used require the aligned input image size to be 112×112 pixels.

For feature extraction, we used two open-source deep-learning-based face feature extraction algorithms implemented in Python:

- (a) ArcFace [33]: This algorithm aims at obtaining highly discriminative face embeddings for automatic face recognition by incorporating margins in the loss function. A collection of face image data sets was used as training data, including the CASIA [34], VGGFace2 [35], MS-Celeb-1 M [36] and DeepGlintFace [37] data sets.
- (b) MagFace [26]: This algorithm leverages the loss function to learn a universal face feature embedding that can be used for both assessing face image quality and face recognition. MagFace has a backbone based on ResNet100 and was trained on MS1MV2 [33], a selected version of MS-Celeb-1 M [36].

For comparing two face feature vectors, we calculated their cosine similarity as similarity score in the range from -1 to 1 (the more similar, the higher). Cosine similarity measures the similarity of the orientation of two feature vectors regardless of their magnitude. ArcFace and MagFace are designed to give best results with cosine similarity.

Each probe face image was compared with each mated and non-mated reference image within the face image data set (see Sect. 4.1.1). Altogether, using each algorithm, 21 327 mated and 5 310 423 non-mated comparisons were performed.

4.1.3 Face image quality assessment

As this article does not aim at a comprehensive comparative performance evaluation of the variety of face image quality assessment algorithms described in literature [38], we chose only some examples to demonstrate utility-prediction performance testing in Sect. 4.2. We use the following open-source deep-learning-based face image quality assessment algorithms implemented in Python (in alphabetical order):

- (a) CR-FIQA [27]: The loss function is extended by a certainty ratio (CR) term that, during training, takes the true class centre and the nearest negative class centre into account to compute a quality score. An additional regression layer is trained to predict quality also for previously unseen samples.
- (b) MagFace [26]: This algorithm extracts feature vectors for face recognition, the magnitudes of which correspond to the utility of the face image. Feature vectors of high-quality images have larger magnitudes than feature vectors of low-quality images.
- (c) SDD-FIQA (Similarity Distribution Distance Face Image Quality Assessment) [11]: The target utility label of each training image was the Wasserstein distance between mated and non-mated comparison score sets.
- (d) SER-FIQ (Stochastic Embedding Robustness Face Image Quality) [25]: This algorithm assesses the quality of a face image based on the robustness of the feature vector of a specific deep-learning-based face comparator against random dropout patterns. We used SER-FIQ fine-tuned for ArcFace [33].

While SDD-FIQA adopts a supervised machine learning-based approach using target labels calculated from comparison scores during training, CR-FIQA, MagFace

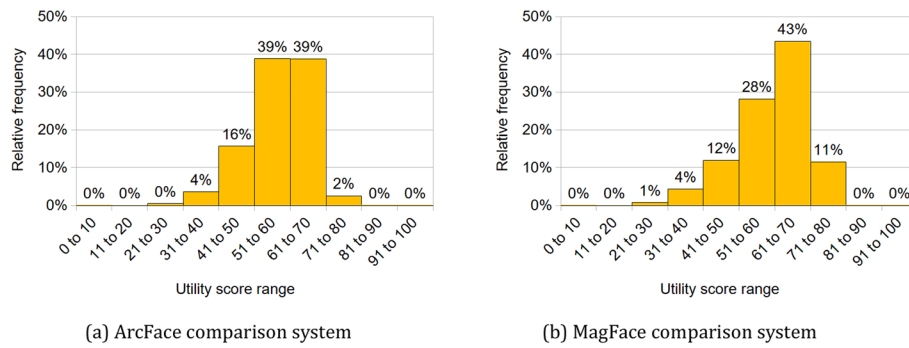


Fig. 1 Utility score distributions in the face image data set with respect to the face comparison systems

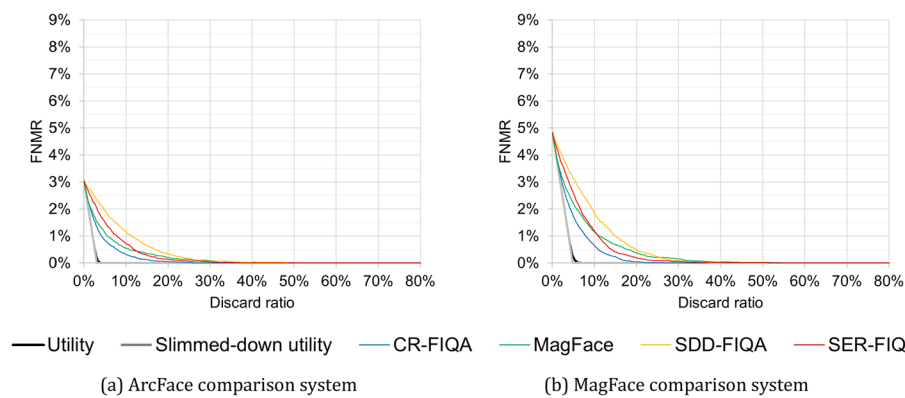


Fig. 2 False non-match EDCs in the face image data set at a fixed decision threshold for an initial FMR of 0.1%

and SER-FIQ adopt unsupervised deep-learning-based approaches not requiring any utility labels for training.

