A Literature Review of Fingerprint Quality Assessment and Its Evaluation

Zhigang Yao, Jean-Marie Le Bars, Christophe Charrier, Christophe Rosenberger

A Literature Review of Fingerprint Quality Assessment and Its Evaluation

Zhigang Yao[1], Jean-Marie Le Bars[1], Christophe Charrier[1], Christophe Rosenberger[2]

[1]Universite de Caen Basse Normandie, Caen, Calvados, France
[2]ENSICAEN, Caen, Calvados, France
[1, 2]UMR 6072 GREYC, Caen, Calvados, France
[1, 2]ENSICAEN Site B, 17 Rue Claude Bloch, Caen 14000, France

Abstract

Fingerprint quality assessment (FQA) has been a challenging issue due to a variety of noisy information contained in the samples, such as physical defect and distortions caused by sensing devices. Existing studies have made efforts to find out more suitable techniques for assessing fingerprint quality but it is difficult to achieve a common solution because of, for example, different image settings. This paper gives a twofold study related to FQA, including a literature review of the prior work in assessing fingerprint image quality and the associated evaluation approaches. First, we categorized some representative studies proposed in last few decades to show how this problem has been solved so far. Second, the paper gives a brief introduction of the associated evaluation approaches, and then contributes an extended evaluation framework based on the enrollment selection, which offers repeatable and statistically convincing measures for evaluating quality metrics. Experimental results demonstrate the usability of the proposed evaluation framework via offline trials.

Index Terms

Fingerprint, quality assessment, evaluation, biometrics.

I. Introduction

In biometrics, fingerprint is recognized as the most common modality among all biometric characteristics that have been studied so far [1], [2]. The advantages of the fingerprint could be attributed to several factors: 1) fingerprint pattern is stable and invariant which completely satisfies the requirement of uniqueness for being a biometric modality and 2) the use of fingerprint is also acceptable to people in comparison with other kinds of biometric modalities [3], [4], [2]. Therefore, besides the traditionally official applications, fingerprint also being adopted in civilian area since its study in modern biometrics had been proposed the 1960s [5], [6]. For instance, the automatic fingerprint identification system (AFIS) can be easily found in office buildings, and some similar applications are also provided on the laptop computers, smart phones, etc.

The performance of these applications largely depends on the reliability of features extracted from the sensed fingerprint image. Existing studies have shown that fingerprint quality is significant to the reliability of the features [7]. A fingerprint, as a common sense, is composed of two kinds of lines namely fingerprint ridge and valley [6] which alternately run on the fingertip surface. The discontinuous flowing of ridge lines generate two types of minor features known as ridge ending and ridge bifurcation which are categorized as the minutiae points [8].

Matching algorithms of fingerprint mainly use minutiae-based features, particularly restricted to these two types of minutiae points [8], [2]. The accuracy of a matching algorithm therefore depends on the reliability of the detected features of fingerprint images. However, these features could be easily affected by noisy information contained in the image, such as noise pixels and abrupt break of ridge lines caused by tiny wrinkles. In addition to such impacts, there are also some additional factors that influence the reliability and precision of the extracted features, and hence reflect the quality of fingerprint images [9]. For these reasons, earlier studies chose to improve the reliability of detected features via postprocessing [8] or to improve the quality of fingerprint images with enhancement and other preprocessing approaches [10], [8], [11], [12]. Figure 1 shows examples of the effect of enhancement (fig. 1c and 1e) and noises (fig. 1b and 1d) to matching results. As a result, fingerprint recognition systems employ quality control to guarantee the quality of captured samples [14]. Alternatively, matching algorithm for low quality image could also be implemented via other features rather than minutiae points [15], [16]. Furthermore, recent studies pay attention to high resolution image which is able to provide higher quality image, and the level-3 features could be available for more secure applications [17]. This paper chiefly discusses quality assessment of the gray-level fingerprint images, which has been the main focus of most of the existing studies [18].

The first study of the FQA had been proposed in the late 1990s [19]. Bolle *et al.* qualify the fingerprint image in terms of the orientation of local block. Soon afterward, Nalini *et al.* defined fingerprint quality in terms of several aspects, including

(a) 11_2 (b) 11_1 (c) 11_7 (d) Noisy 11_1 (e) Enhanced 11_7

Fig. 1: Illustration of the effect of enhancement and noises to matching result. The matching score between 11_2 (from FVC2002DB2A) and the other four are 231, 135, 217 and 196, respectively. NBIS [13] Bozorth3 is used for matching.

the clarity of the fingerprint image, separation of the ridge-valley pattern and the contrast. These factors are basically consistent with some subjective criteria for estimating the quality of an image [20]. However, fingerprint quality is not only determined by image itself, but also affected by matching algorithm as matching is generally dependent on features. Therefore, insufficiency of features can lead to low matching result. Further, it gives negative impacts on the recognition performance. In this case, biometric standards consider three components of the sample quality known as character, fidelity and utility [21]. The utility is a function of both the character and fidelity of a sample [22] and is generally presented as the contribution of a biometric sample to the performance in terms of recognition errors. It means that a good quality sample should be beneficial to matching operations. To achieve such an objective, a good quality fingerprint sample should be not only a clear image that satisfies subjective assessment criteria, but also an image that is suitable for extracting sufficient and reliable features. A state-of-the-art literature has defined that the quality of a biometric sample is a predictor of the matching performance [23]. Grother *et al.* also pointed out that the quality is not linearly predictive to the matching performance (genuine match). Besides, the prediction largely depends on the employed matcher which impacts on the efficacy of the quality metric.

According to the statement above, one can note that FQA is still an open issue and has been involved in some new branches such as quality metric for fingerprint images collected via mobile phone camera [24]. To demonstrate the progress in FQA and its evaluation, this paper firstly contributes a quick up-to-date literature review of the FQA, and then a brief review of evaluation approaches related to the FQA is given. At last, a systematic definition of the proposed validation approach is presented. The general point of view of this paper mainly follows the utility of biometric sample. Besides, the paper is not to give a comparative study between different quality metrics.

The organization of the paper is given as follows: In Section II, the paper presents a literature review of FQA in terms of how they are implemented. Section III addresses independently the evaluation approaches related to fingerprint quality metrics. Furthermore, in this section, a generic evaluation approach based on the enrollment selection [25] is detailed with an extension. Conclusions are presented in Section V.

II. Literature Review of FQA

Fernandez *et al.* [26] proposed a comparative study of FQA prior to 2006, in which they categorized FQA algorithms into several classes known as local feature-based approaches, global feature-based methods and solutions with classifiers [26]. Those quality metrics can be simply summarized in several points: quality metrics based on the orientation of fingerprint pattern; algorithms that rely on the variation of Gabor responses; approaches in frequency domain; measurements based on pixel information and quality indexes rely on classification with multi-feature. In addition, that study also analyzed quality metrics mainly in terms of the linearity between them.

