Quality Summarization

Recommendations on Biometric Quality Summarization across the Application Domain

NISTIR 7422

Elham Tabassi
Patrick Grother

National Institute of Standards and Technology

May 3, 2007

1 Purpose and Scope

The purpose of this document is to recommend procedures for the appropriate aggregation of quality values over a collection of samples, e.g. enterprise-wide summarization. It assumes that the concept-of-operations mandates computation of quality values for enrollment and/or verification samples, and that there is a need for application specific summarization of those values. The result is a summary value which supports monitoring of quality. Quality summarization should be performed across similar usage, e.g. quality summarization over all enrollment samples of an enterprise, or quality summarization over all verification samples of an enterprise. In operations where users frequently interact with a biometric system (e.g. time and attendance applications), quality values may be aggregated on a per user basis. This will reveal the existence of indviduals that consistently yield low quality samples.

2 Performance Related Quality Monitoring

A biometric quality assessment method (BQAM) derives a numerical quality value from an input biometric sample. The quality value is related to the biometric error rates that are likely to be realized when the sample is matched. Even if this predictive function is imperfect it is likely to be valuable nevertheless. Two factors mitigate against the direct use of such numbers:

- the interpretation of the value is specific to the BQAM[1]
- the recognition error rates are usually nonlinearly dependent on the quality values.

Nonetheless, there is a need to summarize quality values computed across all retained samples in an enterprise into a single quality value representing the overall quality of the enterprise. Quality summarizaton supports monitoring

- over time (to expose seasonal variation, or trends),
- for each sensor (to identify defective devices),
- at each site (to identify problem locations)
- of officials or attendants (to assess adherence to operating procedures), and
- per user basis (to identify users that consistently yield low quality samples).

In each case the quality summaries can be used to identify departures from the application specific historical norms, or design targets. These valuable uses of biometric quality assessment algorithms prompt the following two recommendations.

Recommendation 1 - Quality values should be computed across all retained samples in an enterprise.

This may be done online or offline. This will depend on factors such as:

- the computational cost of BQAM execution during enrollment or verification,
- whether or not the samples are retained (in verification, they may not be),
- whether the matching scores or decisions themselves constitute a reportable operational performance measure,
- the timescale for production of quality summaries.

Once values have been collected in a central location, these should be aggregated. In section 3, we show that it is generally not sufficient to simply average those values.

[1] Although there is consensus within the ISO/IEC JTC 1 SubCommitee 37 on Biometrics that quality values should be predictive of performance, the standards to date only make the qualitative associations Poor [1, 25], Average [26, 50], Good [51, 75] and Excellent [76, 100].

Recommendation 2 - The provider of a quality assessment algorithm should supply a function to aggregate values into a summary statistic.

Furthermore, we recommended that such functions compute quality summaries on the standardized range of biometric sample quality values as specified in ISO/IEC 19784-1 BioAPI [3], which requires single sample quality values on $[0, 100]$.

Recommendation 3 - Quality summary statistic of a BQAM should be on the range $[0, 100]$.

The recommended procedure for National Institute of Standards and Technology (NIST) Fingerprint Image Quality NFIQ [1, 2] is given in the next section. This kind of quality aggregation applied here to NFIQ may be appropriate for other quality measures. However, this paper does not prescribe any particular functional form, and developers are free to use any appropriate method. Indeed, we anticipate (and encourage) that such methods will remain the private intellectual property of the provider. The quality summarization function could be the result of a BQAM calibration process conducted by the provider, by a third party laboratory, or by the deploying organization *in-situ*[2].

Recommendation 4 - For verification applications, quality summarization functions should weight the native quality values to reflect mean expected false non-match rate (FNMR).

3 Recommendations for NFIQ

In an operation where fingerprint images are collected and their NFIQ values are computed the overall quality of the collection is given by:

$$\tilde{Q} = 102.75 - 2.75p_1 - 5.37p_2 - 14.38p_3 - 42.25p_4 - 102.75p_5 \tag{1}$$

where p_i is the proportion of the fingerprints with quality value $i = 1 \dots 5$. The weights were determined using the method of Appendix A and they reflect the likelihood that an observed false non-match involved a fingerprint of quality i. As shown there, the poorest quality category, 5, is responsible for better than half the observed error rates. Therefore, the primary recommendation of this document is:

Recommendation 5 - In verification applications users of NFIQ should apply equation (1) for summarization.

