WHAT DO QUALITY MEASURES PREDICT IN BIOMETRICS?

Krzysztof Kryszczuk [†] *and Andrzej Drygajlo* [††]

[†] IBM Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland
email: kkr@zurich.ibm.com
[††] Swiss Federal Institute of Technology Lausanne (EPFL), STI-IEL-LIDIAP, CH-1015 Lausanne, Switzerland
email: andrzej.drygajlo@epfl.ch

ABSTRACT

This paper is discusses the role of quality measures in biometric classification. We challenge a common notion that quality measures are performance predictors of the baseline biometric classifier. Instead, we postulate that quality measures are class-independent classification features, and as such are conditionally relevant class predictors. We present a systematic, probabilistic approach towards error prediction in biometric classification systems, where quality measures play an integral role in a stacked classifier ensemble. We demonstrate the results of error prediction in face verification using the proposed method.

1. INTRODUCTION

Quality of biometric signals and their measurement have recently become a common research topic of many academic and industrial laboratories. The reason for this is that existing large scale deployments of biometric technologies have systematically encountered the problem of inconsistent data quality. The interest in biometric data quality is reflected in the fact that entire scientific workshops have been dedicated to this topic[1]. Biometric conferences have recently attracted a significant volume of research papers on incorporating quality measures into biometric classification. One of the most important questions about the role of quality information is - what can it really be used for?

A prevailing notion is an intuitive one - that quality information predicts classification performance of the baseline classifier. Hsu et al. [1] postulate that the quality measures for face images ought to be predictors of classification performance. Chen et al. [2] propose a quality measure that is explicitly designed to predict performance of an automatic fingerprint matcher. Capelli et al. [3] designed a series of experiments that elicit the relationship between fingerprint quality aspects (acquisition area, resolution accuracy, geometric distortions) and average recognition performance of four different fingerprint recognition algorithms. They showed a non-linear, monotonic relationship between the quality measures and error rates. Ko and Krishnan [4] provide arguments that "better captured fingerprint image quality will have better match accuracy". They also express their conviction that quality metrics are good predictors of fingerprint matching performance. Grother and Tabassi [5] present an extensive study of a popular quality measure algorithm, NFIQ [6]. The authors evaluate the algorithm by analyzing how well it predicts the classification performance of the NFIS fingerprint matching software. Alonso-Fernandez et al. [7] also focus on fingerprint quality metrics. They evaluate a series of quality assessment algorithms designed for fingerprint images. The authors follow the approach of Grother and Tabassi [5] of applying the performance prediction criterion when comparing the algorithms. The notion of a performance-predicting function of quality measures in biometrics seems to be widely accepted.

It is well-known that quality degradation of observed biometric signals may lead to a degradation of classification performance. This effect can be attributed to the fact that a low quality biometric signal contains less *biometric information* than a high quality one [8]. If such a simple relationship is universally present in biometric systems then it would indeed seem like a straightforward conclusion to assume that a quality measure must be a predictor of performance. Here is the problem: the concepts of "high" and "low" quality are at best vague. What is the criterion that allows one to tell a "low quality" signal from a "high quality" one? Human judgement of image quality has often little to do with the discriminative biometric content of the signal [9] and is linked to the perceived signal fidelity. Therefore, if we discard any definitions that involve human judgement, the only criterion remaining is the actual classification performance. But if classification performance is a criterion for labeling a signal's quality, then the argument that quality measures predict performance is clearly circular.

Youmaran and Adler [8] propose to describe the relative change of signal quality in terms of differences in divergence of classification feature distributions for signals of nominal and degraded quality. That is a sound idea, but since the authors employ the very same features for deriving their quality estimator as used for classification (for instance, PCA features for face recognition), this approach is equivalent to looking at the baseline classifier confidence, clad in an information-theoretic wrap-up. As such, their quality measure predicts classifier performance as much as a confidence measure otherwise would.