In this study, we classify the existing studies into several frameworks in terms of their implementation to show the difference and some potential problems that need to be considered. As mentioned above, the quality metrics that had been proposed so far are all dependent on one or several features. According to how they are carried out, this study categorizes them as: 1) segmentation-based approaches; 2) a single feature-based quality index; 3) solutions rely on a combination of multi-features or indexes, which is further divided into methods based on linear fusion and classification. Table I shows a categorization of some of the representative studies in FQA.

In fact, FQA had been involved in some relevant studies, such as the contrast of image block [8] and the region mask of ridge-valley pattern [12]. In table I, work on FQA had been followed by an approach based on the segmentation of the directional block and non-directional block [19]. The quality index is represented via a ratio of the area of qualified blocks

TABLE I: Categories of existing fingerprint quality metrics.

Framework	Solution	Quality index or feature.
Segmentation	Image-based	Directional blocks area of foreground [19] (1999). Poor foreground blocks [27] determined with Gabor features in 8 directions (2001). Sum of directional contrast of local block [28] (2005).
	Template-based	Reasonable informative region [29] measured by Delaunay Triangulation of minutiae (2015).
Single feature	Feature regularity	Cumulative total energy (CTE) in the initial few subbands of the wavelet transform of fingerprint [30] (1999). Sum of pixels standard deviation (STD) of local block[28] (2005). Sub-band energy of fingerprint Fourier spectrum [28] (2005). Weight of block symmetry [31] generated by 2-order orientation tensor (2006). Square root of the absolute value of the Pet's hat wavelet (PHCWT) coefficients [32] (2007). Relative contrast index (CI) [33], a logrithmic ratio of the reflective intensity of the valleys to the one of the ridges (2008). The shape of the probability distribution functions (PDFs) of ridge orientation and ridge-to-valley orientation [34] (2008).
Multi-feature	Linear fusion	Orientation certainty level (OCL); Ridge frequency, ridge-to-valley thickness ratio and ridge thickness; Continuity and uniformity [35] (2002). Global clarity score (GCS) and global orientation quality score (GOQS) [36] (2004). Gabor feature in [27]; Ratio of foreground blocks to fingerprint image blocks; Central position [37] (2005). Residual variances and mainfold topology structure of the PCA of a block feature vector and block Harris-corner strength [38] (2012). BLIINDS; SIFT; Root means square (RMS) of image block [39] (2013).
	Classification	Local SNR of the DFT of a signature (sine wave), unimformity and curvature [40] (2004). 11-dimensional feature vector includes most of the existing features [23] (2004). Auto-correlation and DCT-based features of image block [24] (2013). A histogram of the unit activations of Self-organizing maps (SOM) obtained by training a SOM (Neural Network) with image block [41] (2013).

to the whole image. However, the determination of block prominent direction depends on the threshold which is a thorny problem for a common application. In practical, the segmentation-based measures for FQA are generally used in two ways, one is to represent the quality via the qualified foreground area and another is to segment foreground from the image at first. For instance, Shen *et al.* [27] use the regularity of 8-direction Gabor features to generate quality index. Their Gabor feature is initially used for segmenting foreground from the image, which is also involved in a threshold. In addition, segmentation-based measures proposed in [42] were also used for image quality assessment [28].

The use of segmentation presented above are all associated with fingerprint image. Yao *et al.* [29] proposed an approach (MQF) with minutiae template only, in which the convex-hull and Delaunay triangulation are adopted to measure the area of a reasonable informative region. This algorithm is hence dependent on a minutiae extracting operation. In addition, some bad quality images that own relatively large informative region are more possibly to give outliers. According to these literature, one can note that foreground area is indeed a good factor for qualifying fingerprint image. In this case, multiple segmentation could be a potential solution to generate area-based quality metric.

The second category discusses quality metrics that rely on a single feature which could be applied locally or globally to the image. For example, Nalini *et al.* [30] proposed to use cumulative energy of several subbands of the compressed image in the wavelet domain. Lee *et al.* [28] reviewed three approaches based on the fingerprint image, including local standard deviation [42], directional contrast of local block [43] and the Gabor feature [27]. They proposed a feature via observing the Fourier spectrum of the fingerprint image. Their quality metric depends on the pixels information of the Fourier spectrum image which is a floating measure for different kinds of image settings. Other quality metrics denoted by a single feature could also be found in [31], [33], [34], where the symmetry features decomposed via 2-order orientation tensor [31] depending on scale parameter and threshold, the contrast index (CI) relies on a mean spectrum of ridge-valley measurements, and the difference of kurtosis value of the probability density functions (PDFs) [34] is not distinctive between some convex and concave shapes that are relatively smooth. Trial results in these literature show relatively good performance comparing with baseline algorithm(s). However, threshold values and parameters are unavoidable for most of them and lead to difficulties for achieving a generic application because they are greatly affected by image specification.

In addition, this kind of approach also can be found in [37], [44], [38], where Chen *et al.* [44] estimate the power spectrum ring with Butterworth functions instead of observing directly the pixel information of the spectrum image, and Tao *et al.* [38] observed two regularities from the circle mainfold topology of an order set of block pixels and the associated principle component analysis (PCA). However, in addition to the coefficient problem, there are also constraints of the employed

features. For example, the ridge frequency [44] depends on the resolution and image size.

Many of the existing studies made effort in qualifying fingerprint image with multiple features. This is generally carried out in two aspects: linear fusion with weighted coefficients and classification. Both could be associated with knowledge-based schemes [23], [39]. For instance, Lim *et al.* [35] proposed a quality metric through weighted combination of local and global quality scores that are estimated in terms of several features such as orientation certainty level (OCL) and so on. Their quality metric also involves several thresholds to classify the local blocks into variant levels. Similarly, Chen *et al.* [36] proposed a metric by linearly combine the orientation flow (OF) and the ridge-valley clarity features. Apparently, the weighted coefficients have to be adjusted if a different image setting is involved.

The linear combination of multi-feature could also be illustrated by a regression-based approach [39] which adopts genetic algorithm (GA) optimizing (or maxiaizing) [45] the linear relationship between the quality value and the genuine matching scores of a set of training samples. Maximizing the correlation between the two measures [45] is a solution for qualifying biometric sample. However, the optimization largely depends on the genuine matching results. Likewise, this problem could also be considered for other quality metrics that are associated with a prior-knowledge of matching performance, for matching algorithms can be quite different.