Before applying the quality assessment algorithms, we aligned the face images by applying an affine transformation based on the facial landmarks detected by MTCNN. The required size of aligned input images happens to be 112×112 pixels in all cases.

4.2 Experimental results and discussion

4.2.1 Observed utility

We calculated utility scores u_i^* (Eq. 6) using the face image data set and comparison systems given in Sect. 4.1. Figure 1 shows the distributions of utility scores of the face images in the data set with respect to both face comparison systems. Despite the different poses, illumination conditions and facial expressions, most images in the data set are of adequate quality and were assigned a utility score higher than 50.

4.2.2 Error vs. discard characteristics

Figures 2, 3, 4, 5 show false match EDCs and false non-match EDCs for both face comparison systems with respect to utility scores (bold black lines), with respect to slimmed-down utility scores (bold grey lines) and, for comparison, with respect to the various face image quality scores (colourful lines). While we fixed the decision

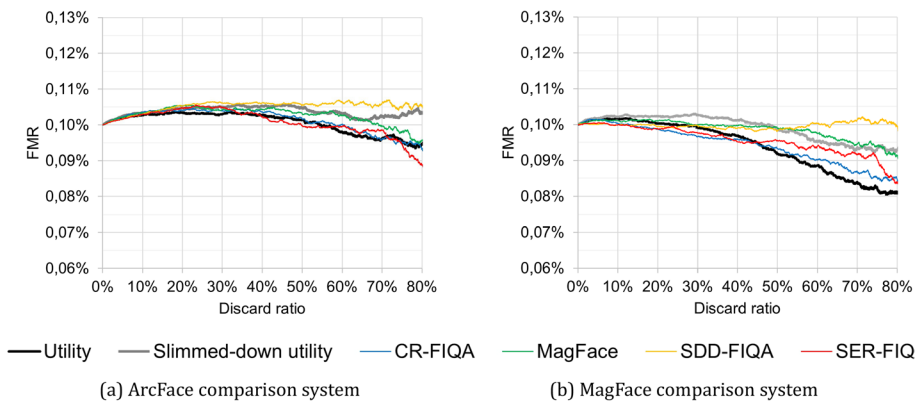


Fig. 3 False match EDCs in the face image data set at a fixed decision threshold for an initial FMR of 0.1%

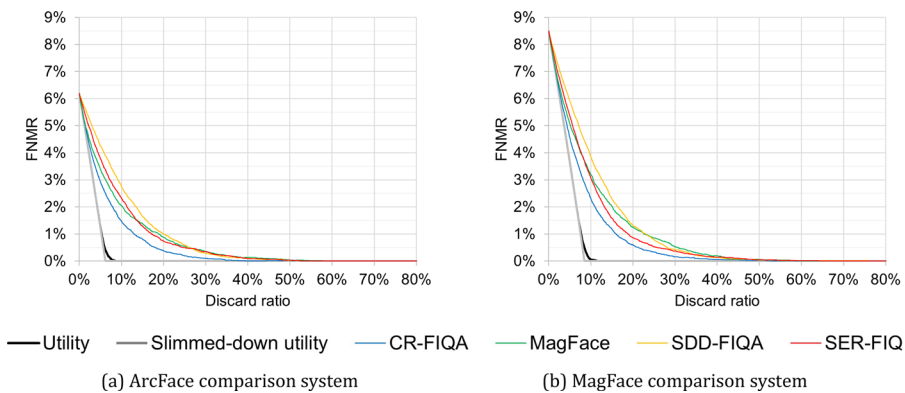


Fig. 4 False non-match EDCs in the face image data set at a fixed decision threshold for an initial FMR of 0.01%

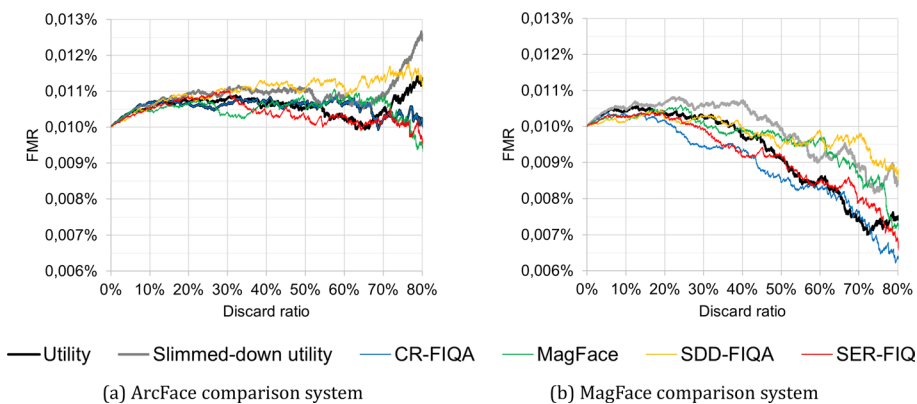


Fig. 5 False match EDCs in the face image data set at a fixed decision threshold for an initial FMR of 0.01%

threshold in Figs. 2 and 3 to give an initial FMR value of 0.1%, in Figs. 4 and 5 the initial FMR value is 0,01%.

