Quality-based Score Normalisation with Device Qualitative Information for Multimodal Biometric Fusion

Norman Poh, Josef Kittler and Thirimachos Bourlai

***Abstract*— As biometric technology is rolled out on a larger scale, it will be a common scenario (known as cross-device matching) to have a template acquired by one biometric device used by another during testing. This requires a biometric system to work with different acquisition devices, an issue known as device interoperability. We further distinguish two sub-problems, depending on whether the device identity is known or unknown. In the latter case, we show that the device information can be probabilistically inferred given quality measures (e.g., image resolution) derived from the raw biometric data. By keeping the template unchanged, cross-device matching can result in significant degradation in performance. We propose to minimise this degradation by using device-specific quality-dependent score normalisation. In the context of fusion, after having normalised each device output independently, these outputs can be combined using the Naive Bayes principal. We have compared, and categorised several state-of-the-art quality-based score normalisation procedures, depending on how the relationship between quality measures and score is modelled, as follows: i) direct modelling, ii) modelling via the cluster index of quality measures, and iii) extending (ii) to further include the device information (device-specific cluster index). Experimental results carried out on the Biosecure DS2 data set show that the last approach can reduce both false acceptance and false rejection rates simultaneously. Furthermore, the compounded effect of normalising each system individually in multimodal fusion is a significant improvement in performance over the baseline fusion (without using any quality information) when the device information is given.**

I. INTRODUCTION

A. Interoperability of Biometric Devices

Person verification using biometrics such as face and fingerprint is becoming an important solution to border control and identity fraud [1]. As the biometric technology is being rolled out on a nation-wide scale, e.g., in the form of passport control, biometric devices may be replaced with newer designs (possibly from a different vendor), or old ones may be replaced with newer ones from the same vendor but having a different parameter configuration. For instance, in the US-VISIT (United States Visitor and Immigration Status Indicator Technology) program, an optical fingerprint sensor is being used during enrolment, but it is not guaranteed that the same type of sensor will be used elsewhere and indefinitely. In fact, the cost of re-enrolling individuals with the actual sensor to be deployed can be very high, making sensor interoperability an important practical requirement [2].

The National Biometric Security Project (NBSP) [1] proposed to improve device interoperability by standardisation efforts, which attempts to ensure that any two biometric devices are capable of producing raw images that can be processed by a matching subsystem. Two types of failure can occur: *failure-to-extract* features or *failure-to-match* two feature samples. Even though two devices may be interoperable (not producing any of the two types of failure just mentioned), matching the biometric raw data they produce (known as "cross-device" matching) can still result in *significantly* sub-optimal performance as compared with matching two biometric samples acquired from the same device ("same-device" matching). This has been demonstrated for at least three different biometric modalities by: Ross and Jain [2] using an optical and a solid-state fingerprint sensor, Malayath [3] using an electret and a carbon microphone (for the speech modality), and Alonso-Fernandez et al [4] using two Tablet PCs (for online signature). Our study using face images captured by a digital camera and a web camera (with significantly lower resolution and in adverse environment) also confirms the degradation in performance.

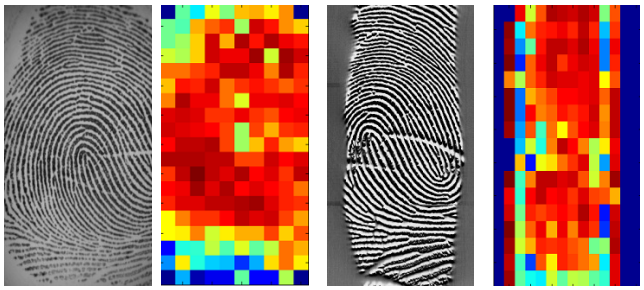
The solutions to biometric sensor interoperability can be grouped into at least three categories: data-level, feature-level and score-level calibration. In the data level calibration, the goal is to model the physics of the distortion process introduced by a sensor, in order to recover the actual (canonical) biometric representation. This can be done using a distortion compensation model, as briefly mentioned in [5]. A second solution is to model the relative distortion between images acquired using two different sensors, as proposed in [5]. In the feature-level calibration, the goal is to compute a common (feature) subspace given a pair of raw biometric samples acquired using two different biometric devices. An established statistical approach in machine learning to solve this problem is called canonical correlation analysis (CCA) [6]. Finally, the score-level calibration aims to map the output of different devices into a canonical score space so that a single decision threshold can be used rather than having to optimise one for each device [7]. In the context of fusion, as will be shown in this paper, score-level calibration has an important role.