Quality assessment approaches with multi-feature carried out in another form is classification. Lim *et al.* [40] extended their work in [35] by classifying a certain amount of fingerprint samples with 3 different classifiers rather than calculating the quality metric. Later, the state-of-the-art quality metric, NFIQ, employs 11-dimension feature to estimate a matching score and classify results to five levels through a trained model of a neural network [23]. Further, in NFIQ 2.0, Olsen *et al.* [41] trained a two-layer self-organizing map (SOM neural network) to obtain a histogram of SOM unit activation with an intensity vector of image block. The histogram is the frequency of the occurrence of the best-matching unit (with respect to the competitive layer) assigned to each block. The trained feature is then threw to a Random Forest (RF) to estimate the binned genuine matching scores (GMS). This is the first study of FQA to generate a learning-based feature by using unsupervised approach and a quite large dataset. However, the RF is to classify samples in terms of a prior-knowledge of matching score and quality is represented by the regularity as well. So far, no studies is able to conduct a perfect matching algorithm because the matching scores between two bad quality genuine or impostor samples are somehow unforeseeable [14].

According to such a statement, one can note that approaches with a single feature is limited to a specified image type and knowledge-based solutions is not absolutely appropriate to cross-use. Besides, it is also possible to consider whether a quality metric based-on multi-feature really makes a robust criterion or takes the advantages of them.

A. Discussion

In previous sections, we investigate some representative literature of FQA to illustrate the solutions that had been proposed so far. A common fact is that biometric quality of fingerprint sample is not completely the same as it is estimated by subjective criteria [20], [14]. The biometric definition should be related to the matching performance which is expected to be benefited from the qualified fingerprint samples. This problem could be simply illustrated by some examples, as given in Figure 2. Figure

Fig. 2: Example of fingerprint samples that are visually different. From left (S1) to right (S4): 73_2, 7, 5, 8 of FVC2002DB2A.

2 shows several genuine fingerprint samples that are visually different, among which we can simply determine the leftmost one is relatively clear and complete, followed by an image with a little bit translation, a fragment print and a scattered-looking image. According to some subjective assessment criteria [20], we believe the quality of them are different as well. There is no ground-truth of them so that we chose several metrics to generate quality values of them, as shown in Table II.

TABLE II: Quality values of the samples in 2.

QM \ Sample	S1	S2	S3	S4
STD	72	74	33	31
OCL	0.66	0.63	0.41	0.40
NFIQ	1	2	2	2

Apparently, sample S3 is only a partial image and sample S4 is not a clear image with relatively small foreground area. These two samples should have relatively bad quality values. NFIQ gives both a quality of level 2, indicating their qualities are better than the average level. As one cannot assert this is not reasonable without the ground-truth of them, we simply calculate the GMS by assigning every of them as the enrollment, results are given in Table III.

TABLE III: Genuine matching scores calculate by using Bozorth3.

Enroll \ Sample	S1	S2	S3	S4
S1 enroll		266	72	230
S2 enroll	263		95	231
S3 enroll	72	95		43
S4 enroll	228	230	43	

In table III, the S3 leads to relatively lower values when it is used as the enrollment. Although the matching is fully dependent on minutiae including minutiae number, but good quality samples should be suitable for matching [14], especially genuine matching.

III. Evaluation of Quality Metric

The validity of a biometric quality metric should be verified via its contribution to the matching performance. Hence, this paper also reviews some evaluation algorithms related to fingerprint quality metric, and most of them are generic approaches that are available for other modalities as well.

A. Review

Philips et al. [46] pointed out the significance of protocols for a biometric test such a system level evaluation. However, some earlier studies of evaluation approach do not provide totally open protocols such as the employed dataset [14]. In addition, some other studies directly related the evaluation of quality metric to the elements of subjective assessment [36], and some approaches proposed later choose to evaluate their metrics with the existing ones [30]. Similarly, Shen *et al.* [27] divided fingerprint images into several classes according to samples' quality types and compared their approach with another via computing the proportion of correctly classified samples. Some others computes a quality benchmark through the observation of automatically detected minutiae of fingerprints [35], [36]. These approaches failed to explicitly reflect the relation between quality metric and matching performance. In addition, these attempts are more or less related to subjective observations when validating quality metric, such as the manual classification of fingerprint quality types, differentiating spurious minutiae or missed minutiae from the templates.

Tabassi *et al.* [23] defined biometric sample quality as a predictor of matching performance, for they observed that biometric samples of good quality should produce relatively high genuine matching scores (GMS) which are well separated from impostor matching scores (IMS). However, the prediction is totally dependent on the matching performance. Chen *et al.* [44] proposed that the Equal Error Rate (EER) should be monotonically decreased after a certain part of bad samples had been pruned. In addition, they also considered that the detected minutiae of good quality samples should not greatly varied after enhancement. Another consideration of their study is that the detected minutiae of good quality samples should have relatively good consistency with manual groundtruth of minutiae. This is somehow a coherent principle when dealing with minutiae-based matching algorithms. Obviously, matching performance could be guaranteed if minutiae are precisely detected. This assumption is restricted by the underlying matching techniques, *i.e.* minutiae-based matching only. Figure 3 illustrates an example of Chen's approach obtained by using NFIQ and QMF [47] on one FVC database.

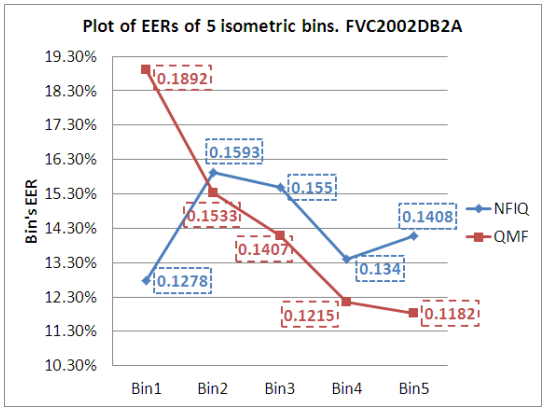


Fig. 3: Plots of 5 isometric bins' EER values, where blue points are obtained by NFIQ and red points correspond to QMF.

As illustrated in 3, this evaluation criterion requires an equivalency of sample number for each bin, otherwise the EER values of the bins could be affected. A valid quality metric is expected to generate monotonically decreasing (or increasing) EER values of the bins. In addition, it is necessary to consider whether this approach is also appropriate to a quality metric that represents biometric sample qualities with several labels, because the difference between samples that have the same label is unknown, and hence it is inappropriate to get samples of variant labels divided into the same bin, *i.e.* the EER values could be seriously contaminated by the outliers of each other. For example, the result of NFIQ in fig. 3 shows a disordered plot of the EER values of the bins. The details about 'isometric bins' could be found in the reference article [44].

Grother *et al.* [14] discussed that the quality measure of a biometric sample is generally employed within 3 different cases, including enrollment phase, verification and identification. They proposed several evaluation approaches associated to matching threshold and quality levels, including rank-ordered DET curve, error versus reject curve and the approach based on the Kolmogorov Smirnov (KS) test. As they are closely related to the quality level, some of them may not be completely suitable for quality metrics with a larger range values, such as KS test-based approach.