In order to understand what, if anything at all, quality measures are capable of predicting, it is necessary to formulate the actual role that quality measures can play in a biometric classification system. In our previous work we have defined quality measures as auxiliary features, used together with baseline classifier scores in $Q-stack$, a stacking-based hierarchical classifier ensemble [10]. This generalized formulation has been demonstrated to encompass previous heuristic methods of classification with quality measures in single-, and multiple-classifier scenarios, including

The research reported in this paper was conducted when K. Kryszczuk was with the Swiss Federal Institute of Technology Lausanne (EPFL).

[1] Biometric Sample Quality Workshops organized by the NIST in 2006 and 2007.

multimodal biometric classification.

In this paper we build on this framework to specifically address the issue of predictive capabilities of quality measures. In particular, we substantiate claims that in the general case, quality measures are not predictors of performance of the baseline biometric classifier. Instead, we show that quality measures, as conditionally-relevant classification features, can be used to predict the accuracy of classification decisions of a second-level Bayesian classifier in a stacking architecture of $Q-stack$. We substantiate this claim using a face matching example.

The rest of the paper consists of an overview of existing intuitive notions of quality and quality measures and their role in biometric classification in Section 2, a summary of the $Q-stack$ framework, with its extension to error prediction in Section 3, and a theoretical discussion of the predictive function of quality measures in Section 4, and an evaluation using a face matching example in Section 5. Section 6 concludes the paper.

2. INTUITIVE NOTIONS OF QUALITY AND QUALITY MEASURES

Figure 1 shows a currently predominant notion of the role of quality measures in biometric classification systems [1, 2, 3, 5, 4].

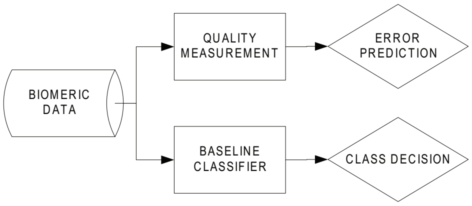


Figure 1: A commonly accepted structure of a biometric classification system with quality measures. Quality measures are used to directly predict the errors of the baseline classifier

In Figure 1 an arbitrary biometric signal is classified using a dedicated baseline classifier. The classification decision is taken based on the output of the baseline classifier. In parallel, the quality of the classified biometric signals is measured. Based on so obtained quality measures, authors cited in Section 1 postulate to predict the performance of the baseline classifier or to predict the individual classification errors. A central concept that surfaces in an overwhelming majority of papers is the notion that biometric signals of *high quality* are more likely to be correctly classified than those of *low quality*. From here the intuitive understanding of the role quality measurement is straightforward. After the biometric signal is recorded, its quality is measured. If the measured quality fails to satisfy some preset criterion, it in a way *predicts* that the baseline classifier is likely to commit a classification error.

Despite its deceiving clarity, this notion of performance *prediction* using quality measures suffers from a logical weakness, a direct consequence of the circular definition discussed in Section 1. We show the consequences of this weakness in two illustrative counterexamples, in Section 4.

3. $Q-STACK$: A SYSTEMATIC FRAMEWORK OF CLASSIFICATION WITH QUALITY MEASURES

Figure 2 shows a diagram of the $Q-stack$ architecture. Identity-related information is composed of n biometric signals $\mathbf{s} = [s_1, s_2, ..., s_n]$, classified by n baseline classifiers, resulting in a score vector $\mathbf{x} = [x_1, x_2, ..., x_n]$. At the same time, the signals undergo quality measurements, resulting in m quality signals $\mathbf{qm} = [qm_1, qm_2, ...qm_m]$. In general, $n \neq m$ is permitted, since one quality measure can be pertinent to multiple signals. Vice-versa, multiple quality measures can be used to characterize one signal. The score vector is concatenated with the quality measure vector to form an evidence vector $\mathbf{e} = [\mathbf{x}, \mathbf{qm}]$. The evidence vector $\mathbf{e}$ becomes a feature vector for the stacked classifier. The architecture presented in Figure 2 is a generalization of single- and multiple-classifier systems, including multimodal approaches. If no quality measures are present, the architecture shown in Figure 2 performs a multimodal score-level fusion.