There are four sub-problems that can be identified in addressing device interoperability, depending on the dichotomies of the following two assumptions: whether the actual acquisition device is known or not; and whether a *set of possible devices* is known or not. We shall address the problem where a set of devices is known, but further distinguish two sub-problems; namely whether the actual device used to acquire a biometric sample within the set may or may not be known. When the actual device is not known, we will show that it is possible to probabilistically infer this information from a set of quality measures derived from the raw biometric data.

Quality measures are an array of measurements quantifying the degree of excellence or conformance of biometric samples to some predefined criteria known to influence the system

The authors NP and JK are with CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK; and TB is with Biometric Center, West Virginia University Morgantown, WV 26506-6109, USA. normanpoh@ieee.org, J.Kittler@surrey.ac.uk, ThBourlai@mail.wvu.edu

[1] http://nationalbiometric.org/



(a) Optical sensor (b) Optical quality (c) Thermal sensor (d) Thermal quality

Fig. 1. Samples of two fingerprints acquired using two fingerprint sensors and their associated local quality maps [8].

performance. Examples of quality measures for the face (still) images are focus, contrast and face detection reliability [9] (see Figure 5). For fingerprint, local image gradients have been reported [8] (see Figure 1). Quality measures have been extensively used in multimodal biometric fusion [10], [11], [12], [13], [14], where the goal is to weigh the output of each biometric device such that biometric samples with higher quality contribute more to the final combined score. In our context, these methods are not used directly in fusion, but rather used as a quality-based score normalisation procedure, hence, considering fusion only after the individual biometric device output has been normalised.

Since the work presented here falls into the category of score-level calibration, only the related work in this domain will be covered in the next section.

B. Related Work in Quality-Based Score Normalisation

To the best of our knowledge, the issue of quality-based score normalisation is rarely discussed. Instead, the focus has always been dominated by the literature on quality-based fusion. We shall distinguish two categories in this literature: those that have the potential to be used as quality-based score normalisation [10], [14], [9], [15] and those that cannot, e.g., [11], [12], [13]. The reason that the latter category of algorithms are not applicable is that they have been designed specifically to consider the joint score space in such a way that it is not possible to factorise the approach for each device independently. We shall, therefore, review only the algorithms in the first category.

Quality-based score normalisation can be categorised according to how the relationship between match scores and quality measures is modelled, which can be: i) direct modelling; ii) modelling via cluster of quality measures, also referred to as *quality cluster* throughout this paper; and, iii) associating each device with a quality cluster. We will show that these methods can be considered in a more general Bayesian framework.

- **Direct modelling**: Nandakumar *et al.* [10] proposed a likelihood ratio-based approach to achieve quality dependent score fusion. This is a generative approach to model the relationship between scores and quality measures of the same modality. The likelihood of scores and quality measures of different biometric modalities are combined using the product rule, hence, realising a naive Bayes classifier. The result is that the less informative modalities will produce likelihood ratio close to one and will therefore not influence the final combined score.
  Kittler *et al.* [14] proposed a framework to incorporate the quality information in fusion from the pattern recognition perspective. In this framework, various levels of system output dependency, i.e., whether scores belong to the same modality or to different modalities, are considered. In the study, match scores are augmented by a vector of quality measures in such a way that the pairwise interaction between these two variables (in the sense of a tensor product) is modelled by a linear discriminative fusion function. It was first shown that quality-based fusion is non-linear with respect to the match scores.
- **Modelling via quality cluster** Poh *et al.* [9] proposed a generative approach to estimate the joint density of scores and quality measures by first clustering the quality measures into discrete hidden states (the quality clusters). This approach assumes that the scores and quality measures are independent given the discrete quality cluster. This approach is sensible because similar quality measures in a cluster will share similar statistical property (i.e., similar combination of factors, e.g., lighting condition, head pose and image resolution for face images) and thus they can be combined by the same fusion classifier, and vice versa for dissimilar quality measures.
  Maurer *et al.* [15] applied a Bayesian belief network to the problem of quality-based fusion. The novelty lies in modelling the density of match scores *conditioned* on the quantised quality measures, rather than the direct modelling approach (between match scores and quality measures). The subtle difference between the two approaches, as will be shown later, is that the former approach involves first of all a more simplified model (in terms of the assumption used) but is also technically difficult to implement when the quality measures are multi-dimensional. Maurer *et al.* 's approach falls into this category because by quantizing/binning the quality measure (which was assumed to be scalar in their case, rather than a vector), they effectively associate a quality measure with a quality cluster.
- **Modelling with device information** Instead of using the hidden quality cluster, Poh *et al.* [7] proposed to use a qualitative device characterisation. This study showed that it is possible to design a quality-based score normalisation procedure in order to mitigate the effect of device mismatch. The basic idea is to identify the device used to acquire a sample and then use a quality-normalisation strategy tailored to the device. The same concept was also independently demonstrated in [16] but in the context of multimodal biometric fusion.