The terms of equation 1 indicate that the errors are dominated by images with NFIQ values 4 and 5, and this implies that a plain averaging of observed values is not an appropriate summary. Thus:

Recommendation 6 - Users of NFIQ should not use the mean or median of a set of quality values as a summary statistic.

Equation 1 produces a NFIQ summary on the range $[0, 100]$. This is achieved by a transformation of a simpler linear quantity (see the development in Appendix A). It is used here to allow standardized range of biometric sample quality values; mainly in keeping with the ISO/IEC 19784-1 BioAPI [3] requirement for single sample quality values on $[0, 100]$.

[2] A representative set of (mated) samples and one or more matching algorithms will be needed for calibration.

3.1 Dependence on Matching Algorithm

Weights in equation 1 are consensus estimates. That is they were estimated using the observed false non-match rates from a set of leading commercial matching algorithms. The result is that the weights are not exactly the weights that would be used for any one algorithm, or for a specified set of algorithms. NIST regards the NFIQ weights above as Best Practice estimates to be used unless other details about the application are known. Thus, we make the following recommendation.

Recommendation 7 - In verification applications, where a specific set of one or more matching algorithms are known and available, users of NFIQ fingerprint quality assessment algorithm should follow the procedure in Appendix A to establish dedicated weights.

3.2 Dependence on Operating Threshold

Weights in equation 1 are estimates of the observed false non-match rates computed at some fixed threshold. The result is that these weights are most accurate for that particular threshold and not as accurate for biometric systems operating at other thresholds. Figure 1 shows the variation of these weights computed at three different thresholds. It appears that weights for NFIQ values of 1 and 2 are quite robust to wide range of thresholds, but weight for NFIQ value 5 varies with threshold. Table 1 shows recommendation for NFIQ summarization at several operation threshold. NIST regards these recommendations as Best Practice estimates and these should be used unless other details about the application are known. Thus, we make the following recommendation.

Recommendation 8 - In verification applications, where operating threshold is fixed at τ, users of NFIQ fingerprint quality assessment algorithm should either use the weights computed at threshold closest to τ (as shown in Table 1) or follow the procedure in Appendix A to establish dedicated weights.

4 Summary

This document provides technical guidance for users of biometric quality algorithms in large-scale enterprise operations. Specifically, it recommends the computation and use of performance-related quality summaries, which, for verification, should serve as measures of the overall expected false non-match rate. As an example, Equation 1 provides a Best Practice estimate for the NFIQ algorithm for those verification applications in which the specific matchers and operating thresholds are unknown. For those operations where such details are known, users of BQAMs (including NFIQ) should follow the procedure in Appendix A to compute dedicated weights tailored to the application.

References

[1] E. Tabassi, C. L. Wilson, and C. Watson *Fingerprint Image Quality, NFIQ*, NISTIR 7151 ed., National Institute of Standards and Technology, 2004.

[2] E. Tabassi, C. L. Wilson, "A novel approach to fingerprint image quality" in *IEEE International Conference on Image Processing ICIP-05*, vol. 2, pp. 37-40, 2005.

[3] ISO/IEC JTC1 / SC37 / Working Group 2, *ISO/IEC 19784-1 Information Technology - Biometric Application Programming Interface - Part 1: BioAPI*, 2006, http://isotc.iso.org/isotcportal.

Table 1: Recommendation for NFIQ summarization at different operating thresholds

False Match Rate	**Recommendation for NFIQ summarization**
0.01	101.91 - 1.91p_1 - 3.97p_2 - 10.24p_3 - 34.03p_4 - 101.91p_5
0.001	102.75 - 2.75p_1 - 5.37p_2 - 14.38p_3 - 42.25p_4 - 102.75p_5
0.0001	105.41 - 5.41p_1 - 9.15p_2 - 23.82p_3 - 55.81p_4 - 105.41p_5