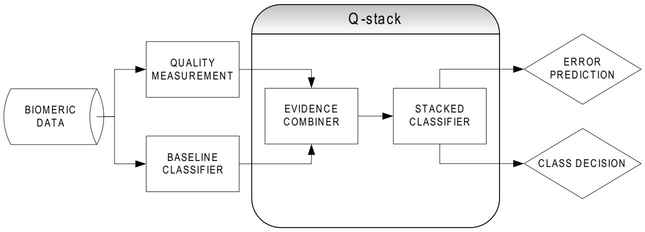


Figure 2: $Q-stack$ architecture, in which baseline classifier scores and quality measures jointly serve as features to a second-level, stacked classifier. Confidence measures of the stacked classifier can be used to predict errors of the stacked classifier.

The proposed method of $Q-stack$ is a generalized framework which encompasses previously reported methods of using quality measures in biometric verification [10]. In particular, previously reported methods can be shown to be case- and data-specific, largely heuristic approximations of an optimal decision boundary in the evidence space. As opposed to the ad-hoc methods which can hardly be generalized to new data sets and classifier architectures, $Q-stack$ attempts to approach the optimal decision function by learning the causal dependencies between quality information and baseline classifier scores from available data.

As shown in Figure 2, in the scheme of $Q-stack$ the error prediction is performed using the output of the stacked classifier. In the following Section 4 we will explain why this is a better and theoretically justified manner of performing error prediction. For a detailed discussion of $Q-stack$, a classifier-stacking approach to classification with quality measures the reader is referred to [10].

4. QUALITY MEASURES AND ERROR PREDICTION

4.1 What do quality measures not predict?

The prevailing notion of the role of quality information is that it predicts the errors of the baseline classifier, as shown in Figure 1. In this section we show why this notion is incorrect. In the classifier ensemble architecture shown in Figure 2 the final classification decision is taken based on the output (score) of the second-level, stacked classifier. Let us assume without a loss of generality [2] that the classification system is devised to assign class labels A and B to observation x. Under the assumption of Bayesian optimality of the baseline classifier, the predicted classification error, and thus the classifier performance, can be expressed as a continuous and monotonic function f of overlap D between distributions of observations x, given their actual class alignment and respective quality measure,

$$ER(qm_i) = f(D(p(\mathbf{x}|A, qm_i), p(\mathbf{x}|B, qm_i))). \qquad (1)$$

The choice of an overlap measure D is immaterial, for instance the likelihood difference $D(p(\mathbf{x}|A, qm_i), p(\mathbf{x}|B, qm_i)) = |p(\mathbf{x}|A, qm_i) - p(\mathbf{x}|B, qm_i)|$ can be used. If the quality measure is a performance predictor then the relationship between the quality measure and the error measure must have a functional character $qm \mapsto ER(qm)$: each possible value of qm is assigned one and only one $ER(qm)$. In order to show that this condition is not necessarily satisfied, we present an example where for one value of qm the error measure given by Equation 1 has more than one value,