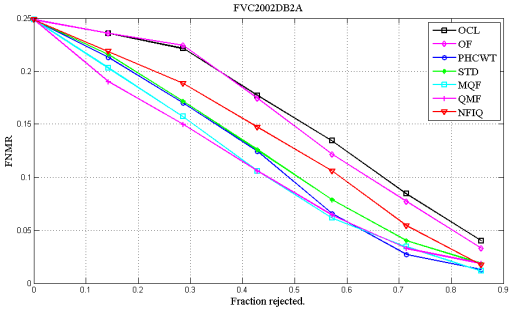


Fig. 4: An illustration of the error versus reject curve. Database is FVC2002DB2A.

Figure 4 shows another form of the error versus reject curve which is independent to the quality level. For example, one database with M individuals and N samples per individual, the verification operation could be simulated via assigning a specified sample (for example, the first one) of each individual as the enrollment, while other $N-1$ samples act as authentication samples. Each time, one authentication sample with the lowest quality is eliminated for each individual, *i.e.* a fixed percentage of low quality genuine samples are removed each time. The FNMR then can be calculated with a threshold which is similar to the definition in [14].

B. Evaluation based on Enrollment Selection

Generally, a system level quality control is performed after the capture session. This process would be executed as a loop to acquire a qualified sample measured by quality assessment module. The re-capture at the same session might result in FTE and the Failure to Acquire (FTA) caused either by algorithm crash or sensor overtime. However, to validate a quality metric, it is avoidable to consider this problem involved in the capture and enrollment sessions as it had just been mentioned before. Without emphasizing the test type, the enrollment can be a supervised process. This study systematically presents an extension of the algorithm proposed in [25], which eliminates the discrepancy caused by the definition of FMR or FAR (another is FNMR or FRR), for there is no error rate like failure to enroll (FTE) in the experiment. The details are presented as the follows.

1) Algorithm Description

Assuming a database of biometric sample contains N samples of each of M individuals, *i.e.* size of $M \times N$. With an employed matcher, one can calculate an objective measure for each sample, which is represented by sample EER value. The sample EER is regarded as an approximation of the ground-truth of the sample within the involved matcher, which we also denote as sample utility. The sample EER, $SEER_{i,j}$, is obtained by specifying the j^{th} sample of the i^{th} individual as the enrollment, while other $N-1$ samples of this individual are viewed as authentication samples. In this way, given a matcher R, a total of $N-1$ genuine matching scores (GMS)

$$gms_{i,j,k} = R\left(S_{i,j}, S_{i,k}\right), \ j \neq k \tag{1}$$

and $N-1 \times M-1$ impostor matching scores (IMS)

$$ims_{i,j,l,k} = R\left(S_{i,j}, S_{l,k}\right), \ i \neq l \ and \ j \neq k \tag{2}$$

are computed for obtaining the $SEER_{i,j}$ of sample $S_{i,j}$. The $SEER_{i,j}$ is computed as the point where $FNMR_{i,j}(t) = FMR_{i,j}(t)$. The $FNMR_{i,j}(t)$ and $FMR_{i,j}(t)$ are computed from a given set of threshold t:

$$\begin{aligned} FNMR_{i,j}(t) &= \frac{card\{gms_{i,j,k} | gms_{i,j,k} \geq t\}}{N-1}, \\ FMR_{i,j}(t) &= \frac{card\{ims_{i,j,l,k} | ims_{i,j,l,k} \leq t\}}{(N-1) \times (M-1)} \end{aligned} \tag{3}$$

where $card$ denote the cardinality of a given set, $k \neq j \ and \ k \leq N-1$ and $l \neq i \ and \ l \leq M-1$.

This calculation results in a total of M-by-N sample EER values for one database, by which one can perform enrollment selection (ES) in two cases. The first case of ES is carried out by choosing the best sample of each individual as the enrollment. The 'best' here means that the $SEER$ of the selected sample is the smallest (best) among all the samples of one individual. This is defined by the consideration that one matcher cannot obtain better performance than this case from a given dataset. Likewise, one can achieve another ES in the opposite way which indicates the worst case. In this paper, the two cases are denoted as the best utility and worst utility, respectively. An illustration of the selection framework is given in Figure 5.

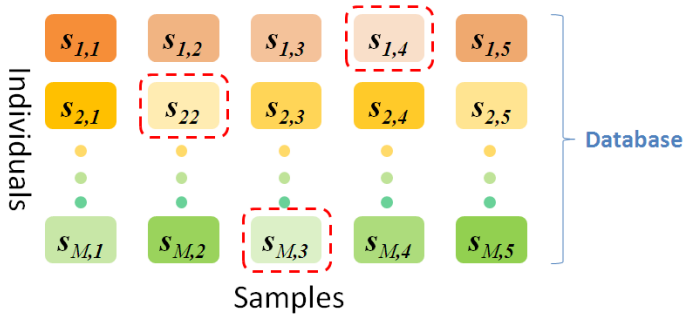


Fig. 5: Illustration of enrollment selection. Pale color represents small value.

The enrollment sample of each individual is determined by doing so. Correspondingly, one can calculate again the intra-class and inter-class matching scores for the whole dataset by the assigned enrollments. Further, one can calculate a global EER value from the new matching scores. The global EER value is simply an indication of the best (or worst) case that a matcher could obtain from the given dataset. In this framework, the global EER value is defined as below:

1 A set of 500 samples are randomly chosen from the $M \times (N-1)$ intra-class matching scores which are calculated after enrollment selection.
2 Similarly, another set of 500 samples are randomly chosen from the $M \times (M-1) \times (N-1)$ inter-class matching scores.
3 A global EER sample could be calculated from the two sets of randomly selected matching scores in a similar way to equation 3.
4 Step 1 to 3 are performed for 1000 iterations, and an average of 1000 global EER samples are preserved at last.

Note that the size of matching score samples could be changed according to the dataset dimension. The changing is to ensure the diversity of the matching scores, especially for impostor matching scores.

Finally, one can obtain the quality values $q_{i,j}$ of one dataset by using a given quality metric. Likewise, according to sample quality, another enrollment selection could be done by choosing the best sample of each individual as the enrollment. Then, by performing steps 1 to 4, a global EER value also could be figured out, which indicates how much the quality metric contributes to reducing the error rate. For a relatively good metric, one can consider that the overall performance (global EER) should be much close to the best case when choosing samples of good quality during the enrollment, and a

valid quality metric would satisfy this condition as much as possible. This property will be illustrated in the experimental section.

In addition to the global EER value, one can use other global measure such as the Area Under Curve (AUC) value. In this paper, we also define a global index based on three AUC values obtained by the best utility, worst utility and the best quality, respectively. This global index is namely AUC ratio, formulated as

$$r_{auc} = \frac{Q_{auc} - W_{auc}}{B_{auc} - W_{auc}}, \tag{4}$$

where Q_{auc} is the global AUC value of the ROC curve computed in terms of the best quality samples as it had just been mentioned, B_{auc} and W_{auc} are the global AUC values correspond to the ROC curves of the best and worst utilities. This definition is to say that the ROC curve obtained by a relatively good quality metric should be much closer to the ROC curve of the best case, see Figure 6.

