Performance of Biometric Quality Measures

Patrick Grother, *Member*, *IEEE*, and Elham Tabassi, *Member*, *IEEE*

Abstract—We document methods for the quantitative evaluation of systems that produce a scalar summary of a biometric sample's quality. We are motivated by a need to test claims that quality measures are predictive of matching performance. We regard a quality measurement algorithm as a black box that converts an input sample to an output scalar. We evaluate it by quantifying the association between those values and observed matching results. We advance detection error trade-off and error versus reject characteristics as metrics for the comparative evaluation of sample quality measurement algorithms. We proceed this with a definition of sample quality, a description of the operational use of quality measures. We emphasize the performance goal by including a procedure for annotating the samples of a reference corpus with quality values derived from empirical recognition scores.

Index Terms—Biometrics, quality measurement, authentication, evaluation, performance measures.

1 Background

Quality measurement algorithms are increasingly deployed in operational biometric systems [1], [2] and there is now international consensus in industry [3], academia [4], and government [5] that a statement of a biometric sample's quality should be related to its recognition performance. That is, a quality measurement algorithm takes a signal or image, $\mathbf{x}$, and produces a scalar, $q = Q(\mathbf{x})$, which is predictive of error rates associated with the verification or identification of that sample. This paper formalizes this concept and advances methods to quantify whether a quality measurement algorithm (QMA) is actually effective.

What is meant by quality? Broadly, a sample should be of good quality if it is suitable for automated matching. This viewpoint may be distinct from the human conception of quality. If, for example, an observer sees a fingerprint with clear ridges, low noise, and good contrast, then he might reasonably say it is of good quality. However, if the image contains few minutiae, then a minutiae-based matcher would underperform. Likewise, if a human judges a face image to be sharp, but a face recognition algorithm benefits from a slight blurring of the image, then the human statement of quality is inappropriate. Thus, the term quality is not used here to refer to the fidelity of the sample, but instead to the utility of the sample to an automated system. The assertion that performance is ultimately the most relevant goal of a biometric system implies that a c. For fingerprint minutiae algorithms, this could be the ease with which minutiae are detected. For face algorithms, it might include how readily the eyes are located.

Prior work on quality evaluation, and of sample quality analysis generally, is limited. Quality measurement naturally lags recognition algorithm development, but has emerged as it is realized that biometric systems fail on certain pathological samples. The primary use of a quality measure is as a means of detecting a bad sample and initiating recapture of the live subject. "Bad" in this context refers to any property or defect associated with a sample that would cause performance degradation.

This paper proposes testing quality measurement algorithms in large scale offline trials which offer repeatable, statistically robust means of evaluating core algorithmic capability. Alonso-Fernandez et al. [6] reviewed five algorithms and used the fingerprints of the multimodal MCYT corpus [7] to compare the distributions of the algorithms' quality assignments with the result that most of the algorithms behave similarly. We note that finer grained aspects of sample quality can be addressed. For instance, Lim et al. [8] trained a fingerprint quality system to predict the accuracy of minutia detection. However, such methods rely on the manual annotation of a data set and this is impractical for all but small data sets, not least because human examiners will disagree in this respect. The virtue of relating quality to performance is that matching trials can be automated and conducted in bulk. We note further that quality algorithms that relate to human perception of a sample quantify performance only as much as the sensitivities of the human visual system are the same as those of a biometric matcher. One further point is that performance related quality evaluation is agnostic on the underlying technology: It would be improper to force a fingerprint quality algorithm to produce low quality values for an image with few minutia when the target matching algorithm is non-minutia-based, as is the case for pattern-based methods [9].

We formalize the concept of sample quality as a scalar quantity that is related monotonically to the performance of biometric matchers, under the constraint that at least two samples with their own qualities (as opposed to a pairwise quality) are being compared. We do this in the context of enrollment, verification, and identification use-cases. We consider the common and useful case of a quality measure tuned to predict performance of one matcher and the more

- *P. Grother is with the Image Group, Information Access Division, Information Technology Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Bldg 225, Room A203, MS 8940, Gaithersburg, MD 20899. E-mail: pgrother@nist.gov.*
- *E. Tabassi is with the Image Group, Information Access Division, Information Technology Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, MS 8940, Bldg 225, Room A207, Gaithersburg, MD 20899. E-mail: tabassi@nist.gov.*

Manuscript received 31 Jan. 2006; revised 22 June 2006; accepted 28 Aug. 2006; published online 13 Feb. 2007.
Recommended for acceptance by S. Prabhakar, J. Kittler, D. Maltoni, L. O'Gorman, and T. Tan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org and reference IEEECS Log Number TPAMISI-0075-0106.
Digital Object Identifier no. 10.1109/TPAMI.2007.1019.

difficult case of one that generalizes to other matchers or classes of matchers.

In Section 2, we consider how sample quality is actually used and this establishes context for the desirable properties of a quality measure that we present in Section 3. This precedes the main contribution on evaluation in Section 4, which discusses the appropriateness of various performance measures as prediction targets for a quality algorithm and then as metrics themselves. In Section 5, we discuss what data should be used for testing a quality algorithm and document a procedure to construct a reference target database. Conclusions follow in Section 6.

The evaluation protocols proposed assume only that the quality algorithm is claimed to predict performance: We do not assume that the algorithm has been standardized nor that its output has any particular distribution. We test the claim by relating quality values to empirical matching results. However, we consider the algorithm to be a black box whose design and intended outputs are determined solely by its author and we make no assumption of its internal operation.

2 Uses of Biometric Quality Values

This section describes the roles of a sample quality measure in the various contexts of biometric operations. The quality value here is simply a scalar summary of a sample that is taken to be some indicator of matchability.

2.1 Enrollment Phase Quality Assessment

Enrollment is usually a supervised process and it is common to improve the quality of the final stored sample by acquiring as many samples as are needed to satisfy either an automatic quality measurement algorithm (the subject of this paper), a human inspector (a kind of quality algorithm), or a matching criterion (by comparison with a second sample acquired during the same session). Our focus on automated systems' needs is warranted regardless of analyses of these other methods, but we do contend that naive human judgment will only be as predictive of a matcher's performance as the human visual system is similar to the matching system's internals and it is not evident that human and computer matching are functionally comparable. Specifically, human inspectors may underestimate performance on overtly marginal samples. Certainly, human inspectors' judgment may be improved if adequate training on the failure modes and sensitivities of the matcher is given to the inspector, but this is often prohibitively expensive or time consuming and not scalable. Immediate matching also might not be predictive of performance over time because same-session samples usually produce unrealistically high match scores. For instance, Fig. 1 shows an example of two same-session fingerprint images that were matched successfully by three commercial vendors despite their obvious poor quality. That said, this paper does not take a position on the merits of doing this. Instead, we answer the question, if a quality apparatus is used, is it actually performing?

In any case, by viewing sample acquisition as a measurement and control problem in which the control loop is closed on the quality measure, a system gains a powerful means of improving overall sample quality. We demonstrate this in Section 4.2.

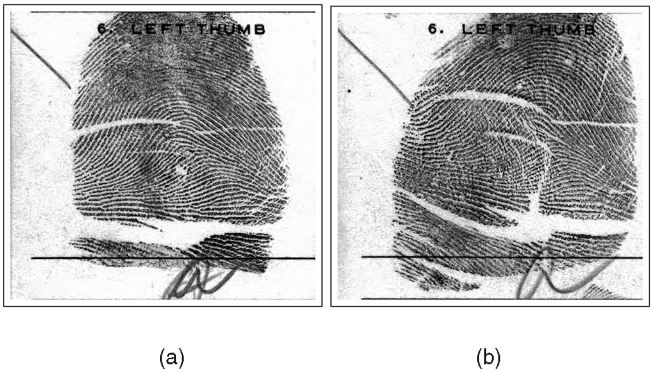


Fig. 1. Example of same session captures of a single finger that, despite their poor quality $(\mathrm{NFIQ} = 5)$, were matched correctly by three leading commercial matchers. (a) First. (b) Second.

2.2 Quality Assurance

QMAs may be used to monitor quality across multiple sites or over time. This is useful to signal possible performance problems ahead of some subsequent matching operation. Quality values may be aggregated and compared with some historical or geographic baselines. Use of quality values in this role has been documented in [1].

2.3 Verification Quality Assessment

During a verification transaction, quality can be improved by closing an acquire-reacquire loop on either a match-score from comparison of new and enrollment samples or on a quality value generated without matching. Indeed, it is common to implement an "up to three attempts" policy in which a positive match is a de facto statement that the sample was of good quality—even if the individual happens to be an impostor. Depending on the relative computational expenses of sample matching, reacquisition, and quality measurement, the immediate use of a matcher may not be the best solution.

The key difference here (as compared to the enrollment-phase) is that quality values of both the enrollment and verification samples can be used to predict performance. This two-dimensional problem is distinct from the enrollment case where only one quality value is used.