C. Our Approach and Contributions

While there is abundant literature on quality-based fusion, e.g., [11], [12], [13], [10], [17], [18], our approach deviates from the mainstream in two ways: First, the focus in the

mainstream is on quality-based fusion, whilst our focus is on quality-based normalisation and treating the fusion as a second stage process. Such an approach is also pursued in [10], [15]. Second, while quality measures are often treated as scalar values, in this paper, we consider them as a *vector* of measurements derived from raw biometric samples. This is very important because for instance, for face images, it is common (and necessary) to use several quality measures to characterise the quality of a cropped face image (e.g., head pose parameters, face detection reliability), and the intrinsic characteristic of image quality (e.g., focus, contrast).

Our contribution is three-fold. First, we propose a new quality-based score normalization procedure, capable of incorporating the qualitative device information. It is a generalization of several quality-based score normalisation procedures, including [10], [15], [9], [7]. Because of the generalization, the method can still be used when the device information is absent in testing (the second sub-problem), but can still take advantage of such information in training. Second, we compare several existing quality-based score normalisation procedures, essentially adapted from the quality-based fusion literature. Third, we demonstrate that quality-based normalisation/calibration with device information can be used effectively in fusion, using the Naive Bayes principal.

For the sub-problem where the device identity is known, the proposed method significantly outperforms all the submitted fusion systems in the past evaluation [19]. This method has the practical advantage of being very generic: it is applicable to any biometric device, and only the output match scores are needed. This greatly facilities integration of multimodal biometrics across different vendors (where the internal functionality of a device is often not disclosed). This is unlike the feature-level or data-level calibration where the raw data is required (although undoubtedly further performance improvement can still be tapped by this approach).

Our proposal was validated on the Biosecure DS2 database [20] using the face and fingerprint modalities.

D. Paper Organisation

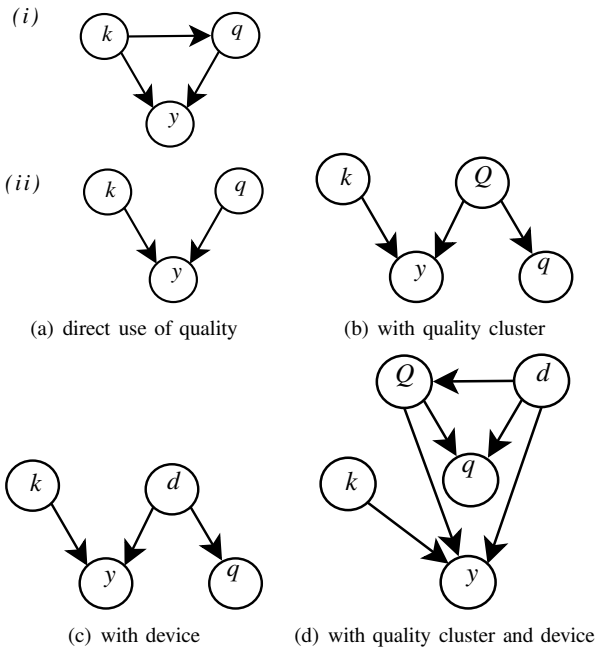
This paper is organised as follows: Section II presents a Bayesian framework unifying several quality-based normalisation procedures derived from the multimodal fusion literature. Section III describes the database that will be used to assess the procedures and the results are presented in Section IV. Section V discusses some possible future research directions. This is followed by conclusions in Section VI.