$$\exists_{qm_i, k \neq l} : (ER(qm_i) = k) \wedge (ER(qm_i) = l). \qquad (2)$$

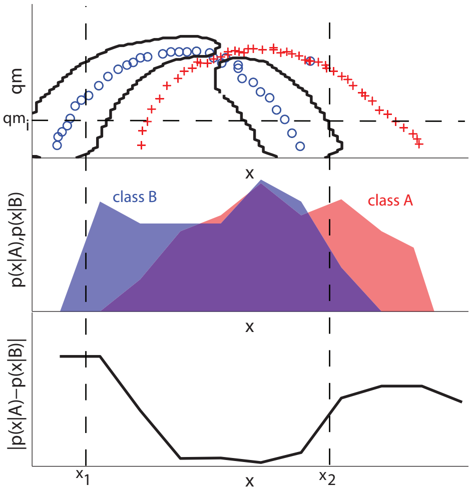


Figure 3: A toy embodiment of the condition given by Equation 2 - there exists a value of qm_i that predicts two different values of $|p(\mathbf{x}|A) - p(\mathbf{x}|B)|$, and consequently of $ER(qm_i)$.

A schematic of such situation is given in Figure 3. One could argue that if the quality measure behaves as in Figure 3 then it is not a *good* quality measure. But such argument is strikingly similar to the circular argument discussed in Section 1: a *good* quality measure must predict performance and therefore a quality measure that does not predict performance is NOT a *good* quality measure. The fact is that the quality measure qm from the example shown in Figure 3 helps reach better class separation than can be achieved in the domain of x alone:

$$D(p(\mathbf{x}|A), p(\mathbf{x}|B)) > D(p(\mathbf{x}, qm_i|A), p(\mathbf{x}, qm_i|B)). \qquad (3)$$

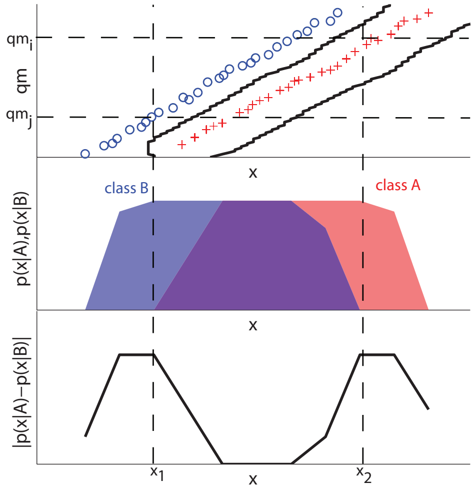


Figure 4: A toy example, where different $qm_i \neq qm_j$ predict the same baseline classifier performance $ER(qm_i) = ER(qm_j)$. The class-independent quality information qm allows for much better class separation than using the baseline classifier alone.

Another illustrative example of the fallacy of considering quality measures as performance predictors of the baseline classifier is shown in Figure 4. Here, the dependence between x and qm is straightforward - close to linear. Two values of $qm_i \neq qm_j$ predict the same baseline classification performance $ER(qm_i) = ER(qm_j)$, which, according to the intuitive definition cited in Section 1, renders this quality measure useless. Yet, in the classification problem shown in Figure 4 the quality measure allows for a clearly better class separation in the evidence space jointly defined by x and qm, than using baseline scores x alone.

Obviously, this makes quality measure qm a useful *classification feature*. Since by default $p(qm|A) = p(qm|B)$ is admissible and actually observed in practice [10], qm becomes a *conditionally relevant feature*. Classification features are not performance predictors, they are *class predictors* [12]. This makes biometric quality measures *conditionally relevant class predictors*.

4.2 What do quality measures actually predict?

The distributions given in Figures 3 and 4 are synthetic examples that demonstrate the fallacy of the statement that the quality measures predict classification performance. Admittedly such distributions are not frequently observed in practice, and one could argue that for practical, engineering purposes quality measures could be used as a coarse *predictor of performance*. In this section we show that even if that is the case, it is not the baseline classifier whose performance one should care to predict.

Compare again the diagrams in Figures 1 and 2. In both figures quality measures play a role in error prediction, but in the former case, it is the errors of the baseline classifier that are being predicted. In the latter case, one strives to predict the performance of the second-layer, stacked classifier. As demonstrated in [10], an appropriately chosen stacked classifier can offer an average classification accuracy superior to that of the baseline classifier, given that the quality measures are statistically dependent on the baseline classifier scores.