2.4 Identification Quality Assessment

Quality measurement in identification systems is important for at least three reasons. First, many users often do not have an associated enrollment sample. So, a one-to-many match will be an inefficient and inconclusive method of stating whether the authentication sample had high quality. Second, in negative identification systems where users with an enrolled sample are motivated to evade detection, quality measurement can be used to detect and prevent submission of samples likely to perform poorly [10], which may help prevent attempts at spoofing or defeating detection. Third, identification is a difficult task: It is imperative to minimize both the false nonmatch rate (FNMR) *and* the false match rate (FMR). To the extent that consistently high quality samples will produce high genuine scores, a high matching threshold can be used and this will collaterally reduce FMR. But, in large populations, FMR becomes dominant and this raises the question: Can a quality apparatus be trained to be directly predictive of false match likelihood? The authors can find no publications in this area.

2.5 Differential Processing

Quality measurement algorithms can be used to alter the subsequent processing of a sample. Such conditional activities are categorized as follows:

1. *Preprocessing Phase*. An identification system might apply image restoration algorithms or invoke different feature extraction algorithms for samples with some discernible quality problem.
2. *Matching Phase*. Certain systems may invoke a slower but more powerful matching algorithm when low-quality samples are compared.
3. *Decision Phase*. The logic that renders acceptance or rejection decisions may depend on the measured quality of the original samples. This might involve changing a verification system's operating threshold for poor quality samples. For example, in multimodal biometrics, the relative qualities of samples of the separate modes may be used to augment a fusion process [11], [12].
4. *Sample Replacement*. To negate the effects of template aging, a quality measurement may be used to determine whether a newly acquired sample should replace the enrolled one. An alternative would be to retain both the old and new samples for use in a multi-instance fusion scheme.
5. *Template Update*. Again, to address template aging, some systems instead combine old and new sample features. Quality could be used in this process.

3 Properties of a Quality Measure

This section gives needed background material, including terms, definitions, and data elements, to support quantifying the performance of a quality algorithm.

Throughout this paper, we use low quality values to indicate poor sample properties. This is at odds with some systems (for example, the NIST Fingerprint Image Quality (NFIQ) algorithm [13]), for which low values indicate good "quality." Accordingly, this paper transforms the raw NFIQ values $1 \dots 5$ using $Q = 6 - \mathrm{NFIQ}$.

3.1 Quality as Summary Statistic

Consider a data set D containing two samples, $d_i^{(1)}$ and $d_i^{(2)}$ collected from each of $i = 1, \dots, N$ individuals. The first sample can be regarded as an enrollment image, the second as a user sample collected later for verification or identification purposes. The appropriate composition of this data set for quality measurement algorithm assessment is discussed later in Section 5. For now, consider that a quality algorithm Q can be run on the ith enrollment sample to produce a quality value

$$q_i^{(1)} = Q\left(d_i^{(1)}\right) \tag{1}$$

and likewise for the authentication (use-phase) sample

$$q_i^{(2)} = Q\left(d_i^{(2)}\right). \tag{2}$$

We have thus far suggested that these qualities are scalars, as opposed to vectors, for example. Operationally, the requirement for a scalar is not necessary: A vector could be stored and could be used by some predictor. The fact that quality has historically been conceived of as scalar is a widely manifested restriction. For example, BioAPI [14] has a signed single byte value, BioAPI_QUALITY; and the headers of the ISO/IEC biometric data interchange format standards [15] have one or two byte fields for quality. We do not further address the issue of vector quality quantities other than to say that they could be used to specifically direct reacquisition attempts (e.g., camera settings), they have been considered (e.g., the defect fields of [3]), and their practical use would require application of a discriminant function.

3.2 Relationship to Matching

We now formalize our premise that biometric quality measures should predict performance. That is, we formalize quality values q_i that are related to recognition error rates. A formal statement of such requires an appropriate, relevant, and tractable definition of performance. Consider K verification algorithms, V_k, that compare pairs of samples (or templates derived from them) to produce match (i.e., genuine) similarity scores,

$$s_{ii}^{(k)} = V_k\left(d_i^{(1)}, d_i^{(2)}\right), \tag{3}$$

and, similarly, nonmatch (impostor) scores,

$$s_{ij}^{(k)} = V_k\left(d_i^{(1)}, d_j^{(2)}\right) \quad i \neq j. \tag{4}$$

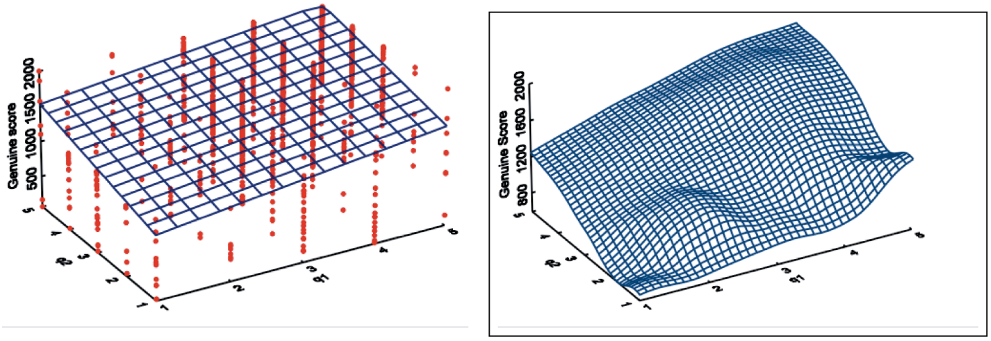
If we now posit that two quality values can be used to produce an estimate of the genuine similarity score that matcher k would produce on two samples

$$s_{ii}^{(k)} = P\left(q_i^{(1)}, q_i^{(2)}\right) + \epsilon_{ii}^{(k)}, \tag{5}$$

where the function P is some predictor of a matcher k's similarity scores, and ϵ_{ii} is the error in doing so for the ith score. Substituting (1) gives

$$s_{ii}^{(k)} = P\left(Q\left(d_i^{(1)}\right), Q\left(d_i^{(2)}\right)\right) + \epsilon_{ii}^{(k)} \tag{6}$$

and it becomes clear that, together, P and Q would be perfect imitators of the matcher V_k in (3) if it was not necessary to apply Q to the samples separately. This separation is usually a necessary condition for a quality algorithm to be useful because, at least half of the time (i.e., enrollment), only one sample is available, see Section 2. Thus, the quality problem is hard, first, because Q is considered to produce a scalar and, second, because it is applied separately to the samples. The obvious consequence of this formulation is that it is inevitable that quality values will imprecisely map to similarity scores, i.e., there will be a scatter of the known scores, s_{ii}, for the known qualities $q_i^{(1)}$ and $q_i^{(2)}$. For example, Fig. 2 shows the raw similarity scores from a commercial fingerprint matcher versus the transformed integer quality scores from the NFIQ algorithm [5], where NFIQ native scores are mapped to $Q = 6 - \mathrm{NFIQ}$. Fig. 2a also includes a least squares linear fit and Fig. 2b shows a cubic spline fit of the same data. Both trend in the correct direction: worse quality gives lower similarity scores. However, even though the residuals in the spline fit are smaller than those for the linear, they still are not small. Indeed, even with a function of arbitrarily high order, it will not be possible to fit the observed scores perfectly if quality values are discrete (as they are for NFIQ). By including the two fits of the raw data, we do not assert that scores should be linearly related to the two quality values (and certainly not



(a) (b)

Fig. 2. Dependence of raw genuine scores on the transformed NFIQ qualities of the two input samples. (a) Linear fit. (b) Spline fit.

locally cubic). Accordingly, we conclude that it is unrealistic to require quality measures to be linear predictors of the similarity scores; instead, the scores should be a monotonic function (higher quality samples give higher scores).

Thus, our conclusion is that it is futile to consider regression methods because the residuals of (5) will not generally have the needed properties for any fit to hold.

3.3 Quantized Quality Values

Biometric standards quite reasonably recommend quality values in the range of [0, 100] with the implication that there are that many distinct values (i.e., between 6 and 7 bits). Practically, this may not be the case and a coarser quantization, corresponding to $L < 100$ statistically separate levels, is usually achieved. Indeed, although BioAPI [14] states that "no universally accepted definition of quality exists," it goes on to specify four ranges ([0, 25], [26, 50], [51, 75], [76, 100]) with associated meanings: *unacceptable*, *marginal*, *adequate*, and *excellent*. This is a tacit acknowledgment that the range [0, 100] is too fine and that an integer quality value on the range [1, 4] is effectively all that may be needed (or possible). If quality algorithms do not provide 100 statistically distinct levels, an evaluation using $L \ll 100$ would be appropriate. Indeed, quantization of a continuous quality metric down to fewer levels may make evaluation easier and/or more robust. For now, we avoid the details of the mapping (i.e., from [1, 100] to [1, L]) and on whether the tester or the algorithm author should have the responsibility for this and instead suggest that BioAPI's use of $L = 4$ is a tractable operational definition.

This is ad hoc, and, clearly, a mathematical rationale for L (for example, a criterion against which L can be optimized) is preferable. This could be something like the knees of the distribution functions of the genuine and impostor scores, or L levels based on the separation of the two distributions. An alternative might be to let L be a free parameter in a fitting process, analogous to some discovered intrinsic precision. Regardless of how L is determined, for a quality algorithm to be effective and operationally meaningful, its L quality levels shall be statistically separate.