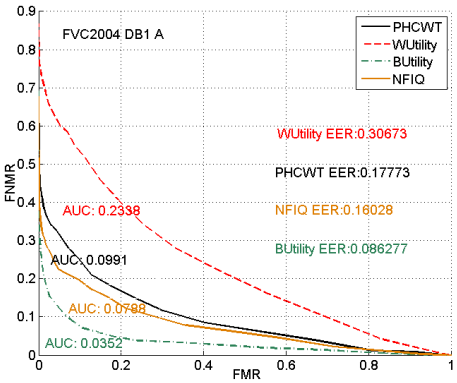


Fig. 6: Illustration of the enrollment selection result via ROC curves, EER and AUC values. The 'BUtility' and 'WUtility' represent the best case and the worst case, respectively.

Figure 6 illustrate how AUC ratio is used in evaluating quality metrics. The quality metrics and the dataset employed are NFIQ, PHCWT and FVC2004DB1A. Obviously, the larger the value of r_{auc}, the better the performance of quality metric is.

At last, an ideal case or a perfect quality metric is expected to be able to generate N monotonically increasing or decreasing global EER values based on the enrollment selection approach. This is done by performing ES for N times on one dataset, *i.e.* the ES is performed in an ascending or descending order of sample qualities of each individual. This is harsher than the 5-bins evaluation strategy, for it not only depends on the quality assessment approach but also relies on the matching algorithm.

2) Confidence Interval (CI)

According to the definition above, we can quantify the performance of a quality metric with several global EER values. However, a further validation needs to be acquired for making a significant comparison between variant quality metrics, especially for comparing them with a certain kind of measurements. To do so, this study calculates a CI [48], [49] at 95% level for the global EER value based on the enrollment selection of quality (the last case in the definition). In biometrics, the bootstrap CIs of two different measurements are able to indicate the statistical difference between them, if their CIs do not overlap each other [50]. By doing so, the difference between quality metrics could be determined statistically. In Section III-B2, the global EER value is an average of 1000 global EER samples. The CI of the global EER value is also calculated with these 1000 samples, formulated as

$$CI = [\overline{X} - \frac{\sigma}{\sqrt{1000}}\mu_{\alpha/2},\ \overline{X} + \frac{\sigma}{\sqrt{1000}}\mu_{\alpha/2}], \tag{5}$$

where X represents the global EER samples, $\alpha = 100\% - 95\%$, $\overline{X}$ is the average of X, σ is the standard deviation of X and $\mu_{\alpha/2}$ is the $\alpha/2$ quantile.

Likewise, one can calculate the CIs of the global AUC and AUC ratio as well. By doing so, the validity of a quality metric could be determined statistically.

IV. Experimental Results

In this section, we simply present a comparison between different quality metrics to exhibit the generality of the evaluation scheme and how it works. One evaluation algorithm, by specifying an evaluation protocol, should be repeatable and feasible to different kinds of quality metrics and is able to demonstrate metrics' contribution to the matching performance in terms of utility. In the experiment, the quality metrics are given anonymously, simply as some existing studies conducted [14].

A. Test Protocol

According to the evaluation approach given in Section III, the experiment is performed by using several quality metrics of fingerprint image on one database.

1) Database

In the experiment, we combined 6 FVC databases, including 00DB2A, 02DB2A and four databases of FVC2004 Set A. The resolution of these datasets are over 500-dpi and the image sizes are different. The combined dataset hence has 600 individuals where 8 sample per individual, 4800 images in total. Figure 7 gives a glance to the samples of the combined dataset.

Fig. 7: Illustration of samples of the trial datasets.

2) Software

First, we use MINDTCT and Bozorth3 [13] of the NIST for generating minutiae templates and matching scores. The minutiae template of the MINDTCT is recorded as $m_i = \{x, y, o, q\}$, where (x, y) is the location of minutia point, o indicates orientation and q is a quality score of minutia point. The Bozorth3 performs matching between templates and generates an integral matching score. In addition, we also used a minutiae extractor and a matching algorithm provided by a commercial SDK (namely ID3Finger toolkit, Version 2.3.5647) [51]. The SDK creates ISO/IEC-19794:2 [52] standard template for our experiments. Correspondingly, M and N in this experiment are 600 and 8, respectively. The experiments involve 600×7 genuine matching operations and $600\times599\times7$ impostor matchings for each sample of one individual.

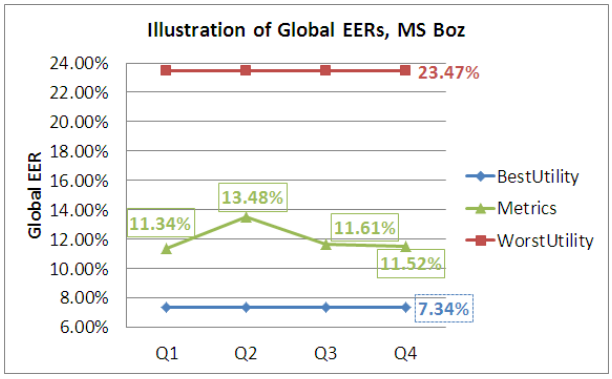
In next section, we simply present a demonstration of the evaluation approach by using four fingerprint quality metrics, anonymously denoted as Q1, Q2, Q3 and Q4. The first 3 metrics generate continuous quality values in the range of [0, 100] after normalization, which denotes an ascending order of sample qualities. The Q4 represents sample quality via discrete labels from 1 to 5, where low value denotes good quality. Results are given as the follows.

B. Results

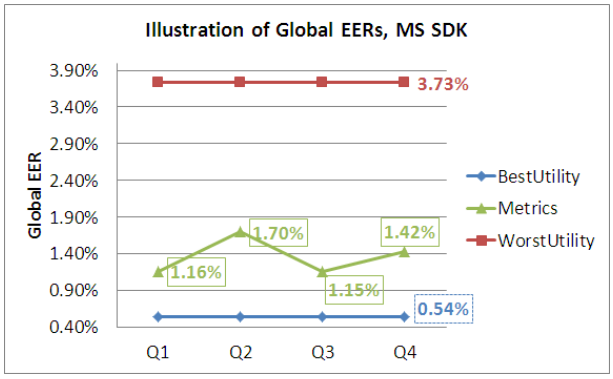
We firstly demonstrate the global EER values obtained from the trial dataset by using the ES with both quality metrics and sample utility, as illustrated together via plots in Figure 8.