II. A BAYESIAN FRAMEWORK

This section presents four different methods of modelling the relationship between match scores and quality measures already mentioned in the introduction.

A. Notation

Throughout this paper, graphical models [21], also known as Bayesian networks [22], are used to compare different existing quality-based fusion algorithms. It is, therefore, indispensable at this stage to present this concept. A graphical model of evidence combination in statistical decision making is a graph with directed arrows representing conditional probabilities. A node in the graph is a variable. An arrow from variable A to variable B specifies their causal relationship, i.e., the conditional probability of B given A, i.e., $p(B|A)$. The graphical models that we shall use are shown in Figure 2. These models will be explained in detail in Section II-B.



(a) direct use of quality (b) with quality cluster

(c) with device (d) with quality cluster and device

Fig. 2. Various graphical models that model the relationship between match scores y and quality measures q via quality cluster Q and device qualitative information d, conditioned on the authenticity of the matching, i.e., the class label k.

In the graphs just shown, the following variables are used:

- $k \in \{\mathtt{C}, \mathtt{I}\}$ is the true class label, i.e., being either a genuine user (also known as a client) or an impostor.
- $q \in \mathbb{R}^{\mathtt{N}_q}$ is the vector of quality measures output by $\mathtt{N}_q$ quality detectors. A quality detector is an algorithm designed to assess the quality of an image, e.g., the number of bits per pixel ratio, contrast and brightness as defined by the MPEG standards. In our case, these quality measures deal with face images describing, for instance, the orientation, illumination and spatial resolution of a face image. Both the general and face-related quality measures will be used in this paper.
- the quality state $Q \in \{1, \ldots, \mathtt{N}_Q\}$ which signifies one of the $\mathtt{N}_Q$ discrete events[2] each describing a composite combination of quality degrading factors. Under well controlled facial recognition experiments, Q can be associated with factors that affect the system performance, for instance, {wearing glasses, back illumination, smile}, {no glasses, left illumination, neutral}, etc. These states are obvious from a direct examination of face images. However, from a computational point of view, Q is *not observable* when a biometric system operates without

[2]Note that $\mathtt{N}_Q$ and $\mathtt{N}_q$ are different. We use the small letter "q" to denote a quality measure and the capital "Q" to denote a *quality cluster*.

human intervention. Note that, in practice, in uncontrolled acqusition environment, it is difficult to associate Q with a semantically meaningful category, a problem akin to assigning a semantic label to a cluster found by a clustering algorithm.

- $d \in \{1, \ldots, \mathtt{N}_d\}$ signifies one of the $\mathtt{N}_d$ devices. Although the set of devices used for data acquisition is *known*, which device was actually used to acquire a particular query biometric sample may or may not be known, hence, constituting two variants of the problem. The solution to the latter problem (with unknown device) can enable "plug-and-play" of biometric devices (hence not requiring any configuration) since the device identity can be probabilistically inferred. In this study, for the face modality, two choices of cameras are possible: web-cam (of low resolution images) and digital camera (giving higher resolution images). For the fingerprint modality, the devices are either optical or thermal sensor. While it is easy to distinguish manually images taken by a web-cam versus a digital camera, or by an optical versus a thermal sensor, from a computational point of view, the device d is *not observable* when a biometric system operates without human intervention.
- $y \in \mathbb{R}^N$ is the vector of scores output by N base biometric authentication systems. They are stacked to form a vector. This form is used often in a fusion framework [23] (and references therein) and is useful to model the dependency among the system outputs. On the other hand, when $N = 1$, i.e., no other systems are involved, estimating the density of a single system output is still beneficial because quality can systematically influence the match score distribution. More rationale and insights are given in Section II-B.

B. Four Different Models and Their Rationale

The first model attempts to model the joint density of quality measures and match scores directly. The majority of quality-based fusion methods are based on this approach, e.g., [24], [25], [14], [9].