[2] Multi-class classifiers can be represented as a combination of dichotomizers [11].

Therefore it is this classifier's errors that should be predicted.

Since in the classification ensemble shown in Figure 2 both quality measures *qm* and baseline classifier scores *x* are features to the stacked classifier, its errors can be predicted in a probabilistic manner, by computing the probability of error given available evidence. If given observation s_i is assigned a class label $\omega_i \in \Omega = [A,B]$ by the stacked classifier using evidence vector $e_i = [x_i, qm_i]$, the probability of correct classification can be expressed by the Bayes rule:

$$P(\omega|\mathbf{x}_i, qm_i) = \frac{p(\mathbf{x}_i, qm_i|\omega)P(\omega)}{\sum_\omega p(\mathbf{x}_i, qm_i|\omega)P(\omega))}. \tag{4}$$

Here, $P(\omega|\mathbf{x}_i, qm_i)$ is a subjective Bayesian degree of belief (credence) in the correctness of a single classifier decision [13].

5. PERFORMANCE PREDICTION WITH QUALITY MEASURES: EXPERIMENTAL EVALUATION

The experiments reported in this section demonstrate that quality measures, as conditionally relevant classification features in the $Q-stack$ architecture, can serve as conditionally-relevant class predictors and consequently help predict the performance of the stacked classifier. For the purpose of this demonstration, we used a face matching experiment, conducted using face images of 200 subjects (training set: 50 subjects, testing set: 150 subjects) from the BioSec database, baseline corpus [14]. The experiments involved one-to-one sample matching. Face matching is a hard classification task with relatively high observed error rates, hence it provided sufficient errors to show the effectiveness of the presented error prediction paradigm. All face images were cropped manually in order to avoid the impact of face localization algorithms on the matching performance. All images were photometrically normalized [15].

In our experiments we used the following two face matchers: 1. *DCT* - local *DCTmod*2 features and a Bayes classifier based on feature distributions approximated by Gaussian Mixture Models (GMM)[16]: scores produced by the *DCT* matcher denoted as x_{f1}, and 2. *PCA* - Mahalanobis distance between global *PCA* feature vectors. The *PCA* projection space was found using all images from the development dataset. The scores produced by the *PCA* matcher are denoted as x_{f2}. The two face matchers were chosen because they both operate on very different features. The local *DCTmod*2 features encode mostly high spacial frequencies, while the projection of the face images on the *PCA* subspace emphasizes lower spacial frequencies.

In the experiments reported here we used two face image quality measures - a normalized two-dimensional cross-correlation coefficient with an average face template, denoted as qm_{f1}, and a probabilistic quality measure which evaluates how well do the used classification models account for the observed data. This quality measure is based on the *DCTmod*2 classification features, and denoted as qm_{f2}. For details on these quality measures the reader is referred to [10].

The experiments reported here were conducted using a classifier ensemble shown in Figure 2. A Bayesian classifier with GMM class models was deployed as the stacked classifier. Sample evidence distributions and decision hypersurfaces created by stacked classifiers are shown in Figure 5. The estimates of posterior probabilities of the stacked classifiers were used as error predictors for single classifier decisions. In order to show that the presence of quality measures allows for precise prediction of the errors of the stacked rather than the baseline classifier, several experiments with, and without the use of quality measures were conducted. In these experiments the baseline classifiers (*DCT* and *PCA*) are used separately, and together in a multi-classifier scenario. For each experiment the actual and predicted error rates were recorded, where the predicted error rates $\overline{R}$ were computed as