In Figure 8, the points plotted in blue and red are global EERs obtained by using the two cases of the utility value generated from the associated matching scores, where in fig. 8a the utility value of samples are calculated from the matching scores of the Bozorth3 and the global EERs are also figured out with the same matching scores. Similarly, 8b demonstrates results based on the matching scores generated via the SDK, where the blue points correspond to the best case and the red points represent the worst case. The global EER values obtained by each quality metric with matching scores of Bozorth3 are 11.34%, 13.48%, 11.61% and 11.52%. Another group of the global EERs based on the matching scores of the SDK are 1.16%, 1.70%, 1.15% and 1.42%, respecitvely. Next, the 95% CI of each global EER is calculated, as given in Table IV.

In Table IV, the CI of the global EER values based on the employed quality metrics statistically differentiates one from the others. With this observation, one can have an objective evaluation of quality metrics and achieve a comparison between them. Meanwhile, one can notice that the evaluation results of Q4 obtained by two different groups of matching scores are



(a) Result based on MS Boz



(b) Result based on MS SDK

Fig. 8: Plots of global EERs obtained by using ES with 4 quality metrics and sample utility. Fig. 8a is the result based on matching scores of the Bozorth3 and 8b is the result calculated with matching scores of the SDK.

quite different. This might be attributed to two reasons: one is the interoperability issue [14] and another is the quality metric itself. The designed of Q4 is tied to a prior-knowledge of matching scores. We simply use this observation to illustrate a problem that quality is not absolutely to predict the overall performance unless it is associated to a particular matcher. This problem will be further revealed in the final experiment.

TABLE IV: The 95% confidence interval of the global EER of each quality metric.

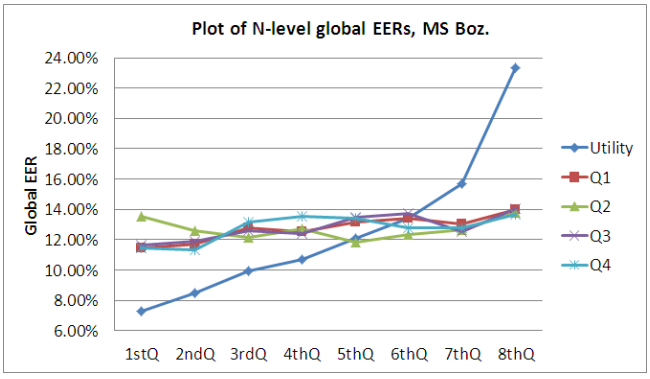
QM \ Matcher	MS Boz	MS SDK
Q1	[0.1126 0.1143]	[0.0113 0.0118]
Q2	[0.1341 0.1356]	[0.0167 0.0173]
Q3	[0.1154 0.1169]	[0.0112 0.0118]
Q4	[0.1144 0.1160]	[0.0139 0.0145]

Secondly, the values of AUC ratio of the employed quality metrics are calculated to demonstrate the evaluation framework. The AUC values involved in calculating the AUC ratio are also average values of 1000 AUCs obtained during the computation of 1000 global EER samples. Results are given in Table V.

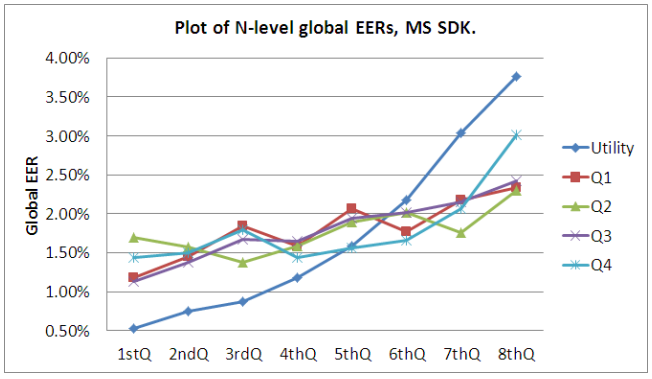
TABLE V: AUC Ratio.

Matcher \ QM	Q1	Q2	Q3	Q4
MS Boz	0.7843	0.6538	0.7830	0.7877
MS SDK	0.8333	0.7133	0.8400	0.7867

Finally, examples of N global EERs generated by orderly performed ES are calculated for each of the trial metrics. The graphical results are given in Figure 9. In each figure, the plots indicated by 'utility' are global EERs obtained by using



(a) Results based on MS Boz



(b) Results based on MS SDK

Fig. 9: 8-level global EER values based on enrollment selection. 9a is the result obtained from matching scores of Bozorth3 and 9b is calculated by using matching scores of SDK.

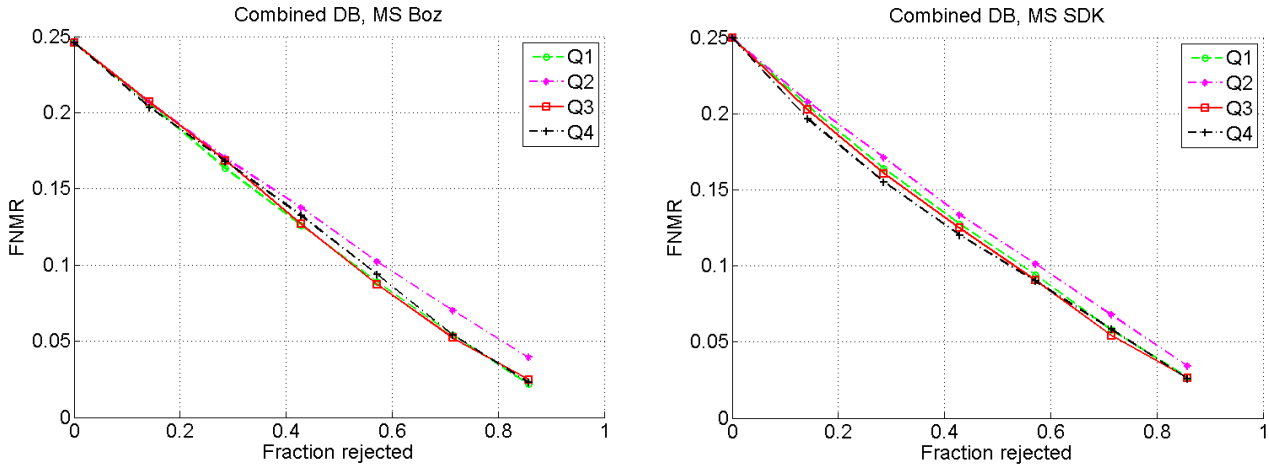
the utility values, while other groups of plots are results of the quality metrics. The utility-based result is simply given as a reference. The first level ($1^{th}Q$) of each plot represents the case that the best quality sample of each user is used as enrollment, followed by the second and so on.