The remaining three models utilise the concept of *quality cluster* (i.e., a cluster of quality measures). It is assumed that quality measures of the same cluster will exhibit statistical properties that are not only similar among themselves (by definition) but are also coherent among the match scores they produce during matching. The concept of quality cluster was first proposed in [9]. In that study, a fusion strategy was devised for each cluster of quality measures, in order to combine the outputs of six face experts. Since only the cluster index is used when modelling the joint score density, the task of modelling the relationship between quality measures and match scores is greatly simplified.

There are at least three possible ways one can derive clusters of quality measures:

1) The first approach considers finding the natural clustering of quality measures derived from as much biometric data as possible, collected using different devices, possibly in different application scenarios. The main characteristic of this approach is that the device qualitative information is not used. For the case of a single device, this model was reported in [9] where, in order to combine several face experts (hence intramodal fusion), a fusion strategy was devised for each cluster of quality measures. In the experimental context of [9], two clusters of quality are expected since the face images are taken with either frontal or side illumination. In an actual application, it is reasonable to expect several clusters of quality which could be associated with the user manner of interaction with a biometric device, and the acquisition environment. For instance, simply providing an impression of a fingerprint or sliding it over a small sensor surface will certainly produce different effects and is likely to exhibit different image quality.
2) In the second approach, one categorises the quality measures according to the device which was used to collect the biometric samples (from which the quality measures have been derived). If there are N_d devices, there will be N_d clusters of quality measures. In this case, one says that the clusters Q are *device-dependent*. This approach was found in [7], [16].
3) The third approach can be considered a further refinement of the second approach. Since there is no guarantee that the quality is consistent for each device, it is reasonable to further find the natural clustering that exists within each device-dependent cluster of quality measures. It will be shown that this approach generalises the above two approaches.

The four graphical models for quality-based biometric fusion are shown in Figure 2. Model (a-i) was proposed in [10] and a theoretical model discussing its effectiveness can be found in [25]. Model (a-ii), with q being a discrete scalar variable as a special case, was found in [15] (the interpretation of graphs will be further discussed in Section II-C).

Different from model (a), models (b), (c) and (d) all utilise the concept of cluster of quality measures. Model (b) clusters the quality measures regardless of the device qualitative information. It has been used successfully [9] in intramodal face fusion in which case a single device was involved. Consequently, the issue of cross-device matching was not of concern[3].

Model (c) [7] explicitly considers the device information. The structures of model (b) and (c) are very similar; they converge to the same model if one associates each cluster of quality measures to an acquisition device. However, this is not always necessarily the case, because the concept of cluster extends easily beyond that of a particular device.

Finally, model (d) considers both cluster and device information. It addresses the issue that a single device can produce varying levels of quality. The variation in quality may not be

[3]The causal relationship shown in Figure 2(b) is $p(q|Q)$ whereas in [9], it is $p(Q|q)$. As will become clear later (see (8), $p(Q|q)$ is indeed needed during inference. Although this seems to be inconsistent, it is not because there exists a deterministic relationship between $p(Q|q)$ and $p(q|Q)$, as shown by (9). By interpreting Q as some latent factors affecting biometric performance, "Q causes q" seems to fit better the picture. The same remark also applies to model (c), as reported in [7].

due to the device itself; it could be the consequence of different ways the user can interact with the device, *and* varying acquisition conditions. By further identifying a number of quality clusters per device, model (d) is a generalisation of models (b) and (c).

It is worth highlighting the generative nature of the above graphical models. For instance, according to model (d) (see Figure 2(d)), the arrow $Q \rightarrow q$ implies that the discrete variables Q and d directly influence the measurement q. In order to understand the significance of this, let us use the following hypothetical example. Suppose that we have a human expert annotating different quality aspects (factors constituting Q) from a face image. These aspects are a pair of binary attributes: {presence of glasses, emotion} where emotion is either neutral or with expression. In this case, Q will have four states and they are manually annotated. If a database is controlled and {presence of glasses, emotion} are the *only* two possible sources of variation, then, by clustering q it is possible to recover the resultant four quality states. The above example shows that Q can capture a higher level of information. However, in reality, the number of states in Q is unknown. Furthermore, in practice, it is also difficult to assign the significance to each cluster; this is an on-going research issue. We shall be content in this paper that once some clusters are found, the above models can be used to design multimodal biometric fusion algorithms.