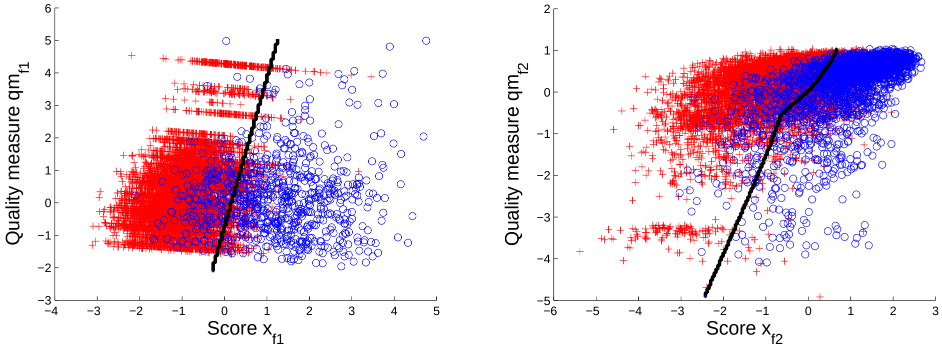
	Average	Imposter (class A)	Genuine (class B)
$\mathbf{e} = [x_{f1}]$			
ER	0.146	0.149	0.142
$\overline{R}$	0.153	0.192	0.113
δ	**-0.007**	**-0.043**	**0.029**
$\mathbf{e} = [x_{f1}, qm_{f1}]$			
ER	0.132	0.149	0.115
$\overline{R}$	0.130	0.145	0.116
δ	**0.002**	**0.005**	**-0.001**
$\mathbf{e} = [x_{f2}]$			
ER	0.272	0.209	0.334
$\overline{R}$	0.253	0.288	0.217
δ	**0.019**	**-0.079**	**0.117**
$\mathbf{e} = [x_{f2}, qm_{f1}]$			
ER	0.220	0.176	0.264
$\overline{R}$	0.184	0.192	0.176
δ	**0.036**	**-0.016**	**0.088**
$\mathbf{e} = [x_{f2}, qm_{f2}]$			
ER	0.220	0.176	0.264
$\overline{R}$	0.184	0.192	0.176
δ	**0.036**	**-0.016**	**0.088**
$\mathbf{e} = [x_{f1}, x_{f2}]$			
ER	0.127	0.127	0.127
$\overline{R}$	0.124	0.150	0.097
δ	**0.003**	**-0.024**	**0.031**
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$			
ER	0.117	0.118	0.116
$\overline{R}$	0.107	0.119	0.094
δ	**0.010**	**-0.001**	**0.022**

Table 1: Evaluation of credence estimates using the accountability criterion. The mean difference between actual observed error and the mean credence estimates after 100 experimental iterations is given by δ

$$\overline{R} = \frac{1}{I_\omega} \sum_{I_\omega} (1 - P(\omega|\mathbf{e}_i)), \tag{5}$$

where I_ω is the actual number of samples in the evaluation data set, per class ω, and e_i is an evidence vector containing either scores alone, or scores and quality measures. The combinations of scores and quality measures used in the reported experiments are given in Table 1. If the estimates of $P(\omega|\mathbf{e}_i)$ are accurate, then on average taken over a sufficiently large data set, the *a priori* estimates of error probabilities of individual classifier decisions must account for the actual error rates ER, observed *a posteriori*. This notion,referred to as *accountability criterion* for estimating error prediction accuracy [13], can be expressed in terms of $\delta = ER - \overline{R}$. The values of ER, $\overline{R}$ and δ are reported per class, and as a total average. The results of the experiments are given in Table 1.

The baseline classifier error rates are given by evidence combinations $\mathbf{e} = [x_{f1}]$, $\mathbf{e} = [x_{f2}]$ and $\mathbf{e} = [x_{f1}, x_{f2}]$. The results gathered in Table 1 clearly show that quality measures can be effectively used for improving the classification performance in $Q - stack$ architecture, as shown in Figure 2, with respect to the baseline classifier results. For all evidence configurations, the proposed error prediction method worked well and accurately predicted observed error rates with er-



(a) $\mathbf{e} = [x_{f1}, qm_{f1}]$

(b) $\mathbf{e} = [x_{f2}, qm_{f2}]$

Figure 5: Classification in the evidence space: sample distributions and stacked classifiers for face matching, Biosec database using (a) *DCT* and (b) *PCA* baseline classifiers and corresponding quality measures.

rors given by the corresponding δ. The values of δ are at least one order of magnitude smaller than the predicted error rates.