With such plots, one can also observe the variation of the evaluation results of each metric obtained from different matching scores. For instance, the 4^{th} EER values of both Q1 and Q4 in fig. 9b show an typical problem that outliers is unavoidable for most of the existing metrics. However, it is almost a different case in fig. 9a for these two metrics. This kind of results clearly reveal that the effect of matching algorithm to the utility-based evaluation. Some evaluation approaches measure a FNMR by considering only the GMS of samples, which neglect the effect of IMS to the performance estimation of quality metrics because it is relatively easier for a metric to be correlated with the GMS [26]. Although the real applications chiefly focus on genuine match, we recommend to perform metric evaluation with the overall performance. In addition, we provide this enrollment selection approach with a Matlab script[1].

C. Discussion

In this section, we discuss briefly what should be considered in a biometric evaluation, even if it is an evaluation of a biometric quality metric. According to Section III-A, one can note that a technology-level [46] evaluation is generally achieved in terms of two factors, genuine matching error and the overall error rate which considers both the genuine and impostor errors. In real applications, biometric authentication mainly focuses on genuine matching but a biometric test is to measure the performance of either a prototype system or an algorithm. Such a test involves both the genuine and impostor matching. In this case, one can question that whether a genuine error rate can give a clear evaluation of the target system or algorithm. For instance, there is no evidence proving that samples of good quality justified by a quality metric cannot lead to false-match errors [14], even if the quality metric performs quite well in terms of the genuine matching errors. In this case, it is necessary to consider the overall error rate when estimating the performance of a biometric quality metric because it is essentially a biometric test. On the other hand, a unilateral evaluation can be subjected by other impacts.

The evaluation result of the 'error vs reject' approach [26], [14], for example, relies first on specified threshold value(s). Nevertheless, the FNMR is related to an orderly sorting in terms of quality. The applicability of quality to other approach is unknown [14] in this case. Similarly, the evaluation based on the equal-size bins [44] of samples needs also a similar sorting. Such a problem has been discussed in Section III-A. The proposed evaluation approach in this paper considers only the quality of the enrollment sample which greatly decreases the effect of sorting to the evaluation result. Because of this, we use the 'err vs reject' approach defined in Section III-A to the employed metrics. The graphical results are given in Figure 10.



(a) Results based on MS Boz

(b) Results based on MS SDK

Fig. 10: Error versus reject results of the employed metrics. Figure 10a and 10b are results obtained by using the matching scores of the two matchers, respective. x-aixs: fraction rejected; y-axis: FNMR.

Apparently, according to Figure 10 (both 10a and 10b), the results are not sufficiently clear unless metrics are distinctively different from each other, such as the difference between Q2 and and other three metrics illustrated by Figure 10b. Meanwhile, one can note that the threshold of the FNMR can lead to different evaluation result once it is changed.

V. Conclusion

This paper makes a literature review for most of the existing studies of FQA to demonstrate the representative solutions that had been proposed so far. Secondly, with a short review of performance evaluation approaches of fingerprint quality metric,

[1]Matlab script URL: http://www.mathworks.com/matlabcentral/fileexchange/48132-biometric-modality-quality-metric-validation-evaluation

we propose an extension of a generic evaluation framework in terms of utility. The evaluation framework is able to provide a statistic measure to demonstrate how much the quality metric contributes to the improvement of the overall performance. By making efforts in these two aspects, we noted that several questions should be answered or considered in a further study: 1) Are those fingerprint quality metrics based-on multi-feature really able to make the fused metric complementary? and 2) To achieve a common solution, it is necessary to consider whether learning a prior-knowledge of matching performance such as GMS is reasonable or not? This is not to claim that quality is not predictive to the matching performance but one should note this limitation [29] as existing matching approaches are not perfect or robust to all image settings, even though image resolutions are quite close to each other. In addition, according to the literature [14], it is agnostic that whether two samples produce low impostor score when they are of low quality. Similarly, in Section II-A, it is dubious as well for the genuine matching score between two genuine samples if one of them has an unexpected quality.

To the end, we recommend as well to perform evaluation via the overall performance, for this is more objective to a biometric test. Besides, without considering the modeling approach, we recommend that a generic quality metric should be independent from the prior-knowledge of matching performance unless it is applied to one or several specific scenario(s) or the training dataset and the employed knowledge are sufficient and reliable enough. In addition, we recommend to perform off-line quality evaluation via relative difficult dataset such as the CASIA test database because the effect of quality to the verall performance could be blurred if the matcher is relatively robust, *i.e.* it is relatively easier to achieve a lower EER.

Our next study will focus on a comparative study between quality metrics from different frameworks, because some metrics based on a single feature can outperform those with multiple features. Meanwhile, perceptively, we can consider that the quality could be measured in terms of the quality desired by a system rather than estimating samples as good ones or not.