The device qualitative information d certainly influences the quality cluster Q. This knowledge is shown by adding an arrow from d to Q in model (d). Figure 1 illustrates the local fingerprint gradient produced by two fingerprint sensors. The global fingerprint quality is defined as the average of all local fingerprint gradients. It is based on [8] and is used in all our experiments. Because the devices are different, the resulting verification performance is also different for these two devices. This is also well supported by our experiments for the face modality (see Figure 3). This justifies the causal relationship $p(Q|d)$ and its use across different biometric modalities.

We conjecture that the varying performance of different devices (keeping in mind the matching algorithm, for the same biometric modality, remains identical throughout this paper) is due to the changing nature of the class-conditional match score distributions from one device to another, i.e., given devices d_1 and d_2, and all other conditions being equal, $p(y|k, d_1)$ can be significantly different from $p(y|k, d_2)$ for each of the class labels $k \in \{\mathtt{C}, \mathtt{I}\}$. This phenomenon is, again, well supported by our experiments (see Figure 4). It is beyond the scope of this paper to explore the reasons for the existence of distribution differences. We have, however, two plausible explanations. First, the input modes, i.e., either sliding or impressing a finger on a device, are completely different. Second, the feature extraction module might have been tuned to perform optimally for a particular device, and hence is sub-optimal for another one[4]. The above illustration, again, justifies why one should model $p(y|k, d)$. Similar justification extends to the cases $p(y|k, Q)$ for model (b) and $p(y|k, Q, d)$ for model (d).

To simplify the discussion, we will treat the case where both y and q are derived from a *single* biometric trait. We will then augment these variables with the subscript m in order to handle the case of multiple biometric traits. The reason for not introducing the subscript m at this point is that one does not need to model the relationship between match score $y_{m'}$ and quality measures q_m where m', m are two different biometric traits. The discussion involving different modalities will be deferred until Section II-G.

Section II-C will first explain the first approach, i.e., methods that use quality measures directly. The remaining techniques based on quality clusters are explained in Section II-D.

C. Direct use of quality measures

Model (a), as shown in Figure 2, has two variants: (a-i) and (a-ii). They can be expressed by the following joint probabilities[5]:

$$p(y, k, q) = p(y|k, q)\underbrace{p(q|k)}p(k) = p(y, q|k)p(k) \quad (1)$$

$$p(y, k, q) = p(y|k, q)p(k)p(q) \quad (2)$$

respectively. It should be noted that (2) is more *restrictive* than (1) because the former does not include the under-braced term $p(q|k)$. This term is the class-conditional density of quality measures. If such a causal relationship existed, than quality measures would have some discriminative power in distinguishing genuine users from impostors. Since this is *not* the case (as will be backed by our experiments; see Figure 5(a)), this implies that modelling $p(q|k)$ is *not* necessary. In comparison, since q has $\mathtt{N}_d$ dimensions, by not modelling $p(q|k)$, model (a-ii) will have much lower number of parameters. Therefore, in this case, although being more *restrictive*, model (a-ii) is an equally effective solution.

One way to realise a classifier from models (a-i) and (a-ii) are by using the Neyman-Pearson lemma, i.e., taking the log-ratio of two hypotheses:

$$y_{a_1}^{norm} = \log \frac{p(y, q|\mathtt{C})}{p(y, q|\mathtt{I})} \quad (3)$$

$$y_{a_2}^{norm} = \log \frac{p(y|\mathtt{C}, q)}{p(y|\mathtt{I}, q)}, \quad (4)$$

respectively. Based on (1) and (2), both classifiers are related by:

$$y_{a_1}^{norm} = y_{a_2}^{norm} + \log \frac{p(q|\mathtt{C})}{p(q|\mathtt{I})}$$

Since q is not discriminative in distinguishing genuine users from impostors, $\log \frac{p(q|\mathtt{C})}{p(q|\mathtt{I})} \approx 0$, and as a result, (3) and (4) converge.

Classifiers (3) and (4) should be compared with the one that does not use any quality measures at all:

$$y_0^{norm} = \log \frac{p(y|\mathtt{C})}{p(y|\mathtt{I})}. \quad (5)$$

[4]The matching algorithm was used out-of-the-box with no tuning whatsoever.