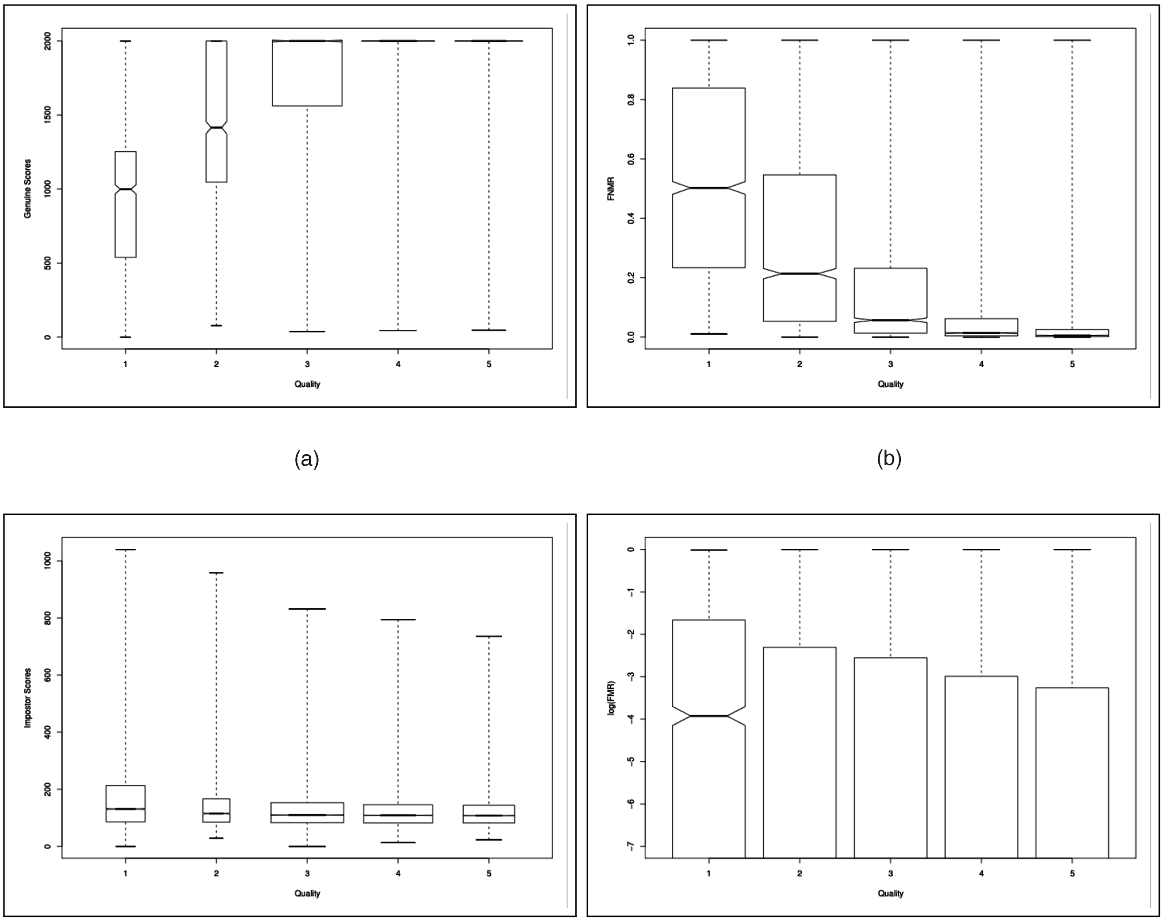
4 Evaluation

This paper's main assertion, that quality should be predictive of performance, has stood so far without a formal specification of how performance should be quantified and whether such performance measures are viable and appropriate. This paper's assumption is that quality measurement algorithms are designed to target application-specific performance variables. For verification, these would be the false match and nonmatch rates. For identification, the metrics would usually be FNMR and FMR [16], but these may be augmented with rank and candidate-list length criteria. Closed-set identification is operationally rare and is not considered here.

Verification is a positive application, which means samples are captured overtly from users who are motivated to submit high quality samples. For this scenario, the relevant performance metric is the false nonmatch rate (FNMR) for genuine users because two high quality samples from the same individual should produce a high score. For FMR, it should be remembered that false matches should occur only when samples are biometrically similar (with regard to a matcher) as, for example, when identical twins' faces are matched. So, high quality images should give very low impostor scores, but low quality images should also produce low scores. Indeed, it is an undesirable trait for a matching algorithm to produce high impostor scores from low quality samples. In such situations, quality measurement should be used to preempt submission of a deliberately poor sample (see the uses discussion in Section 2).

For identification, FNMR is of primary interest. It is the fraction of enrollee searches that do not yield the matching entry on the candidate list. At a fixed threshold, FNMR is usually considered independent of the size of the enrolled population because it is simply dependent on one-to-one genuine scores. However, because impostor acceptance, as quantified by FMR, is a major problem in identification systems, it is necessary to ascertain whether low or high quality samples tend to cause false matches.

For a quality algorithm to be effective, an increase in FNMR and FMR is expected as quality degrades. The plots in Fig. 3 show the relationship of transformed NFIQ quality levels to FNMR and FMR. Figs. 3a and 3c are boxplots of the raw genuine and impostor scores for each of the five NFIQ quality levels. The scores were obtained by applying a commercial fingerprint matcher to left and right index finger impressions of 34,800 subjects. Also shown are boxplots of FNMR and FMR. The result, that the two error rates decrease as quality improves, is expected and beneficial. The FMR shows a much smaller decline. The nonoverlap of the notches in plots of Figs. 3a and 3b demonstrates "strong evidence"



(a)

(b)

(c)

(d)

Fig. 3. Boxplots of genuine scores, FNMR, impostor scores, and FMR for each of five transformed NFIQ quality levels for scores from a commercial matcher. Each quality bin, q, contains scores from comparisons of enrollment images with quality $q^{(1)} \geq q$ and subsequent use-phase images with $q^{(2)} = q$, per the discussion in Section 4.2. The boxplot notch shows the median; the box shows the interquartile range and the whiskers show the extreme values. Notches in (d) are not visible because the medians of FMRs are zero and, therefore, outside the plot range. (a) Genuine. (b) FNMR. (c) Impostor. (d) FMR.

that the medians of the quality levels differ [25]. If the QMA had more finely quantized its output, to $L > 5$ levels, this separation would eventually disappear. This issue is discussed further in Section 4.6.

4.1 Combining Two Samples' Quality Values

Biometric matching involves at least two samples and the challenge is then to relate performance to quality values $q^{(1)}$ and $q^{(2)}$. This empirical dependence of performance on two values was shown in Fig. 2. We simplify the analysis by combining the two qualities

$$q_i = H\Big(q_i^{(1)}, q_i^{(2)}\Big). \tag{7}$$

As discussed in Section 2, it is usually the case that, operationally, a QMA can be used to ensure that an enrollment sample is of high quality. This will be compared later with a sample that typically is of less controlled quality. To capture this concept, we consider $H(x, y) = \min(x, y)$, i.e., the worse of two samples drives the similarity score. We also consider the arithmetic and geometric means, $H(x, y) = (x + y)/2$ and $H(x, y) = \sqrt{xy}$ (see [17]), and the difference function $H(x, y) = |x - y|$ to investigate dependence of similarity score on samples of different quality. We acknowledge that choices for H() are not limited to min(), arithmetic, and geometric mean. We considered those for their relevance to operational scenarios and ease of implementation. We note that, whatever H is used, it should be well defined for allowed values of x and y (e.g., positive values for the geometric mean).

We now describe four methods for the evaluation of quality. All four consider the use of combination functions, H, which are specifically compared in Section 4.3.

4.2 Rank-Ordered Detection Error Trade-Off Characteristics

A quality algorithm is useful if it can at least give an ordered indication of an eventual performance. For example, for L discrete quality levels, there should notionally be L DET characteristics.[1] In the studies that have evaluated

[1] The DET used here plots FNMR versus FMR on log scales. It is unconventional in that it does not transform the data by the CDF of the standard normal distribution. The receiver operating characteristic plots 1-FNMR on a linear scale instead. These characteristics are used ubiquitously to summarize verification performance.

quality measures [4], [13], [16], [17], [23], [24], DETs are the primary metric. We recognize that DET's are widely understood, even expected, but note three problems with their use: Being parametric in threshold, t, they do not show the dependence of FNMR (or FMR) with quality at fixed t, they are used without a test of the significance of the separation of L levels, and partitioning of the data for their computation is underreported and nonstandardized.