The bold black and grey lines in Figs. 2 and 4 show the desired behaviour: discarding images with low utility scores leads to a steep drop in FNMR. Starting from an initial

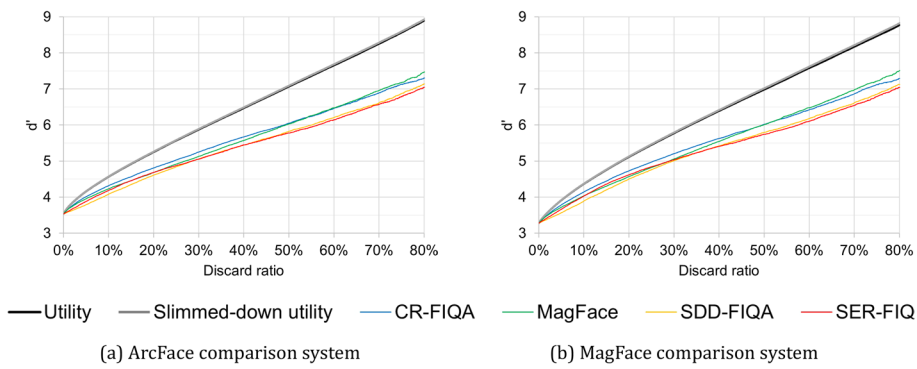


Fig. 6 d' vs. discard characteristics in the face image data set with respect to the face comparison systems

value of about 3% in Fig. 2a or about 5% in Fig. 2b, the FNMR approaches 0% after discarding the worst 3% or 5% of the images. In Figs. 2 and 4, the bold grey lines largely cover the bold black lines. For the face image quality assessment experiments with canonical references, discarding images based on utility scores without consideration of non-mated comparisons (Eq. 3) leads to a slightly steeper decline in FNMR than with consideration of non-mated comparisons (Eq. 1). The deviation is visible only for FNMR values below 1%. The bold black and grey lines in Figs. 3 and 5 show that discarding images based on utility scores with consideration of non-mated comparisons (Eq. 1) slightly improves the FMR while discarding images based on utility scores without consideration of non-mated comparisons (Eq. 3) improved the FNMR in Figs. 2 and 4 slightly more. The consideration of non-mated comparison scores leads to the assignment of lower utility scores to samples responsible for false matches and consequently to their earlier discard. As false matches of face images are caused by look-alikes and cannot be prevented by means of quality assessment, the slimmed-down utility formula without consideration of non-mated comparison scores provides more suitable target utility labels.

The colourful lines in Figs. 2 and 4 show that all face image quality assessment algorithms under consideration achieve a more or less steep drop in FNMR. The colourful lines in Figs. 3 and 5 show that discarding images based on low-quality scores has various small effects on FMR with different face image quality assessment algorithms. Therefore, to ensure a fair comparison of quality assessment algorithms, false non-match and false match EDCs are recommended to be explored together. Note, however, the different orders of magnitude of FNMR and FMR.

4.2.3 d' vs. discard characteristics

Figure 6 shows the d' vs. discard characteristics for both face comparison systems with respect to utility scores (bold black lines), slimmed-down utility scores (bold grey lines) and, for comparison, face image quality scores (colourful lines). For each quality assessment algorithm, we used the same data set and comparison system to enable the comparison of the performance of the quality assessment algorithms. The d' vs. discard characteristics show that exclusion of images with low-quality scores leads to a noticeable improvement in the separability of the mated and non-mated comparison score

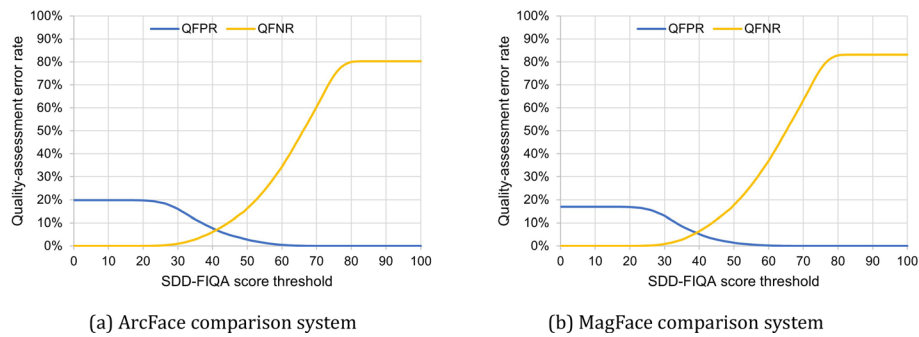


Fig. 7 Quality-assessment error rates in the face image data set over the SDD-FIQA score threshold

distributions. Among the face image quality assessment algorithms under test, CR-FIQA and MagFace achieve the best results.

4.2.4 Prediction errors

We consider the quality scores delivered by the face image quality assessment algorithms introduced in Sect. 4.1.3 as is. Of the studied algorithms only SDD-FIQA covers the desired score range from 0 to 100 [1]. In the studied data set, for the SDD-FIQA scores (which range from 13.0 to 85.4), RMSE is as low as 9.5 with respect to utility for the ArcFace comparison system and 9.4 with respect to utility for the MagFace comparison system.