6. CONCLUSIONS

In this paper we have presented arguments against the commonly encountered, intuitive interpretation of quality measures as performance predictors in biometric classification. We have proposed an alternative, rigorous view on the role of quality measures in biometric classification and error prediction, in which the quality measures are conditionally relevant class predictors. We have proposed an error prediction architecture based on a stacked classifier ensemble, where a Bayesian classifier returns posterior probabilities used as credences in single decisions' correctness. We have instantiated the proposed error prediction system using a face verification example, using two different classifiers and two different quality measures.

REFERENCES

[1] R.-L. V. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *Proc. of the 2006 Biometrics Symposium*, Baltimore, MD, USA, 2006.

[2] Y. Chen, S. Dass, and A. Jain, "Fingerprint quality indices for predicting authentication performance," in *Proc. of Audio- and Video-based Biometric Person Authentication*, Rye Brook, NY, July 2005, pp. 160–170.

[3] R. Cappelli, M. Ferrara, and D. Maltoni, "The quality of fingerprint scanners and its impact on the accuracy of fingerprint recognition algorithms," in *Multimedia Content Representation, Classification and Security*, ser. LNCS, vol. 4105. Springer Verlag, 2006, pp. 0302–9743.

[4] T. Ko and R. Krishnan, "Monitoring and reporting of fingerprint image quality and match accuracy for a large user application," in *Proc of the 33rd Applied Imagery Pattern Recognition Workshop*, Washington, DC, USA, 2004.

[5] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, April 2007.

[6] E. Tabassi, C. Wilson, and C. Watson, "NIST fingerprint image quality," NIST, Tech. Rep. NISTIR 7151, August 2004.

[7] F. Alonso-Fernandez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "A review of schemes for fingerprint image quality computation," in *Proc. of COST 275 Workshop - Biometrics based recognition of people over the internet*, Hatfield, UK, 2005.

[8] R. Youmaran and A. Adler, "Measuring biometric sample quality in terms of biometric information," in *Proc. of the IEEE Biometrics Symposium*, Baltimore, Maryland, USA, September 2006.

[9] A. Adler and T. Dembinsky, "Human vs. automatic measurement of biometric sample quality," in *Proc. Canadian Conference on Computer and Electrical Engineering*, Ottawa, Canada, May 2006.

[10] K. Kryszczuk and A. Drygajlo, "Improving classification with class-independent quality measures: Q-stack in face verification," in *Proc. of the 2nd International Conference on Biometric ICB'07*, Seoul, Korea, 2007.

[11] D. M. J. Tax and R. P. W. Duin, "Using two-class classifiers for multiclass classification," in *Proc. 16th International Conference on Pattern Recognition*, vol. 2, Quebec, Canada, 2002, pp. 124– 127.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2001.

[13] K. Kryszczuk and A. Drygajlo, "Credence estimation and error prediction in biometric identity verification," *Signal Process.*, vol. 88, no. 4, pp. 916–925, 2008.

[14] J. Fierrez, J. Ortega-Garcia, D. Torre-Toledano, and J. Gonzalez-Rodriguez, "BioSec baseline corpus: A multimodal biometric database," *Pattern Recognition*, vol. 40, no. 4, pp. 1389–1392, April 2007.

[15] R. Gross and V. Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *Proc. of the 4th Intl. Conf. on Audio- and Video-Based Biometric Person Authentication*, Guilford, UK, 2003.

[16] C. Sanderson, "Automatic person verification using speech and face information," Ph.D. dissertation, Griffith University, Queensland, Australia, 2002.