We examine three methods for the quality-ranked DET computation. All three use N paired matching images with integer qualities $q_i^{(1)}$ and $q_i^{(2)}$ on the range [1, L]. Associated with these are N genuine similarity scores, s_{ii}, and up to $N(N-1)$ impostor scores, s_{ij}, where $i \neq j$, obtained from some matching algorithm. All three methods compute a DET characteristic for each quality level k. For all thresholds s, the DET is a plot of $\mathrm{FNMR}(s) = M(s)$ versus $\mathrm{FMR}(s) = 1 - N(s)$, where the empirical cumulative distribution functions $M(s)$ and $N(s)$ are computed, respectively, from sets of genuine and impostor scores. The three methods of partitioning differ in the contents of these two sets. The simplest case uses scores obtained by comparing authentication and enrollment samples whose qualities are both k. This procedure (see, for example, [18]) is common but overly simplistic. By plotting

$$\mathrm{FNMR}(s,k) = \frac{\left|\left\{s_{ii} : \; s_{ii} \leq s, \quad q_i^{(1)} = q_i^{(2)} = k\right\}\right|}{\left|\left\{s_{ii} : \; s_{ii} \leq \infty, \; q_i^{(1)} = q_i^{(2)} = k\right\}\right|},$$
$$\mathrm{FMR}(s,k) = \frac{\left|\left\{s_{ij} : \; s_{ij} > s, \quad q_i^{(1)} = q_j^{(2)} = k, \; i \neq j\right\}\right|}{\left|\left\{s_{ij} : \; s_{ij} > -\infty, \; q_i^{(1)} = q_j^{(2)} = k, \; i \neq j\right\}\right|}, \tag{8}$$

the DETs for each quality level can be compared. Although a good QMA will exhibit an ordered relationship between quality and error rates, this DET computation is not operationally representative because an application cannot usually accept only samples with one quality value. Rather, the DET may be computed for verification of samples of quality k with enrollment samples of quality greater than or equal to k,

$$\mathrm{FNMR}(s,k) = \frac{\left|\left\{s_{ii} : \; s_{ii} \leq s, \quad q_i^{(1)} \geq k, \; q_i^{(2)} = k\right\}\right|}{\left|\left\{s_{ii} : \; s_{ii} \leq \infty, \; q_i^{(1)} \geq k, \; q_i^{(2)} = k\right\}\right|},$$
$$\mathrm{FMR}(s,k) = \frac{\left|\left\{s_{ij} : \; s_{ij} > s, \quad q_i^{(1)} \geq k, \; q_j^{(2)} = k, \; i \neq j\right\}\right|}{\left|\left\{s_{ij} : \; s_{ij} > -\infty, \; q_i^{(1)} \geq k, \; q_j^{(2)} = k, \; i \neq j\right\}\right|}, \tag{9}$$

we model the situation in which the enrollment samples are at least as good as the authentication (i.e., user submitted) samples. Such a use of quality would lead to failures to acquire for the low quality levels.

If, instead, we compare performance across *all* authentication samples against enrollment samples of quality greater than or equal to k,

$$\mathrm{FNMR}(s,k) = \frac{\left|\left\{s_{ii} : \; s_{ii} \leq s, \quad q_i^{(1)} \geq k\right\}\right|}{\left|\left\{s_{ii} : \; s_{ii} \leq \infty, \; q_i^{(1)} \geq k\right\}\right|},$$
$$\mathrm{FMR}(s,k) = \frac{\left|\left\{s_{ij} : \; s_{ij} > s, \quad q_i^{(1)} \geq k, \; i \neq j\right\}\right|}{\left|\left\{s_{ij} : \; s_{ij} > -\infty, \; q_i^{(1)} \geq k, \; i \neq j\right\}\right|}, \tag{10}$$

we model the situation where quality control is applied only during enrollment. If repeated enrollment attempts fail to produce a sample with quality above some threshold, a failure-to-enroll (FTE) would be declared. This scenario is common and possible because enrollment, as an attended activity, tends to produce samples of better quality than authentication.

The considerable differences between these three formulations are evident in the DETs of Fig. 4 for which the NFIQ algorithm [5] for the predicting performance of a commercial fingerprint system was applied to over 61,993 genuine and 121,997 impostor comparisons (NFIQ native scores were transformed to $Q = 6 - \mathrm{NFIQ}$). In all cases, the ranked separation of the DETs is excellent across all operating points. We recommend that (9), as shown in Fig. 4b, be used because of it is a more realistic operational model.

However, as relevant as DET curves are to expected performance, we revisit here a very important complication. Because DET characteristics quantify the separation of the genuine and impostor distributions and combine the effect of quality on both genuine and impostor performance, we lose sight of the separate effects of quality on FNMR and FMR.

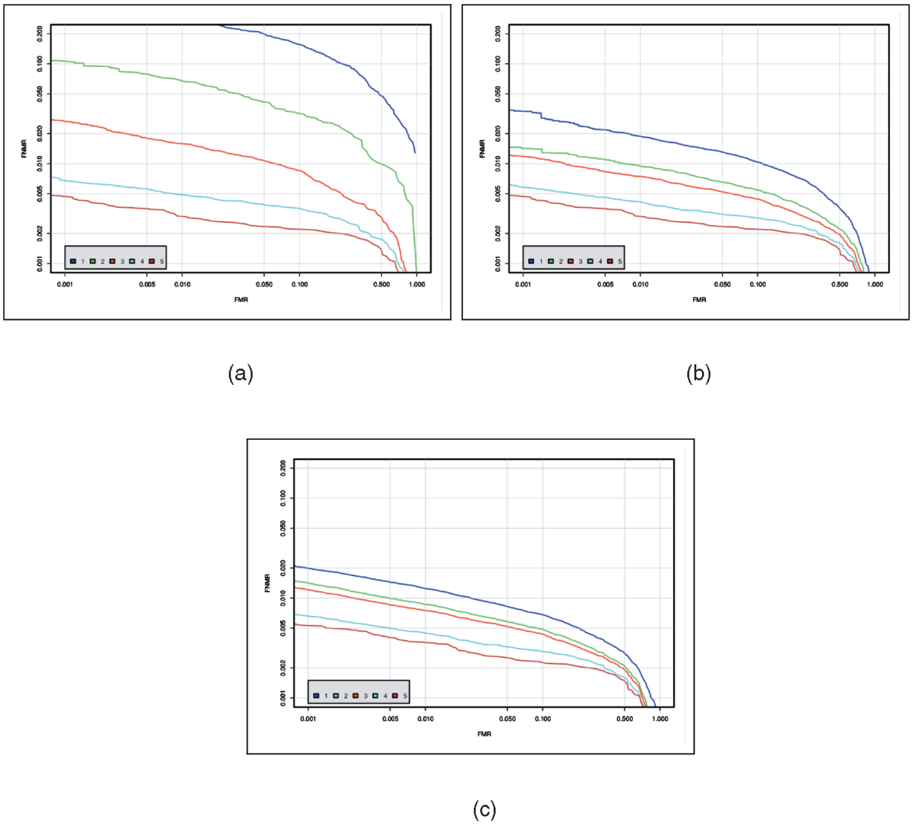
That quality should be evaluated at all in relation to impostor performance (i.e., FMR) is dubious. For example, does a biometric recognition system produce a low impostor score when the two samples are of low quality? Perhaps, but does it also produce lower impostor scores when the samples are of high quality? Under what circumstances are the impostor scores high? (Such questions may be simpler to answer for a fingerprint quality apparatus that predicts a minutiae-based matcher's performance on the basis of number and type, etc., of minutia.)

In any case, we conclude that DETs, while familiar and highly relevant, confound genuine and impostor scores. The alternative is to look at the specific dependence of the error rates on quality at some fixed threshold. Indeed, for verification applications, the variation in FNMR with quality is key because the majority of transactions are genuine attempts. For negative identification systems (e.g., watchlist applications) in which users are usually not enrolled, the variation of FMR with quality is critical. This approach is followed in the next section.

4.3 Error versus Reject Curves

In this section, we propose using error versus reject curves as an alternative means of evaluating QMAs. The goal is to state how efficiently rejection of low quality samples results in improved performance. This again models the operational case in which quality is maintained by reacquisition after a low quality sample is detected. Consider that a pair of samples (from the same subject), with qualities $q_i^{(1)}$ and $q_i^{(2)}$, are compared to produce a score $s_{ii}^{(k)}$, and this is repeated for N such pairs.

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



(a)

(b)

(c)

Fig. 4. Quality ranked detection error trade-off characteristics. Each plot shows five traces corresponding to five transformed NFIQ levels. (a) $q^{(1)} = i$, $q^{(2)} = i$. (b) $q^{(1)} \geq i$, $q^{(2)} = i$. (c) $q^{(1)} = i$, $q^{(2)} \geq -\infty$.

$$R(u, v) = \left\{ j \;:\; q_j^{(1)} < u, \quad q_j^{(2)} < v \right\}. \tag{11}$$

The FNMR is the fraction of genuine scores below threshold computed for those samples *not* in this set

$$\mathrm{FNMR}(t, u, v) = \frac{\left|\{s_{jj} : s_{jj} \leq t, j \notin R(u, v)\}\right|}{\left|\{s_{jj} : s_{jj} \leq \infty, j \notin R(u, v)\}\right|}. \tag{12}$$

The value of t is fixed[2] and u and v are varied to show the dependence of FNMR on quality.