The CR-FIQA, MagFace and SER-FIQ quality scores do not cover the range from 0 to 100. For instance, MagFace scores, in general, range from 0 to 40. Therefore, they are not amenable to comparison with respect to RMSE. We refrain from calibrating the quality scores ourselves. They can be calibrated on training data by selecting meaningful score values marking borders between deficient, marginal, adequate or excellent quality as described in Sect. 2.3 for utility scores.

4.2.5 Quality assessment error rates

Figure 7 shows QFNR and QFPR for SDD-FIQA on the chosen data set over the quality score threshold.

With a higher quality score threshold, the QFPR decreases and the QFNR increases. The QFPR value at the quality score threshold of zero indicates the percentage of images in the data set whose utility score is below 50 with respect to the chosen comparison system. It is up to the system designer to select an appropriate quality score threshold that results in a trade-off between QFPR and QFNR, i.e. between mistakenly retaining objectionable samples and mistakenly discarding unobjectionable samples.

5 Fingerprint image quality assessment experiments

5.1 Experimental setup

5.1.1 Fingerprint image data sets

We used the following publicly available data sets, which contain multiple images per finger instance:

- (a) Subset A of data set DB3 from the First Fingerprint Verification Competition (FVC2000) [39], which consists of 800 plain fingerprint images of 100 different fingers (left and right index and middle fingers), eight images per finger. The fingerprint images were captured using an optical fingerprint sensor (Identicator Technology DF-90). The image size is 448×478 pixels. All fingerprint images are 8-bit grey-scale images with a spatial sampling rate of 500 dpi (i.e. 19,685 pixels per centimetre).
- (b) A subset of the CASIA Fingerprint Image Database Version 5.0 (CASIA-FingerprintV5) [40], restricted to 2500 plain fingerprint images of 500 different right index fingers, five images per finger. The images were captured using an optical fingerprint sensor (URU4000) all in one session. To generate noticeable intra-class variation within a single session, the data subjects rotated their fingers with various levels of pressure. The image size is 328×356 pixels. All fingerprint images are 8-bit grey-scale.bmp files with a spatial sampling rate of 512 dpi (i.e. 20,157 pixels per centimetre).

5.1.2 Fingerprint comparison

We extracted minutiae using the open-source FingerNet framework [41], which deploys a deep neural network for minutia detection. No failures to extract occurred. Minutia positions and angles were converted to MCC (Minutia Cylinder-Code) format [42] for comparison. Each comparison yielded a dissimilarity (or distance) score (the more similar, the lower).

In FVC2000 DB3, each fingerprint image was compared with each other image in the data set (7 mated and 792 non-mated comparisons). In the subset of CASIAFingerprintV5, each fingerprint image was compared with each mated fingerprint image (4 mated comparisons) and with 1245 non-mated fingerprint images.

5.1.3 Fingerprint image quality assessment

To assess the fingerprint image quality, we used the following open-source fingerprint image quality assessment algorithms:

- (a) NFIQ (NIST Fingerprint Image Quality) version 2.2 [6]: NFIQ 2 is the reference implementation of ISO/IEC 29794-4 [7]. NFIQ 2 was trained on fingerprint images captured using optical sensors based on frustrated total internal reflection and fingerprint images scanned from inked fingerprint cards, all with a spatial sampling rate of 500 dpi. For fingerprint images from other sensor technologies or with a different spatial sampling rate, the NFIQ 2 score is in general not predictive of utility.
- (b) MiDeCon [43]: This algorithm builds upon MinutiaeNet [44] for robust minutiae extraction from fingerprint images. By computing multiple forward passes with activated dropout layers in the neural network, the mean and variance of the prediction are computed and turned into a reliability score for each minutia. The fingerprint quality score is then derived from the mean of the reliability scores of the extracted minutiae.

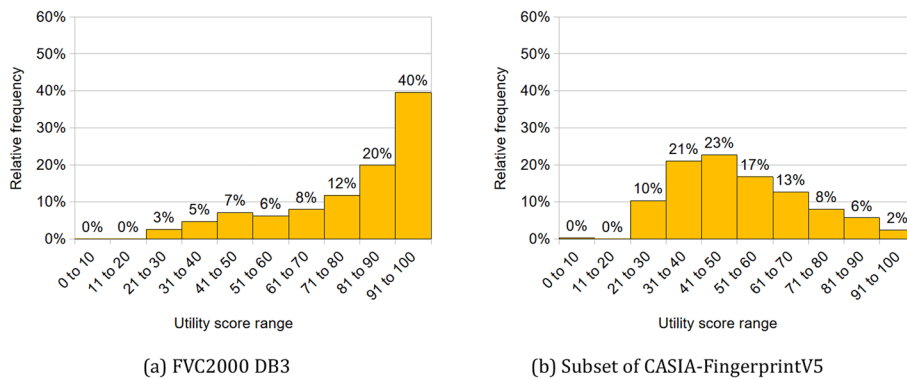


Fig. 8 Utility score distributions in the fingerprint image data sets

To be able to use NFIQ 2, we chose a data set consisting of fingerprint images of 500 dpi (FVC2000 DB3) and re-sampled the fingerprint images in the other data set (subset of CASIA-FingerprintV5) with a spatial sampling rate of 500 dpi, using the free Cognaxon WSQ viewer, version 4.1.