[5]Note that in our notation, we do not distinguish between discrete probability that is usually written with a capital "p" from the continuous one.

This classifier is the simplest among all the models presented in this paper, and is reported in [24], [25], for instance.

A crucial issue related to model (a-ii) is that estimating $p(y|k,q)$ is non trivial compared to $p(y,q|k)$ in model (a-i). This is because the conditioning variable q in model (a-ii) is multivariate and continuous. This calls for a multivariate regression solution. In [15], q was one dimensional. However, since it is continuous, the authors quantised q into several states using a histogram. During inference, based on the histogram, an observed q is assigned to a histogram bin. Since the binning process is deterministic and the transformed q, say q', is discrete (corresponding to the bin index), the density $p(y|k,q')$ can be estimated easily.

In the case of multivariate q, a direct generalization of this approach is to cluster the quality measures into, say $\mathtt{N}_Q$ clusters. In this way, an approximate solution is to model $p(y|k,Q)$ instead of $p(y|k,q)$. This represents a significant savings in terms of the number of parameters since the variable Q is one dimensional. As a result, the model $p(y|k,Q)$ remains in N dimensions. In comparison, $p(y,q|k)$, which is needed in model (a-i), has $N+\mathtt{N}_q$ dimensions. This implies that for quality-based fusion with large $\mathtt{N}_q$, e.g., $\mathtt{N}_q \gg N$, the methods relying on cluster quality are much more scalable since the match score density is independent of the dimension of q.

This is treated in the next section.

D. Quality Clusters with Models (b) and (c)

The joint densities represented by models (b) and (c) are given as follows:

$$p(y,k,q,Q) = p(y|k,Q)p(q|Q)p(Q)p(k) \quad (6)$$

$$p(y,k,q,d) = p(y|k,d)p(q|d)p(d)p(k) \quad (7)$$

In model (b), one assumes *conditional independence* between y and q given Q [21], i.e., if the state Q were known, one could estimate $p(y|Q)$ without the knowledge of q. The consequence, in practice, is that one can first cluster the quality measures, and then estimate the density of match scores y for each cluster. By so doing, according to model (b), one indirectly estimates the relationship between y and q, which is the ultimate goal.

Model (c) can be explained in a similar manner by replacing Q with d, hence attributing a cluster to a device.

In order to design a classifier from model (b) using the Neyman-Pearson lemma, one should first estimate $p(y|k,q)$ from (6):

$$p(y|k,q) = \frac{\sum_Q p(y,k,q,Q)}{p(k,q)} = \sum_Q p(y|k,Q)p(Q|q) \quad (8)$$

where $p(Q|q)$ is the posterior probability of Q given q, i.e.,

$$p(Q|q) = \frac{p(q|Q)p(Q)}{\sum_{Q'} p(q|Q')p(Q')}. \quad (9)$$

Similar to (4), the final classifier takes the following form:

$$y_b^{norm} = \log \frac{\sum_Q p(y|\mathtt{C},Q)p(Q|q)}{\sum_Q p(y|\mathtt{I},Q)p(Q|q)} \quad (10)$$

An important property of model (b) is the concept of "conditional independence" [21] between y and q. Given that the state of Q is *known*, y and q becomes independent of each other. This property can be exploited when learning the model parameters (see Section II-F).

Similar to model (b), the classifier realised using model (c) is:

$$y_{c_1}^{norm} = \log \frac{\sum_d p(y|\mathtt{C},d)p(d|q)}{\sum_d p(y|\mathtt{I},d)p(d|q)} \quad (11)$$

where $p(d|q)$ is defined similarly to (9) except that the variable Q is replaced by d.

When the device is known, we have the following variant:

$$y_{c_2}^{norm} = \log \frac{p(y|\mathtt{C},d_*)}{p(y|\mathtt{I},d_*)} \quad (12)$$

since $p(d_*|q)=1$ and $p(d|q)=0$ for all the remaining devices $d \neq d_*$.