For the one-dimensional case when only one quality value is used (see Section 4.1), the rejection set is

$$R(u) = \left\{ j \;:\; H(q_j^{(1)}, q_j^{(2)}) < u \right\}. \tag{13}$$

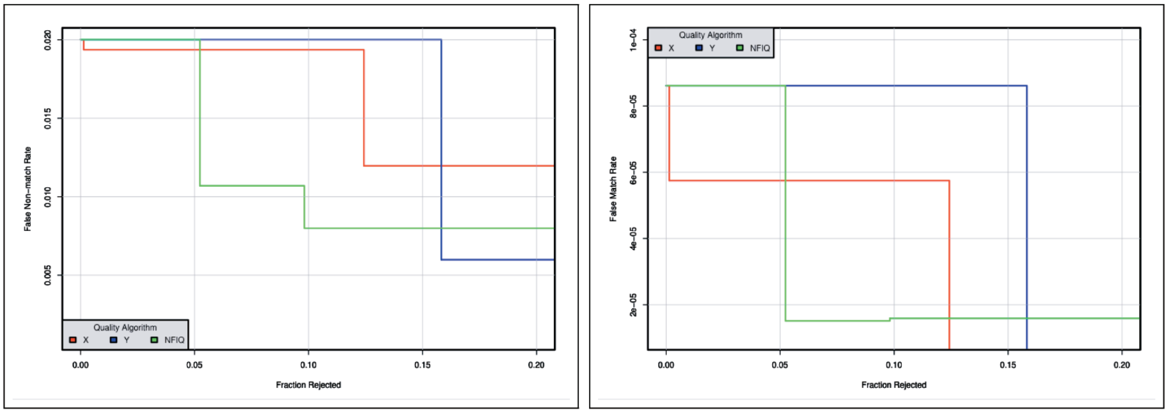
FNMR is false nonmatch performance as the proportion of nonexcluded scores below the threshold.

$$\mathrm{FNMR}(t, u) = \frac{\left|\{s_{jj} : s_{jj} \leq t, j \notin R(u)\}\right|}{\left|\{s_{jj} : s_{jj} \leq \infty, j \notin R(u)\}\right|}. \tag{14}$$

If the quality values are perfectly correlated with the genuine scores, then when we set t to give an overall FNMR of x and then reject proportion x with the lowest qualities, a recomputation of FNMR should be zero. Thus, a good quality metric correctly labels those samples that cause low genuine scores as poor quality. For a good quality algorithm, FNMR should decrease quickly with the fraction rejected. The results of applying this analysis are shown in Fig. 5. Note that the curves for each of the three fingerprint quality algorithms trend in the correct direction, but that, even after rejection of 20 percent, the FNMR value has fallen only by about a half from its starting point. Rejection of 20 percent is probably not an operational possibility unless an immediate reacquisition can yield better quality values for those people. Yoshida and Hara, using the same approach, reported similar figures [19]. Note, however, that, for NFIQ, the improvement is achieved after rejection of just 5 percent. In verification applications such as access control, the prior probability of an impostor transaction is low and, thus, the overall error rate is governed by false nonmatchers. In such circumstances, correct detection of samples likely to be falsely rejected should drive the design of QMAs.

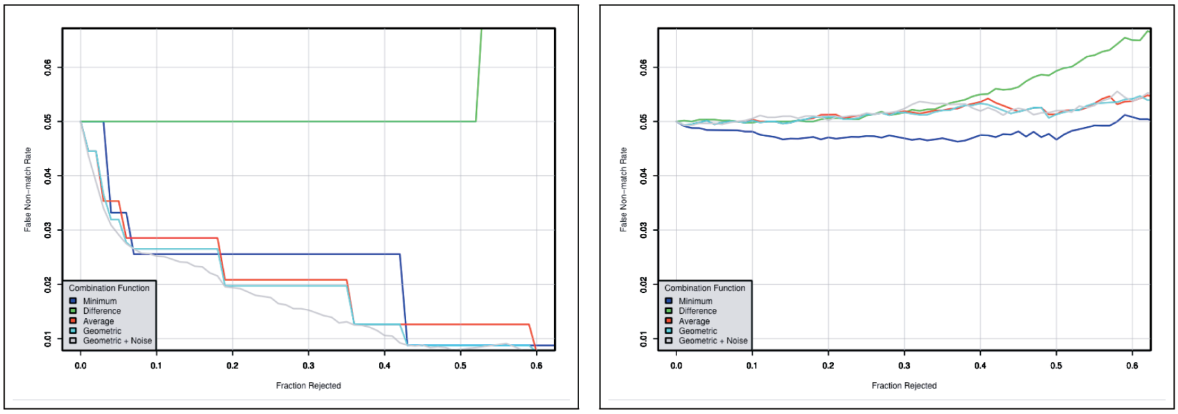
Fig. 6 shows error versus reject behavior for the NFIQ quality method when the various $H(q_1, q_2)$ combination functions of Section 4.1 are used. Between the minimum, mean, and geometric mean functions there is little difference. The geometric mean is best (absent a significance test) with steps occurring at values corresponding to the square roots of the product of NFIQ values. The gray line in the figure shows $H = \sqrt{q_1 q_2} + N(0, 0.01)$, where the Gaussian noise serves to randomly reject samples within a quality level and produces an approximation of the lower convex hull of the geometric mean curve. The green line result, for $H = |q_1 - q_2|$, shows that the transformed genuine comparison score is unrelated to the difference in the qualities of the samples. Instead, the conclusion is that FNMR is related to monotonic functions of the two values. The applicability of this result to other quality methods is not known.

[2] Any threshold may be used. Practically, it will be set to give some reasonable false nonmatch rate, f, by using the quantile function, the empirical cumulative distribution function of the genuine scores, $t = M^{-1}(1 - f)$.



(a) (b)

Fig. 5. Error versus reject performance for three fingerprint quality methods. (a) and (b) show reduction in FNMR and FMR at a fixed threshold as up to 20 percent of the low quality samples are rejected. The similarity scores come from a commercial matcher. (a) Finger-FNMR. (b) Finger-FMR.



(a) (b)

Fig. 6. Dependence of the error versus reject characteristic on the quality combination function H(.). The plots show, for a fixed threshold, the decrease in FNMR as up to 60 percent of the low quality values are rejected. The similarity scores come from commercial matchers. The steps in (a) are the result of the discrete quality metric. Continuous quality metrics such as in (b) do not usually exhibit such steps. (a) Finger. (b) Face.

4.4 Generalization to Multiple Matchers

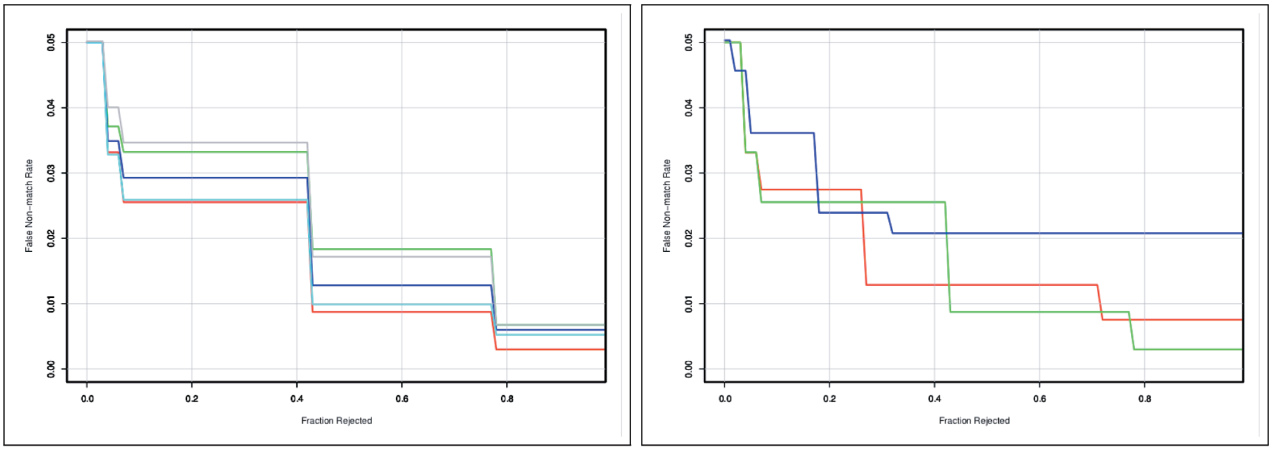
It is a common contention that the efficacy of a quality algorithm is necessarily tied to a particular matcher. We observe that this one-matcher case is commonplace and useful in a limited fashion and should therefore be subject to evaluation. However, we also observe that it is possible for a quality algorithm to be capable of generalizing across *all* (or a class of) matchers and this too should be evaluated.

Generality to multiple matchers can be thought of as an interoperability issue: Can supplier A's quality measure be used with supplier B's matcher? Such a capability will exist to the extent that pathological samples do present problems to both A and B's matching algorithms. However, the desirable property of generality exposes another problem: We cannot expect performance to be predicted absolutely because there are good and bad matching systems. A system here includes all of the needed image analysis and comparison tasks. Rather, we assert that a quality algorithm intended to predict performance generally need only be capable of giving a relative or rank ordering, i.e., low quality samples should give lower performance than high quality samples.