5.2 Experimental results and discussion

5.2.1 Observed utility

We calculated utility scores u_i^* (Eq. 6) using the comparison scores from the fingerprint image data sets and comparison system given in Sect. 5.1. Figure 8 shows the distribution of utility scores of the fingerprint images in the fingerprint image data sets. Most images in FVC2000 DB3 are of unobjectionable quality, i.e. easily recognizable by the comparison software used. This is due to quality assurance measures during data collection. In the subset of the CASIA-FingerprintV5 data set, the utility scores are more evenly distributed because many fingerprints in this data set are not easily recognizable. The presence of low-utility images makes CASIA-FingerprintV5 especially suitable for studying fingerprint image quality assessment.

5.2.2 Error vs. discard characteristics

Figure 9 through Fig. 12 show the false match and false non-match EDCs for both fingerprint image data sets with respect to utility scores and slimmed-down utility scores (bold black and grey lines) and, for comparison, with respect to NFIQ 2.2 and MiDeCon scores (colourful lines). While we fixed the decision threshold to give an initial FMR value of 1% in Figs. 9 and 10, in Figs. 11 and 12 we fixed it to give an initial FMR value of 0.1%. The bold black and grey lines in Figs. 9 and 11 show the desired behaviour as well: discarding images with low utility scores leads to a clear drop in FNMR. In Fig. 9 through Fig. 12, the bold grey lines and the bold black lines overlap. For the fingerprint image quality assessment experiments without canonical references, discarding images based on utility scores with or without consideration of non-mated comparisons makes no significant difference.

The colourful lines in Figs. 9 and 11 show that both fingerprint image quality assessment algorithms under consideration achieve some drop in FNMR. The colourful lines in Figs. 10 and 12 show that discarding images based on low-quality scores has

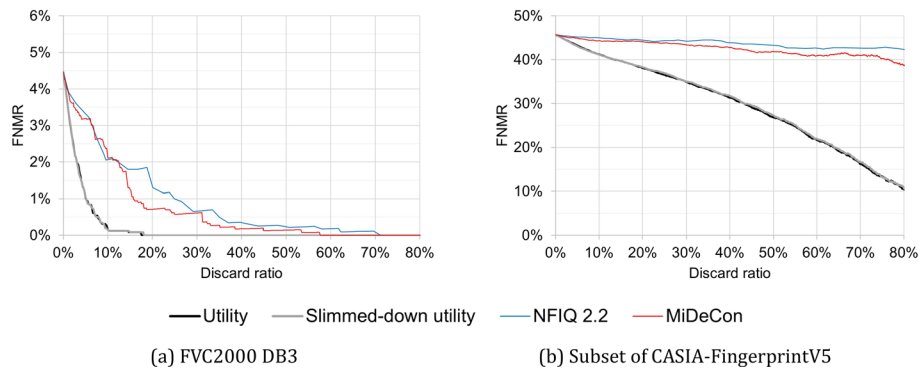


Fig. 9 False non-match EDCs in the fingerprint image data set at a fixed decision threshold for an initial FMR of 1%

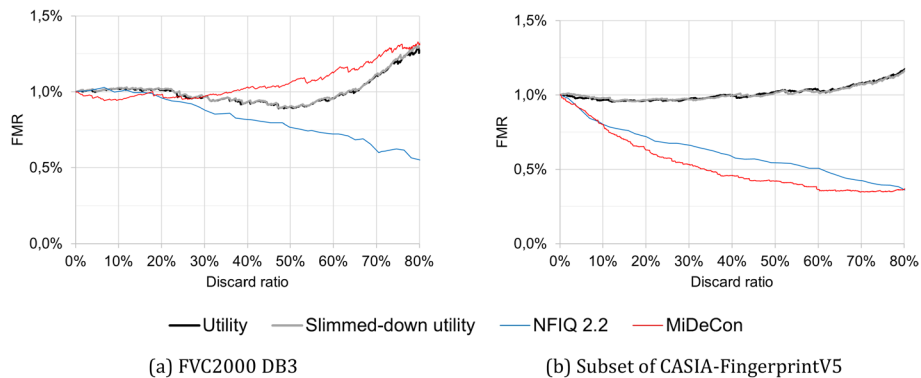


Fig. 10 False match EDCs in the fingerprint image data set at a fixed decision threshold for an initial FMR of 1%

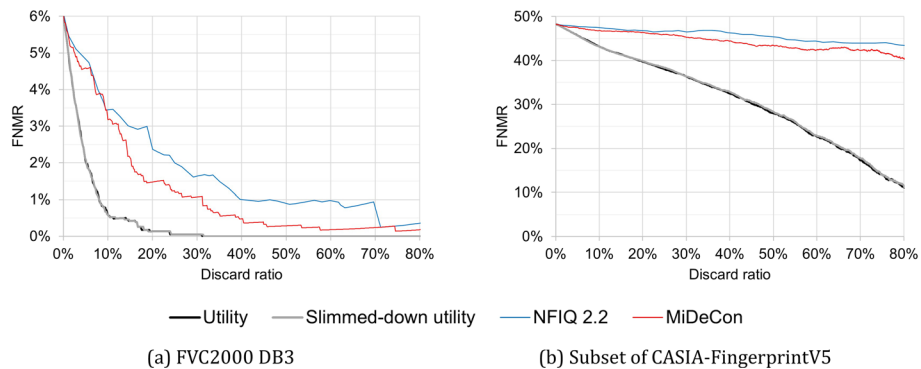


Fig. 11 False non-match EDCs in the fingerprint image data set at a fixed decision threshold for an initial FMR of 0.1%

various effects on FMR with different fingerprint image quality assessment algorithms. Discarding images with low NFIQ 2.2 scores leads not only to a decreasing FNMR, but also to a decreasing FMR in both data sets.