REFERENCES

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2d and 3d face recognition: A survey," vol. 28, no. 14. Elsevier, 2007, pp. 1885–1906.
[2] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*. springer, 2009.
[3] N. K. Ratha, K. Karu, S. Chen, and A. K. Jain, "A real-time matching system for large fingerprint databases," vol. 18, no. 8. IEEE, 1996, pp. 799–813.
[4] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 1, pp. 4–20, 2004.
[5] J. Wegstein, *A semi-automated single fingerprint identification system*. US Department of Commerce, National Bureau of Standards, 1969.
[6] A. Jain, L. Hong, and S. Pankanti, "Biometric identification," vol. 43, no. 2. ACM, 2000, pp. 90–98.
[7] D. H. McMahon, G. L. Johnson, S. L. Teeter, and C. G. Whitney, "A hybrid optical computer processing technique for fingerprint identification," vol. 24, no. 4. IEEE, 1975, pp. 358–369.
[8] N. K. Ratha, S. Chen, and A. K. Jain, "Adaptive flow orientation-based feature extraction in fingerprint images," vol. 28, no. 11. Elsevier, 1995, pp. 1657–1672.
[9] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Quality measures in biometric systems," vol. 10, no. 6. IEEE, 2012, pp. 52–62.
[10] L. Coetzee and E. C. Botha, "Fingerprint recognition in low quality images," vol. 26, no. 10. Elsevier, 1993, pp. 1441–1460.
[11] K. Karu and A. K. Jain, "Fingerprint classification," vol. 29, no. 3. Elsevier, 1996, pp. 389–404.
[12] L. Hong, Y. Wan, and A. Jain, "Fingerprint image enhancement: algorithm and performance evaluation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 777–789, 1998.
[13] C. I. Watson, M. D. Garris, E. Tabassi, C. L. Wilson, R. M. Mccabe, S. Janet, and K. Ko, "User's guide to nist biometric image software (nbis)," 2007.
[14] P. Grother and E. Tabassi, "Performance of biometric quality measures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 531–543, 2007.
[15] A. J. Willis and L. Myers, "A cost-effective fingerprint recognition system for use with low-quality prints and damaged fingertips," vol. 34, no. 2. Elsevier, 2001, pp. 255–270.
[16] K. Ito, A. Morita, T. Aoki, T. Higuchi, H. Nakajima, and K. Kobayashi, "A fingerprint recognition algorithm using phase-based image matching for low-quality fingerprints," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2. IEEE, 2005, pp. II–33.
[17] D. Zhang, F. Liu, Q. Zhao, G. Lu, and N. Luo, "Selecting a reference high resolution for fingerprint recognition using minutiae and pores," vol. 60, no. 3. IEEE, 2011, pp. 863–871.
[18] Q. Zhao, F. Liu, and D. Zhang, "A comparative study on quality assessment of high resolution fingerprint images," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 3089–3092.
[19] R. M. Bolle, S. U. Pankanti, and Y.-S. Yao, "System and method for determining the quality of fingerprint images," Oct. 5 1999, uS Patent 5,963,656.
[20] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Kernel-based multimodal biometric verification using quality signals," in *Defense and Security*. International Society for Optics and Photonics, 2004, pp. 544–554.
[21] I. 29794-1:2009, "Information technology – biometric sample quality – part 1: Framework," October 2009.
[22] A. K. Jain and S. Z. Li, *Encyclopedia of Biometrics: I-Z*. Springer, 2009, vol. 1.
[23] E. Tabassi, C. Wilson, and C. Watson, "Nist fingerprint image quality," *NIST Res. Rep. NISTIR7151*, 2004.
[24] G. Li, B. Yang, and C. Busch, "Autocorrelation and dct based quality metrics for fingerprint samples generated by smartphones," in *Digital Signal Processing (DSP), 2013 18th International Conference on*. IEEE, 2013, pp. 1–5.
[25] Z. YAO, C. Charrier, and C. Rosenberger, "Utility validation of a new fingerprint quality metric," in *International Biometric Performance Conference 2014*. National Insititure of Standard and Technology (NIST), April 2014.
[26] F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, J. Gonzalez-Rodriguez, H. Fronthaler, K. Kollreider, and J. Bigun, "A comparative study of fingerprint image-quality estimation methods," *Information Forensics and Security, IEEE Transactions on*, vol. 2, no. 4, pp. 734–743, 2007.
[27] L. Shen, A. Kot, and W. Koo, "Quality measures of fingerprint images," in *IN: PROC. AVBPA, SPRINGER LNCS-2091*, 2001, pp. 266–271.
[28] B. Lee, J. Moon, and H. Kim, "A novel measure of fingerprint image quality using the Fourier spectrum," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, A. K. Jain and N. K. Ratha, Eds., vol. 5779, Mar. 2005, pp. 105–112.
[29] Z. Yao, J. Le Bars, C. Charrier, and C. Rosenberger, "Quality assessment of fingerprints with minutiae delaunay triangulation," in *International Conference on Information Systems Security and Privacy (ICISSP)*, Feb 2015.
[30] N. K. Ratha and R. Bolle, *Fingerprint image quality estimation*. IBM TJ Watson Research Center, 1999.
[31] H. Fronthaler, K. Kollreider, and J. Bigun, "Automatic image quality assessment with application in biometrics," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 30–30.

[32] L. Nanni and A. Lumini, "A hybrid wavelet-based fingerprint matcher," vol. 40, no. 11. Elsevier, 2007, pp. 3146–3151.

[33] J. D. Humphreys, G. Porter, and M. Bell, "The quantification of fingerprint quality using a relative contrast index," vol. 178, no. 1. Elsevier, 2008, pp. 46–53.

[34] S. Lee, H. Choi, K. Choi, and J. Kim, "Fingerprint-quality index using gradient components," vol. 3, no. 4. IEEE, 2008, pp. 792–800.

[35] E. Lim, X. Jiang, and W. Yau, "Fingerprint quality and validity analysis," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, 2002, pp. I–469–I–472 vol.1.

[36] T. Chen, X. Jiang, and W. Yau, "Fingerprint image quality analysis," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, vol. 2, 2004, pp. 1253–1256 Vol.2.

[37] J. Qi, D. Abdurrachim, D. Li, and H. Kunieda, "A hybrid method for fingerprint image quality calculation," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*. IEEE, 2005, pp. 124–129.

[38] X. Tao, X. Yang, Y. Zang, X. Jia, and J. Tian, "A novel measure of fingerprint image quality using principal component analysis(pca)," in *Biometrics (ICB), 2012 5th IAPR International Conference on*, March 2012, pp. 170–175.

[39] M. El Abed, A. Ninassi, C. Charrier, and C. Rosenberger, "Fingerprint quality assessment using a no-reference image quality metric," in *European Signal Processing Conference (EUSIPCO)*, 2013, p. 6.

[40] E. Lim, K.-A. Toh, P. Suganthan, X. Jiang, and W.-Y. Yau, "Fingerprint image quality analysis," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 2. IEEE, 2004, pp. 1241–1244.

[41] M. Olsen, E. Tabassi, A. Makarov, and C. Busch, "Self-organizing maps for fingerprint image quality assessment," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 138–145.

[42] A. M. Bazen and S. H. Gerez, "Segmentation of fingerprint images," in *Proc. Workshop on Circuits Systems and Signal Processing (ProRISC 2001)*. Citeseer, 2001, pp. 276–280.

[43] M. Ballan, F. A. Sakarya, and B. L. Evans, "A fingerprint classification technique using directional images," in *Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*, vol. 1. IEEE, 1997, pp. 101–104.

[44] Y. Chen, S. C. Dass, and A. K. Jain, "Fingerprint quality indices for predicting authentication performance," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2005, pp. 160–170.

[45] R.-L. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the*. IEEE, 2006, pp. 1–6.

[46] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction evaluating biometric systems," vol. 33, no. 2. IEEE, 2000, pp. 56–63.

[47] Z. Yao, J. M. Le Bars, C. Charrier, and C. Rosenberger, "Fingerprint quality assessment combining blind image quality, texture and minutiae features," in *International Conference on Information Systems Security and Privacy*, 2015.

[48] R. Giot, M. El-Abed, and C. Rosenberger, "Fast computation of the performance evaluation of biometric systems: Application to multibiometrics," vol. 29, no. 3. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., Mar. 2013, pp. 788–799. [Online]. Available: http://dx.doi.org/10.1016/j.future.2012.02.003

[49] R. M. Bolle, N. K. Ratha, and S. Pankanti, "Error analysis of pattern recognition systemsthe subsets bootstrap," vol. 93, no. 1. Elsevier, 2004, pp. 1–33.

[50] T. Dunstone and N. Yager, *Biometric system and data analysis: Design, evaluation, and data mining*. springer, 2008.

[51] User friendly toolkit for easy integration of state of the art fingerprint recognition technology. ID3 Technologies. [Online]. Available: http://www.id3.eu

[52] ISO/IEC. (2005, 09) Information technology biometric data interchange formats part 2: Finger minutiae data. ISO/IEC.