The plots of Fig. 7 quantify this generalization for the NFIQ system using the error versus reject curves of Section 4.3. Fig. 7a includes five traces, one for each of five verification algorithms. The vertical spread of the traces indicates some disparity in how well NFIQ predicts the performance of the five matchers. A perfectly general QMA would produce no spread.

4.5 Number of Levels of Quality

A quality metric is more useful if, operationally, it may be thresholded at one of many distinct operating points. Thus, a discrete-valued quality measure is better if performance is significantly different for level q_k than for q_{k-1} for all levels $1 \leq k \leq K$. Having already stated that FNMR should be monotonic in the quality value, $\text{FNMR}_k \leq \text{FNMR}_{k-1}$, we now additionally require that the quality levels be statistically distinct. If they are not, they could be mapped to fewer levels that are statistically distinct. Real values can be quantized. Formally, we propose testing this by using the Kolmogorov Smirnov (KS) test to determine whether the distribution of the genuine scores for level q_k is distinct from that of q_{k-1}. The KS test is nonparametric, distribution-free, and simple. The KS statistic is simply the maximum absolute difference



(a)

(b)

Fig. 7. Error versus reject characteristics showing how NFIQ generalizes across (a) five verification algorithms and (b) three operational data sets. The steps in (a) occur at the same rejection values because the matchers were run on a common database.

TABLE 1
KS Test for Separation of Quality-Specific Genuine Score Distributions

Quality	$q =$	No. Scores at Level		KS Statistic Between Genuine Distribution of	p
Method	$min(q^{(1)}, q^{(2)})$	$k-1$	k	$\{s_{ii} : q_i^{(1)} = q_i^{(2)} = q\}$ and	value
				$\{s_{ii} : q_i^{(1)} = q_i^{(2)} = q-1\}$	
X	2	3647	3191	0.20	0
X	3	3191	20553	0.34	0
X	4	20553	24297	0.24	0
X	5	24297	17975	0.08	0
Y	2	11023	4878	0.30	0
Y	3	4878	6637	0.05	0
Y	4	6637	8440	0.07	0
Y	5	8440	10751	0.05	0
Y	6	10751	11902	0.02	0.006

The data apply to 69,663 genuine fingerprint comparisons.

between the two distributions' cumulative distributions functions.

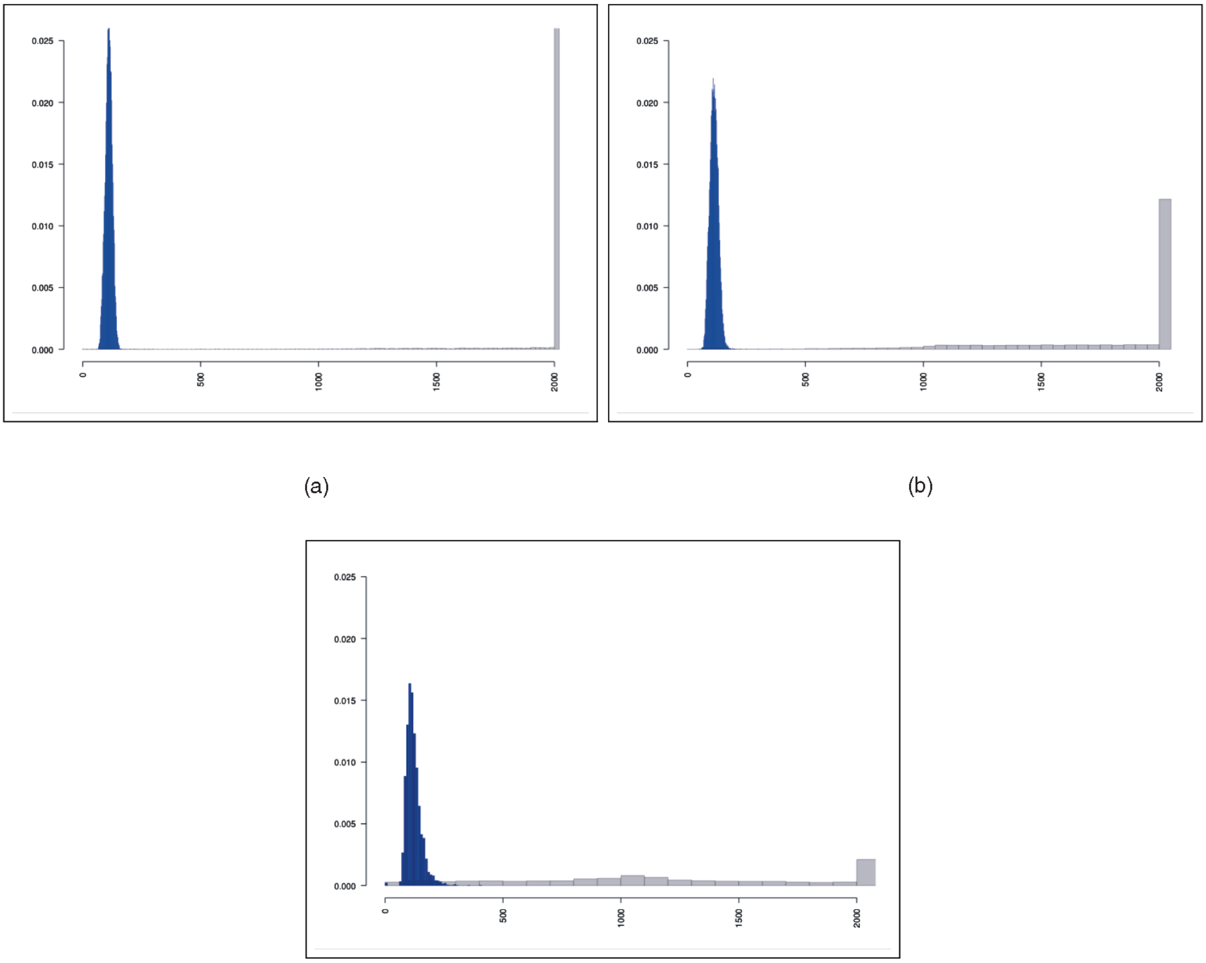
Table 1 shows example results for two fingerprint quality methods. In both cases, the observed KS statistic values are smaller for higher quality levels (where performance is always very high) and are significant: The p-value exceeds 10^{-7} on only one occasion. A higher p-value there would have indicated that quality method Y's level 5 and 6 are insignificantly different. The results do not demonstrate such behavior, presumably because the algorithms were created with a reasonable number of levels as a design parameter.

4.6 Measuring Separation of Genuine and Impostor Distributions

We can evaluate quality algorithms on their ability to predict how far a genuine score will lie from its impostor distribution. This means, instead of evaluating a quality algorithm solely based on its FNMR (i.e., genuine score distribution) prediction performance, we can augment the evaluation by including a measure of FMR because correct identification of an enrolled user depends both on correctly finding the match and on rejecting the nonmatches. Note also that a quality algorithm could invoke a matcher to compare the input sample with some internal background samples to compute sample mean and standard deviation.

The plots of Fig. 8 show, respectively, the genuine and impostor distributions for adjusted NFIQ values 1, 3, and 5. The overlapping of genuine and impostor distributions for the poorest NFIQ means higher recognition errors for that NFIQ level and vice versa; the almost complete separation of the two distribution for the best quality samples indicates lower recognition error. NFIQ was trained to specifically exhibit this behavior.

We again consider the KS statistic. For better quality samples, a larger KS test statistic (i.e., higher separation between genuine and impostor distribution) is expected. Each row of Table 2 shows KS statistics for one of the three quality algorithms that we tested. KS statistics for each quality levels $u = 1, , 5$ are computed by first computing the genuine (i.e., $\{s_{ii} : (i, i) \in R(u)\}$) and impostor (i.e., $\{s_{ij} : (i, j) \in R(u), i \neq j\}$) empirical cumulative distributions, where $R(u) = \{(i, j) : H(q_i^{(1)}, q_j^{(2)}) = u\}$. Thereafter, the largest absolute difference between the genuine and impostor distributions of quality u is measured and plotted. (Note that, to keep quality algorithm providers anonymous, we only reported the KS statistics of the lowest four quality levels.)



(a)

(b)

(c)

Fig. 8. There is a higher degree of separation between the genuine and impostor distribution for better quality samples as measured by NFIQ. (a) Best. (b) Middle. (c) Worst.

5 Quality Reference Data Sets

This section addresses two issues: what data should be used for testing a quality apparatus and how to annotate the samples of a reference corpus with quality values.