For the FVC2000 DB3 data set, MiDeCon performs better than NFIQ 2 in the false non-match EDCs (see Figs. 9a and 11a) but worse than NFIQ 2 in the false match EDCs

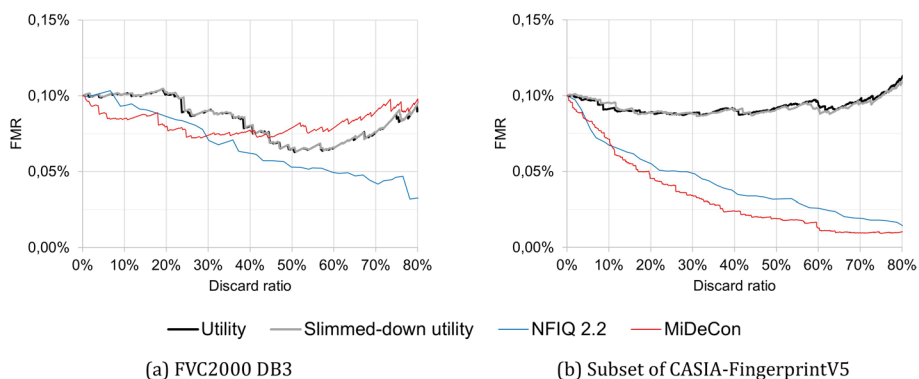


Fig. 12 False match EDCs in the fingerprint image data set at a fixed decision threshold for an initial FMR of 0.1%

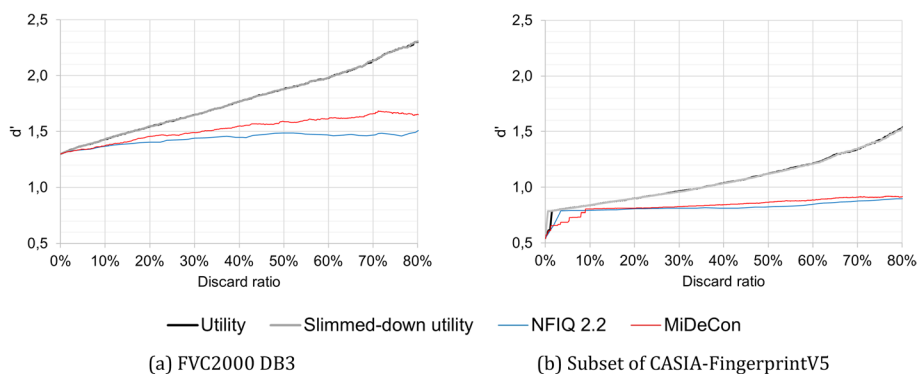


Fig. 13 d' vs. discard characteristics in the fingerprint image data sets

(see Figs. 10a and 12a). Which algorithm performs better overall? This can be seen in the d' vs. discard characteristics (Fig. 13).

5.2.3 d' vs. discard characteristics

Figure 13 shows the d' vs. discard characteristics for both fingerprint image data sets with respect to utility scores and slimmed-down utility scores (bold black and grey lines) and, for comparison, with respect to fingerprint image quality scores (colourful lines). Figure 13 shows that exclusion of images with low-quality scores leads to an improvement in the separability of the mated and non-mated comparison score distributions though not to the extent that would be possible if all hard-to-recognize fingerprint images could be discovered. The results show that newer machine-learning approaches such as MiDeCon can outperform NFIQ 2.

5.2.4 Prediction errors

The RMSE between the NFIQ 2.2 scores (which range from 0 to 91 in FVC2000 DB3) and the utility scores observed in the FVC2000 DB3 data set of mostly good quality is 46.6. The RMSE between the NFIQ 2.2 scores and the utility scores observed in the subset of the CASIA-FingerprintV5 data set is 26.5. As there are merely 100 score points, this is quite a big deviation. One reason is that NFIQ 2 was not trained for predicting

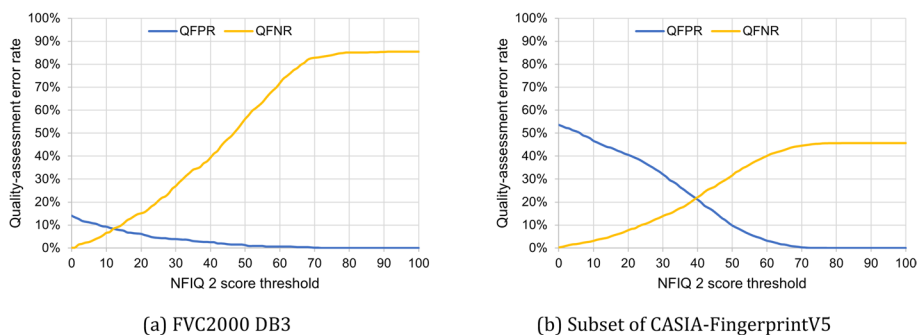


Fig. 14 Quality-assessment error rates over the NFIQ 2 score threshold in the fingerprint image data sets

continuous utility scores but for binary classification into two utility classes (high or low utility).