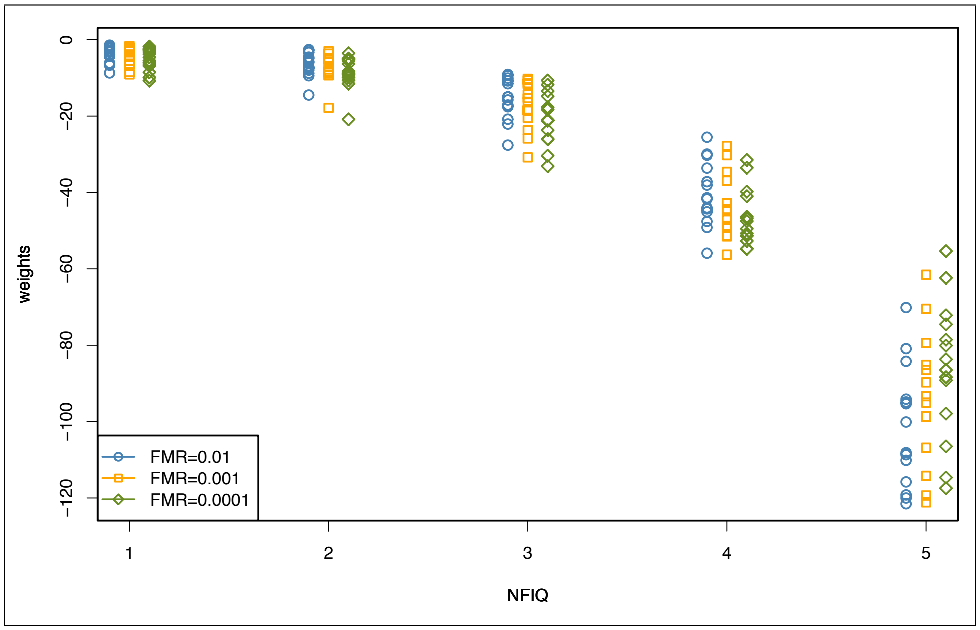

Figure 1: Dependance of NFIQ weights on operating threshold. Weights for NFIQ values 1, and 2 are quite robust to variation of the computing threshold. Thresholds are set at overall false-match-rates of 0.01, 0.001, and 0.0001. Each point corresponds to NFIQ weight estimated using similarity scores of a commercial matching algorithm on large operational fingerprint datasets. NFIQ weights in Table 1 are means of six matching algorithms with the highest performance.

Appendix

A Determination of Quality Weights

This section advances a procedure for assigning weights to the output values of a BQAM. We assume quality values are quantized into L levels so that (without loss of generality) $q = 1 \dots L$, where $q = 1$ and $q = L$ indicate lowest and highest quality values respectively. This is the case with NFIQ for which $L = 5$ and other commercial BQAMs for which $L = 8$ and $L = 10$. The strategy is to assign weights u_q that are directly related to the error rate observed for samples of quality q.

Suppose some enterprise collects fingerprints and measures the quality of each. If the number of prints collected over some interval in an operational situation is n and this is composed of n_q prints of quality q then we could compute the mean quality across all n samples. However, arithmetic mean is not the preferred method of summarizing quality scores because all samples regardless of their quality values are given the same weight. If instead the expected utility of a fingerprint of quality q is $u_q = U(q)$, then a better summary statement of quality is

$$\bar{q} = \frac{\sum_{q=1}^{L} u_q n_q}{\sum_{q=1}^{L} n_q} \tag{2}$$

If the utility u_q is actually an estimate of the false reject rate for samples of quality q of a reference fingerprint verification system operating at some reasonable threshold then $\bar{q}$ will be an estimate of the expected error rate. We proceed by introducing a procedure to compute utility u_q for different levels of a BQAM such that the summarized quality value is an estimate of the expected error rate.

Consider a biometric corpus contains $2N$ pairs of images from N persons. The first sample represents an enrollment sample, and the second represents the authentication sample. The samples have integer qualities $q_j^{(1)}$ and $q_j^{(2)}$ for $j = 1, \dots, N$. Applying V matching algorithms to the samples, we get

- ▷ N genuine similarity scores, $s_{jj}^{(v)}$, and
- ▷ up to $N(N-1)$ impostor scores, $s_{jk}^{(v)}$ with $j \neq k$

where $v = 1, \dots, V$ and $V \geq 1$.