5.1 Data to Be Used for Testing

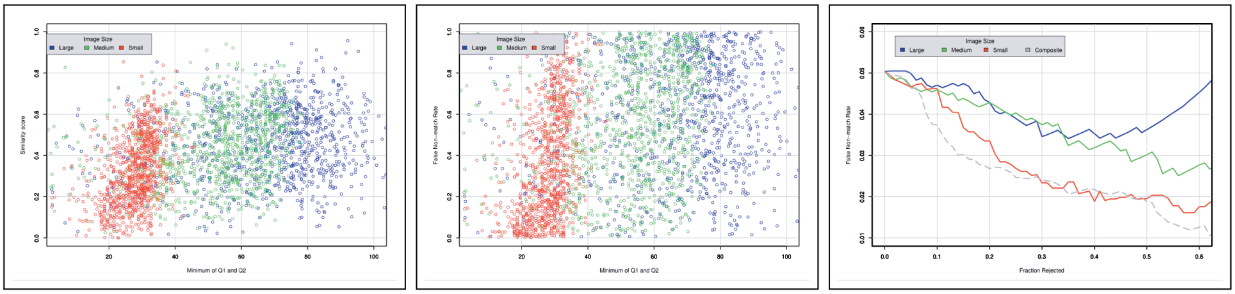
A quality measurement algorithm could be evaluated using data specifically collected with deliberate defects. For example, quality could be degraded by misfocusing the camera. Such data have several notable uses: development of a quality measurement algorithm, teaching best practice by counterexample, and assessing the performance of a product intended to test the conformity of an image or signal to an underlying standard.[3] However, we argue that this type of data should not be used for evaluation for four reasons. First, such data is, by definition, laboratory data and therefore would lack application-specific operational relevance. Second, by applying certain kinds of degradation to the images, the evaluator is making assumptions about the performance sensitivities of matching algorithms. For example, if the chin is cropped from a face image, then this may be immaterial to a face recognition algorithm. Third, it would be difficult or impossible to collect samples that express all possible combinations of quality defects and particularly with their natural frequency and to their natural degree. Finally, the laboratory data may not ordinarily be available in large quantities.

Instead, this paper considers the use of operationally representative data, i.e., samples harvested during real-world usage or from a relevant scenario test [20]. By definition, this has the advantage of having relevance to the operation. We showed examples of such data in Section 4.3.

TABLE 2
KS Statistics for Quality Levels of Three Quality Algorithms

Quality Algorithm	$Q = 1$	$Q = 2$	$Q = 3$	$Q = 4$
Quality Algorithm 1	0.649	0.970	0.988	0.993
Quality Algorithm 2	0.959	0.995	0.996	0.997
Quality Algorithm 3	0.918	0.981	0.994	0.997

[3] For example, the ISO/IEC 19794-5 Face Recognition Interchange Format standard puts quantitative limits on the amount of quality related degradation such as from blur, nonfrontal pose, and the number of gray levels.



(a) (b) (c)

Fig. 9. Scatter plots of scores and FNMR values versus quality and the error versus reject curves for a face quality metric applied to a face database composed of images at full (blue), half (green), and quarter size (red). (a) Score versus $\min(q_1, q_2)$. (b) FNMR versus $\min(q_1, q_2)$. (c) Error versus Reject.

However, if a test compares quality algorithms or is making a more general assessment of the technology, then an aggregated corpus that spans the quality spectrum might be employed. Such a set might include fingerprint images gathered from employees during an access control enrollment and, subsequently, authentication and also samples collected outdoors and from persons detained in adverse law enforcement environments. This construction, unlike the dedicated laboratory collection described above, does not manipulate the sample acquisition process.

To illustrate the importance of using an aggregated corpus for evaluation, we use the Color FERET database [21]. The frontal *fa* and *fb* images from each of 852 subjects were used at full, half, and quarter resolutions. These are input to a quality algorithm and a matching algorithm from the same supplier. The reduction in image size forcibly induces the reductions in both quality and match scores evident in Fig. 9. Note, however, that, for any one of the three point clouds in Fig. 9a, there is large variation in score in relation to quality—a trend that is not improved by plotting $M(s)$ instead (Fig. 9b). This reflects the difficulty of the face quality problem.

The final graph, Fig. 9c, shows the error versus reject performance for each of the image sizes separately and for the aggregate data set. This latter curve, in gray, is lower than the others. This demonstrates the value of using composite sets for evaluation purposes. Also worthy of note is that the error versus reject performance at any of the three sizes is superior to that in Fig. 6b, which uses the same algorithm on a more uniform data set. Those images are about the same size as the half-size FERET images but are more consistently posed (i.e., frontal), sized, and compressed and all subjects do not wear eyeglasses. The suggestion then is that the more homogenous the corpus, the more difficult it is for a quality algorithm to predict variation in similarity scores. We should emphasize that the algorithm was provided to the authors without any claim of efficacy or recommended domain of use.

5.2 Construction of a Reference Data Set

In this section, we advance a procedure for annotating a sample corpus with target quality values. The strategy is to assign values that are directly related to the results of matching those samples. This is achieved by taking the similarity scores from $K \geq 1$ matching algorithms, classifying them, and, in the case of $K > 1$, taking a consensus. The result is a reference set useful to quality algorithm developers. It would be of use for the tuning of an operational quality algorithm when the matcher and kind of data are known.

The input to the procedure is a representative sample database. The output is an annotation of each sample with a scalar quality target. The method presumes the availability of a representative matching algorithm, which will be used to compare samples to produce both genuine and impostor similarity scores. It is therefore implied that two or more samples per person are available.

5.2.1 Data

Data gathered in a target operational application would be most realistic. Contemporary matchers perform extremely well on most images and it is therefore necessary to preferentially stack the reference set with samples that are naturally problematic to the matcher. For example, for a reference fingerprint data set to span the quality spectrum, it should be balanced in terms of finger position (right/left index/thumb/middle), finger impression (roll/plain/flat), sex, age, and capture device. Lack of data often renders it difficult to create such a balanced data set.

5.2.2 Target Quality Assignment

We seek to assign a ground-truth quality score to each image in a reference data set. We ensure that the quality values are representative of performance by associating the image with similarity scores as follows: Consider a biometric corpus containing two samples, $d_i^{(1)}$ and $d_i^{(2)}$, for each of N individuals, $i = 1, \ldots, N$. The first samples represent enrollment samples and the second samples represent those for authentication. The following procedure assigns quality values $q_i^{(1)}$ and $q_i^{(2)}$ to all images in the corpus.

For each person i:

1. Compare the first and second samples using the kth matcher to produce a genuine score. Repeating (3):

$$s_{ii}^{(k)} = V_k\left(d_i^{(1)}, d_i^{(2)}\right). \qquad (15)$$

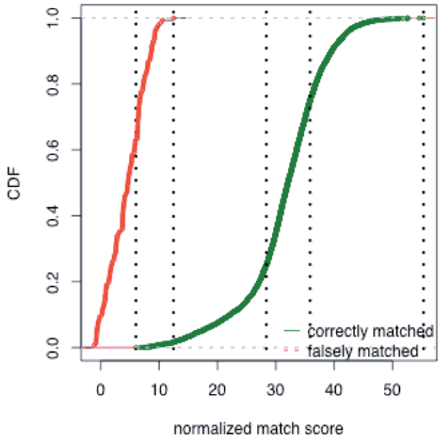


Fig. 10. Empirical cumulative distribution functions for the top-ranked genuine scores and for the impostor scores. The vertical lines are one possible way of binning normalized match score. Samples are assigned quality numbers corresponding to the bin of their normalized match score.

2. Use the kth matcher to compare the first sample from person i with the second sample from all $j = 1, \ldots, N$ and $i \neq j$ other persons. The result is $J = N - 1$ impostor scores:

$$s_{ij}^{(k)} = V_k\left(d_i^{(1)}, d_j^{(2)}\right). \tag{16}$$

   (This is essentially(4).)
3. Insert i into set $\mathcal{T}$ if its genuine score is larger than all of its impostor scores, i.e., $s_{ii}^{(k)} > s_{ij}^{(k)}\ \forall j$. This is a rank 1 condition.
4. For the first sample of each person $d_i^{(1)}$, compute the sample mean and standard deviation of its J associated impostor scores

$$m_i = \qquad J^{-1} \sum\nolimits_{j=1}^{J} s_{ij}^{(k)}, \tag{17}$$

$$\sigma_i = (J-1)^{-1} \sum\nolimits_{j=1}^{J} \left(s_{ij}^{(k)} - m_i\right)^2. \tag{18}$$

5. Normalize the genuine score from (15) using the impostor statistics

$$z_i = (s_{ii} - m_i)/\sigma_i. \tag{19}$$

Once all normalized similarity scores have been computed:

1. Compute two empirical cumulative distribution functions: one for the top-ranked genuine scores of set $\mathcal{T}$,

$$C(z) = \frac{|\{z_i : i \in \mathcal{T}, z_i \leq z\}|}{|\{z_i : i \in \mathcal{T}, z_i \leq \infty\}|}, \tag{20}$$

   and another for those not in that set.

$$W(z) = \frac{|\{z_i : i \notin \mathcal{T}, z_i \leq z\}|}{|\{z_i : i \notin \mathcal{T}, z_i \leq \infty\}|}. \tag{21}$$

   These cumulative distribution functions are plotted in Fig. 10 for live-scan images of the right-index fingers of 6,000 individuals and scores of a commercial fingerprint matcher. These were produced in a US Government test using sequestered operational data.