The MiDeCon scores do not cover the range from 0 to 100. Therefore, without calibration MiDeCon is not amenable to comparison with respect to RMSE. We refrain here from calibrating the quality scores. They can be calibrated in the same way as described in Sect. 2.3 for utility scores.

5.3 Quality assessment error rates

Figure 14 shows QFNR and QFPR for the chosen fingerprint image data sets and comparison algorithm over the quality score threshold. The quality score threshold for discarding low-quality samples depends on the intended use of the retained samples. For instance, for public sector applications in Germany, a technical guideline [45] specifies that the NFIQ 2 score threshold for plain left and right index-finger images captured for enrolment purposes is 30. At the quality-score threshold of 30, for the subset of CASIAFingerprintV5 about 32% of the retained finger images are of objectionable (low) quality, and about 14% of the discarded finger images are of unobjectionable (high) quality. For data sets containing fewer low-quality finger images from the beginning (like FVC2000 DB3), QFPR is much lower. Therefore, technical and organizational measures are needed to capture few bad fingerprint images in the first place.

The quality-score threshold values in [45] are the result of a cost–benefit analysis considering the costs of the two types of errors, i.e. of mistakenly discarding unobjectionable samples and of mistakenly retaining objectionable samples.

6 Conclusion

Defining the utility of a biometric sample based on comparison scores for this sample allows the effective detection of samples responsible for false non-match errors. Discarding samples with low utility scores results in a rapidly declining FNMR in face image as well as fingerprint image quality assessment experiments.

The experiments showed no contribution of non-mated comparison scores to the detection of samples responsible for false non-matches. In case of canonical references, the consideration of non-mated comparison scores contributed to detecting samples responsible for false matches but slightly delayed the detection of samples responsible for false non-matches. In case of physiological biometric characteristics such as faces

and fingerprints, false matches are caused by biometric look-alikes and cannot be prevented by means of biometric sample quality assessment. Therefore, we conclude that the slimmed-down utility formula without consideration of non-mated comparison scores is preferable in case of physiological biometric characteristics such as faces and fingerprints. Non-mated comparison scores do, however, contribute to utility score calibration by putting mated and non-mated comparison scores in relation to each other.

In future work, utility scores observed in a training data set covering a wide range of potential quality issues could be used as target utility labels for training biometric sample quality assessment algorithms. The utility scores observed in a testing data set can be used as a benchmark for evaluating the performance of biometric sample quality assessment algorithms.

Discarding low-quality samples affects not only the FNMR, but also the FMR even if not intended. For reasons of fairness, both false non-match EDC and false match EDC should be considered in comparative performance evaluations of biometric sample quality assessment algorithms. Summarizing utility-prediction performance in a single plot or a single figure of merit makes the comparison of quality assessment algorithms easier.

Acknowledgements

Not applicable.

Author contributions

OH made substantial contributions to the conception and design of the work. BF and AK made substantial contributions to the analysis and interpretation of the data. AK created the new Python scripts used in the work. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

Availability of data and materials

The lists of samples from the public data sets used in this study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 17 May 2024 Accepted: 21 August 2024

Published online: 09 September 2024

References

1. Information technology—Biometric sample quality—Part 1: Framework. International Standard ISO/IEC 29794-1 (2024)
2. Portrait quality (reference facial images for MRTD). ICAO Technical Report (2018)
3. E. Tabassi, C.L. Wilson, C.I. Watson, Fingerprint image quality. NIST Interagency Report 7151, NIST (2004)
4. O. Henniger, B. Fu, C. Chen, Utility-based performance evaluation of biometric sample quality assessment algorithms. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG) (2022)
5. O. Henniger, Utility prediction performance of finger image quality assessment software. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG) (2023)
6. E. Tabassi, M. Olsen, O. Bausinger, C. Busch, A. Figlarz, G. Fiumara, O. Henniger, J. Merkle, T. Ruhland, C. Schiel, M. Schwaiger, NFIQ 2.0—NIST fingerprint image quality. NIST Interagency Report 8382, NIST (2021)
7. Information technology—Biometric sample quality—Part 4: Finger image data. International Standard ISO/IEC 29794-4 (2017)
8. Happold, M.: Learning to predict match scores for iris image quality assessment. In: Proceedings of the International Joint Conference on Biometrics (IJCB) (2014)