1. For all matching algorithms v and quality values q compute FNMR$^v(\tau, i)$ of authentication samples of quality i with enrollment samples of quality better than or equal to i at operating threshold τ using genuine scores of matching algorithm v. Note that we assumed higher quality values indicate better quality. For BQAMs which low values indicate good quality (for example, NFIQ), $q_j^{(1)} \leq i,\ q_j^{(2)} = i$ should replace $q_j^{(1)} \geq i,\ q_j^{(2)} = i$ in the computation of FNMR$^v(\tau, i)$ below.

```
for   $(v = 1, \dots, V)$
    for   $(i = 1, \dots, L)$
        $\mathrm{FNMR}^v(\tau, i) = \frac{\left|\left\{s_{jj}^{(v)} :\ s_{jj} \leq \tau,\ \ q_j^{(1)} \geq i,\ q_j^{(2)} = i\right\}\right|}{\left|\left\{s_{jj}^{(v)} :\ s_{jj} \leq \infty,\ q_j^{(1)} \geq i,\ q_j^{(2)} = i\right\}\right|}$
    end
end
```

which results in the following array

$$\begin{pmatrix} \mathrm{FNMR}^1(\tau,1) & \mathrm{FNMR}^2(\tau,1) & \dots & \mathrm{FNMR}^V(\tau,1) \\ \mathrm{FNMR}^1(\tau,2) & \mathrm{FNMR}^2(\tau,2) & \dots & \mathrm{FNMR}^V(\tau,2) \\ \dots & \dots & \dots & \dots \\ \mathrm{FNMR}^1(\tau,L) & \mathrm{FNMR}^2(\tau,L) & \dots & \mathrm{FNMR}^V(\tau,L) \end{pmatrix}$$

2. compute weight u_i

$$u_i = \frac{\sum_{v=1}^{V} \text{FNMR}^v(\tau, i)}{\sum_{q=1}^{L} \sum_{v=1}^{V} \text{FNMR}^v(\tau, q)}$$

Thus the aggregated quality across an enterprise is

$$Q = \sum_{i=1}^{L} u_i p_i \tag{3}$$

where u_i are estimated posterior probabilities above. As probabilities, these values will not be on a range familiar to users. For example, the NFIQ summary is

$$Q = 0.016p_1 + 0.032p_2 + 0.086p_3 + 0.252p_4 + 0.613p_5 \tag{4}$$

such that if all samples were of NFIQ $= 1$ (i.e. the best quality) the result would be $Q = u_1 = 0.016$. Similarly the worst case is when all samples in the enterprise are of NFIQ $= 5$, which results in $Q = u_5 = 0.613$. Thus this formulation would result in NFIQ summaries on the range $[u_1, u_5]$, which is $[0.016, 0.613]$. Users should regard equation 4 as a measure of expected overall FNMR . However, this document recommends transformation from $[u_1, u_5]$ to the more familiar BioAPI [3] range $[0, 100]$ which has 0 as the lowest quality and 100 as the best. This can be accomplished by:

1. Either relating the quality summary number Q (i.e. expected error rate) back to the native quality range by using the inverse of the utility function:

$$\tilde{Q} = U^{-1}(Q) = U^{-1}\left(\sum_{i=1}^{L} u_i p_i\right) \tag{5}$$

   where U^{-1} is a function approximation (e.g. piece-wise linear interpolation) of pairs (i, u_i),

2. Or by mapping (e.g. linear mapping) $[u_1, u_5]$ to $[0, 100]$. Thus, NFIQ summaries mapped to $[0, 100]$ are given by

$$\tilde{Q} = \frac{100u_5}{u_5 - u_1} - \sum_{i=1}^{5} \frac{100u_i}{u_5 - u_1} p_i \tag{6}$$

   which forms equation 1 in this document. Table 1 shows recommendation for NFIQ summarization at different thresholds where utility $u_i \ i = 1 \ldots 5$ (i.e. five levels of NFIQ) is computed at different operating thresholds, and linearly mapped to $[0, 100]$ using equation 6.