TABLE 3
Binning Normalized Match Score

Category	Description	Range of normalized match score
1	poor	$\{z_i : -\infty \leq z_i < C^{-1}(0)\}$
2	fair	$\{z_i : C^{-1}(0) \leq z_i < W^{-1}(1)\}$
3	good	$\{z_i : W^{-1}(1) \leq z_i < C^{-1}(0.25)\}$
4	very good	$\{z_i : C^{-1}(0.25) \leq z_i < C^{-1}(0.75)\}$
5	excellent	$\{z_i : C^{-1}(0.75) \leq z_i\}$

2. Bin the normalized match score range into K bins based on quantiles of the normalized match score distribution. One strategy, for $K = 5$, is shown in Table 3 in which F^{-1} is the quantile function and $F^{-1}(0)$ and $F^{-1}(1)$ denote the empirical minima and maxima, respectively. If $W^{-1}(1) \geq C^{-1}(0.25)$, an appropriate quartile of $C(z)$ must be selected.
3. Sample d_i is assigned target quality q_i corresponding to the bin of its normalized match score z_i from (19).
4. The procedure is repeated for sample $d_i^{(2)}$ by swapping indices 1 and 2 in (15) and (16). Since one sample will have an impostor distribution different from another, two different samples of the same subject may have different normalized match scores and, therefore, different quality values.
5. The procedure is repeated for scores of all V matchers.
6. Samples with identical quality assignments from *all* V matchers become members of the Quality Reference Data Set. Those without unanimity are discarded.
7. If, for some quality bins, no consensus was made among *all* V matchers, the procedure could be started from Step 2 above with modified bin boundaries.

This procedure has been used to form NFIQ training and compliance set [22], only with different bin boundaries. These were set by manual inspection to give useful categorization of the normalized match score statistic.

6 Conclusion

Biometric quality measurement is an operationally important and difficult problem that is nevertheless massively under-researched in comparison to the primary feature extraction and pattern recognition tasks. In this paper, we enumerated the ways in which it is useful to compute a quality value from a sample. In all cases, the ultimate intention is to improve matching performance. We asserted, therefore, that quality algorithms should be developed to explicitly target matching error rates and not human perceptions of sample quality. To this end, we defined a procedure for the annotation of a reference sample set with target quality values. We gave several means for assessing the efficacy of quality algorithms. We reviewed the existing practice, cautioned against the use of detection error trade-off characteristics as the primary metrics, and, instead, advanced boxplots and error versus reject curves as preferable. We suggest that algorithm designers should target false nonmatch rate as the primary performance indicator.

In conclusion, we posit that quality summarization as a predictor of recognition performance is a difficult problem and we encourage the academic community to consider the problem and extend the quantitative methods of this paper in advancing their work.

References

[1] T. Ko and R. Krishnan, "Monitoring and Reporting of Fingerprint Image Quality and Match Accuracy for a Large User Application," *Proc. 33rd Applied Image Pattern Recognition Workshop,* pp. 159-164, 2004.

[2] *Proc. NIST Biometric Quality Workshop,* Mar. 2006, http://www.itl.nist.gov/iad/894.03/quality/workshop/presentations.html.

[3] D. Benini et al., *ISO/IEC 29794-1 Biometric Quality Framework Standard,* first ed., JTC1/SC37/Working Group 3, Jan. 2006, http://isotc.iso.org/isotcportal.

[4] Y. Chen, S. Dass, and A. Jain, "Fingerprint Quality Indices for Predicting Authentication Performance," *Proc. Audio- and Video-Based Biometric Person Authentication,* pp. 160-170, July 2005.

[5] E. Tabassi, *Fingerprint Image Quality, NFIQ,* NISTIR 7151 ed., Nat'l Inst. of Standards and Technology, 2004.

[6] F. Alonso-Fernandez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "A Review of Schemes for Fingerprint Image Quality Computation," *COST 275—Biometrics-Based Recognition of People over the Internet,* Oct. 2005.

[7] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro, "MCYT Baseline Corpus: A Bimodal Biometric Database," *Proc. IEE Conf. Vision, Image, and Signal Processing,* vol. 150, no. 6, pp. 395-401, Dec. 2003.

[8] E. Lim, X. Jiang, and W. Yau, "Fingerprint Quality and Validity Analysis," *Proc. IEEE Conf. Image Processing,* vol. 1, pp. 469-472, Sept. 2002.

[9] Bioscrypt Inc., *Systems and Methods with Identify Verification by Comparison and Interpretation of Skin Patterns Such as Fingerprints,* June 1999, http://www.bioscrypt.com..

[10] L.M. Wein and M. Baveja, "Using Fingerprint Image Quality to Improve the Identification Performance of the U.S. Visit Program," *Proc. Nat'l Academy of Sciences,* 2005, www.pnas.org/cgi/doi/10.1073/pnas.0407496102.

[11] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Discriminative Multimodal Biometric Authentication Based on Quality Measures," *Pattern Recognition,* vol. 38, no. 5, pp. 777-779, May 2005.

[12] E. Tabassi, G.W. Quinn, and P. Grother, "When to Fuse Two Biometrics," *IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR '06),* June 2006.

[13] E. Tabassi, "A Novel Approach to Fingerprint Image Quality," *Proc. IEEE Int'l Conf. Image Processing,* Sept. 2005.

[14] C. Tilton et al., *The BioAPI Specification.* Am. Nat'l Standards Inst., Inc., 2002.

[15] ISO/IECJTC1/SC37/Working Group 3, *ISO/IEC 19794 Biometric Data Interchange Formats,* 2005, http://isotc.iso.org/isotcportal.

[16] A.J. Mansfield, *ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework,* FDIS ed., JTC1/SC37/ Working Group 5, Aug. 2005, http://isotc.iso.org/isotcportal.

[17] J. Fierrez-Aguilar, L. Muñoz-Serrano, F. Alonso-Fernandez, and J. Ortega-Garcia, "On the Effects of Image Quality Degradation on Minutiae and Ridge-Based Automatic Fingerprint Recognition," *Proc. IEEE Int'l Carnahan Conf. Security Technology,* Oct. 2005.

[18] D. Simon-Zorita, J. Ortega-Garcia, J. Fierrez-Aguilar, and J. Gonzalez-Rodriguez, "Image Quality and Position Variability Assessment in Minutiae-Based Fingerprint Verification," *IEE Proc. Vision, Image and Signal Processing,* special issue on biometrics on the Internet, vol. 150, no. 6, pp. 395-401, Dec. 2003.

[19] A. Yoshida and M. Hara, "Fingerprint Image Quality Metrics that Guarantees Matching Accuracy," *Proc. NIST Biometric Quality Workshop,* Mar. 2006, http://www.itl.nist.gov/iad/894.03/quality/workshop/presentations.html.

[20] M. Thieme, *ISO/IEC 19795-2 Biometric Performance Testing and Reporting: Scenario Testing,* cd2 ed., JTC1/SC37/Working Group 5, Aug. 2005, http://isotc.iso.org/isotcportal.

[21] *The Color FERET Face Database,* Nat'l Inst. Standards and Technology, http://www.nist.gov/humanid/feret, Mar. 2002.

[22] E. Tabassi, *NFIQ Compliance Test, NISTIR 7300,* Nat'l Inst. of Standards and Technology, http://fingerprint.nist.gov/NFIQ. 2006.

[23] A. Martin, G.R. Doddington, T. Kamm, M. Ordowski, and M.A. Przybocki, "The DET Curve in Assessment of Detection Task Performance," *Proc. Eurospeech,* pp. 1895-1898, 1997.

[24] A.J. Mansfield and J.L. Wayman, "Best Practices in Testing and Reporting Performance of Biometric Devices," Report CMSC 14/02, Nat'l Physics Laboratory, Aug. 2002, http://www.cesg.gov.uk/site/ast/biometrics/media/BestPractice.pdf.

[25] J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey, *Graphical Methods for Data Analysis,* p. 62, Wadsworth and Brooks/Cole, 1983.

Patrick Grother is a staff scientist at the US National Institute of Standards in Technology (NIST) responsible for biometrics testing, standards development, and analysis. He is currently involved in the development of government, US and international standards for the Personal Identity Verification program, performance and interoperability testing, and data interchange formats for biometric data and support of fusion processes. At NIST, he is involved in various ongoing biometric performance assessments, an activity for which he received a US Department of Commerce Gold Medal in 2003. Educated at Imperial College, London, he is interested in biometric algorithms, evaluation, fusion, image quality, pattern recognition, data mining, and image processing. He has published papers most recently on biometric interoperability, fusion, testing of sample quality algorithms, large population identification performance, and 3D face, fingerprint, and gait recognition. He is a member of the IEEE and the IEEE Computer Society.

Elham Tabassi received the MS degree in electrical engineering from Santa Clara University in 1994. She is a staff member at the US National Institute of Standards and Technology (NIST) working on various biometric research projects including biometric sample quality, fusion, and performance assessment. She developed the NIST Fingerprint Image Quality (NFIQ), which has won national and international acceptance and has become a de facto standard, it is included in the Electronic Fingerprint Transmission Specification (EFTS), which is a required standard for doing business with the FBI. She received the US Department of Commerce Gold Medal in 2003 for her work on biometric system performance assessment. Her research interests are in biometrics technology, biometric sample quality, pattern recognition, data mining, and signal processing. She is a member of the IEEE and the IEEE Computer Society.