9. L. Best-Rowden, A.K. Jain, Learning face image quality from human assessments. *IEEE Trans Inf Forens Secur* **13**(12), 3064–3077 (2018)
10. J. Hernandez-Ortega, J. Galbally, J. Fierrez, L. Beslay, Biometric quality: review and application to face recognition with FaceQnet. *CoRR abs/2006.03298* (2020)
11. F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, Y.-G. Wang, SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). Computer Vision Foundation / IEEE
12. J. Chen, Y. Deng, G. Bai, G. Su, Face image quality assessment based on learning to rank. *IEEE Signal Process. Lett.* **22**(1), 90–94 (2015)
13. P. Wasnik, K.B. Raja, R. Ramachandra, C. Busch, Assessing face image quality for smartphone-based face recognition system. In: *Proceedings of the 5th International Workshop on Biometrics and Forensics IWBF* (2017)
14. O. Henniger, B. Fu, C. Chen, On the assessment of face image quality based on handcrafted features. In: *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)* (2020)
15. Guest, R., Henniger, O.: Assessment of the quality of handwritten signatures based on multiple correlations. In: *Proceedings of the International Conference on Biometrics (ICB)* (2013)
16. P. Grother, A. Hom, M. Ngan, K. Hanaoka, Ongoing face recognition vendor test (FRVT)—Part 5: Face image quality assessment. Draft NIST Interagency Report (for public comment), NIST (2022)
17. R.M. Bolle, S. Pankanti, N.K. Ratha, Evaluation techniques for biometrics-based authentication systems (FRR). In: *Proceedings of the 15th International Conference on Pattern Recognition (ICPR)* (2000). IEEE Computer Society
18. J. Daugman, Biometric decision landscapes. Technical Report UCAM-CL-TR482, University of Cambridge—Computer Laboratory (2000)
19. T.K. Ho, M. Basu, Measuring the complexity of classification problems. In: *Proceedings of the 15th International Conference on Pattern Recognition (ICPR)* (2000). IEEE Computer Society
20. Information technology—Biometric data interchange formats. International Standards Series ISO/IEC 19794
21. Information technology—Extensible biometric data interchange formats. International Standards Series ISO/IEC 39794
22. P. Grother, E. Tabassi, Performance of biometric quality measures. *IEEE TPAMI* (2007). <https://doi.org/10.1109/TPAMI.2007.1019>
23. Information technology—biometric application programming interface—Part 1: BioAPI specification. International Standard ISO/IEC 19784-1 (2018)
24. T. Schlett, C. Rathgeb, J. Tapia, C. Busch, Considerations on the evaluation of biometric quality assessment algorithms. *IEEE Trans. Biometr. Behav. Identity Sci.* **6**(1) (2024)
25. P. Terhörst, J.N. Kolf, N. Damer, F. Kirchbuchner, A. Kuijper, SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
26. Q. Meng, S. Zhao, Z. Huang, F. Zhou, MagFace: a universal representation for face recognition and quality assessment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). Computer Vision Foundation / IEEE
27. F. Boutros, M. Fang, M. Klemt, B. Fu, N. Damer, CR-FIQA: Face image quality assessment by learning sample relative classifiability. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2023)
28. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. Recommendation ITU-T P.1401 (2020)
29. C. Yan, T. Teng, Y. Liu, Y. Zhang, H. Wang, X. Ji, Precise no-reference image quality evaluation based on distortion identification. *ACM Trans. Multimed. Comput. Commun. Appl.* (2021). <https://doi.org/10.1145/3468872>
30. Information technology—biometric performance testing and reporting—Part 1: principles and framework. International Standard ISO/IEC 19795-1 (2021)
31. R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE. In: *8th IEEE International Conference on Automatic Face & Gesture Recognition* (2008)
32. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**(10), 1499–1503 (2016)
33. J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019)
34. D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014)
35. Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, VGGFace2: a dataset for recognising faces across pose and age. In: *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 67–74 (2018)
36. Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: *14th European Conference on Computer Vision*, pp. 87–102 (2016). Springer
37. Trillionpairs. <http://trillionpairs.deepglinternational.com/overview>
38. T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, C. Busch, Face image quality assessment: a literature survey. *ACM Comput. Surv.* (2022). <https://doi.org/10.1145/3507901>
39. D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman, A.K. Jain, FVC2000: fingerprint verification competition. *IEEE TPAMI* (2002). <https://doi.org/10.1109/34.990140>
40. CASIA Fingerprint Image Database (CASIA-FingerprintV5). <http://biometrics.idealtest.org/> (2009)
41. Y. Tang, F. Gao, J. Feng, Y. Liu, FingerNet: a unified deep network for fingerprint minutiae extraction. In: *Proceedings of the International Joint Conference on Biometrics (IJCB)* (2017)
42. R. Cappelli, M. Ferrara, D. Maltoni, Minutia cylinder-code: a new representation and matching technique for fingerprint recognition. *IEEE TPAMI* (2010). <https://doi.org/10.1109/TPAMI.2010.52>
43. P. Terhörst, A. Boller, N. Damer, F. Kirchbuchner, A. Kuijper, MiDeCon: unsupervised and accurate fingerprint and minutia quality assessment based on minutia detection confidence. In: *Proceedings of the International Joint Conference on Biometrics (IJCB)* (2021)

44. D.-L. Nguyen, K. Cao, A. Jain, Robust minutiae extractor: integrating deep networks and fingerprint domain knowledge. In: Proceedings of the International Conference on Biometrics (ICB) (2018)
45. Biometrics for public sector applications—Part 3: application profiles, function modules and processes. BSI Technical Guideline TR-03121-3

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.