E. Quality Clusters with Model (d)

The last model, as shown in Figure 2 has the following joint density:

$$p(y,k,q,Q,d) = p(y|k,d,Q)p(q|d,Q)p(Q|d)p(d)p(k) \quad (13)$$

Model (d) can be viewed as a refinement of model (b) by further considering the device qualitative variable d with the following three additional arrows:

- $d \rightarrow Q$, hence the causal relationship $p(Q|d)$
- $d \rightarrow q$, resulting in $p(q|Q,d)$
- $d \rightarrow y$, implying the need to estimate $p(y|k,Q,d)$

The justification for the above causal relationships has already been given in Section II-C.

The conditional density useful for implementing the Neyman-Pearson lemma, when the device information is known, is as follows

$$\begin{aligned} p(y|k,q,d) &= \sum_Q p(y|k,q,Q,d) \\ &= \sum_Q p(y|k,Q,d)P(Q|d,q)P(d|q) \\ &= \sum_Q p(y|k,Q,d)P(Q|d,q) \end{aligned} \quad (14)$$

since $P(d|q)=1$ (recall that d is known). Using (14), the final output of the classifier from model (d) is:

$$y_d^{norm} = \log \frac{p(y|\mathtt{C},q,d)}{p(y|\mathtt{I},q,d)} \quad (15)$$

On the other hand, if the device information is not given, we have the following variant of (14):

$$\begin{aligned} p(y|k,q) &= \sum_d \sum_Q p(y|k,q,Q,d) \\ &= \sum_d \sum_Q p(y|k,Q,d)P(Q|d,q)P(d|q) \\ &= \sum_d \sum_Q p(y|k,Q,d)P(Q,d|q) \\ &= \sum_{Q'} p(y|k,Q')P(Q'|q), \end{aligned} \quad (16)$$

where we introduced a composite variable $Q' \equiv \{Q, d\}$. This converges to model (b), and the final output is computed using (10). In this case, the two classifiers are *identical* during inference.

A subtle difference between model (b) and model (d) is that the training strategies are different. When training model (d), the quality measures q are partitioned into different devices. For each d, one then clusters the quality measures. On the other hand, when training model (b), one pools all quality measures of all devices into a big data set first and then identifies the quality clusters from the data set. Were there many more devices available, model (b) would be a viable solution to address the problem of identifying a biometric device from an *unknown* set of devices. Testing this conjecture is not possible given that only two devices are available in our database setting (to be presented in Section III).

F. Models Fitting

We shall discuss the issue of learning the model parameters for model (b) only. The discussion extends to model (d) easily since the difference between both models is the existence of the discrete conditioning variable d, e.g., from $p(y|k, Q)$ to $p(y|k, Q, d)$.

An important property of model (b) is the concept of "conditional independence" [21] between y and q. That is, given that the state of Q is *known*, y and q becomes independent of each other. We will exploit this property in learning the model parameters. Before doing so, it is instructive to illustrate this property using the hypothetical example already mentioned in Section II-B.

Suppose that a human expert has annotated all face images in a database with the following binary attributes: {presence of glasses, emotion}. By considering all possible combinations of these two binary attributes, Q will have four states. Parallel to this, for each face image, an array of automatically derived quality measures is also made available. Since in this hypothetical scenario, Q is observed in training, one can first learn $p(y|k, Q)$ and then learn the relationship between q and Q, i.e., $p(Q|q)$ based on (8), in two *distinctive stages*. Note that the order in which the stages are realised is of no importance. Therefore, thanks to the *conditional independence* property, when Q is observed, the practical implication is that one can train $p(y|k, Q)$ and $p(Q|q)$ independently.

In reality, however, it is often not possible to annotate all the quality of the training samples, and furthermore it is not a viable solution when the system is operational. One has to resort to discovering the natural clusters by using a clustering algorithm. Given that Q is unknown, a well known learning framework with missing observation, called Expectation Maximisation (EM) [26], can be used. The learning goal is to maximise the expectation of (6) (in logarithmic scale), while marginalising over the unknown variable Q, assuming that samples are independently and identically distributed:

$$\text{criterion} = E_{p(Q|q)}\left[\sum_n \log p(y^{(n)}, k^{(n)}, q^{(n)}|Q)\right]$$

where the superscript (n) is the index of a sample in the training set, and $E_p[\cdot]$ denotes expectation over the distribution p. This consists of the following two steps: In the expectation step, one estimates the posterior of Q for each sample via (8) (with already initialised or calculated parameters). In the second step, one maximises the likelihood $p(y, k, q|Q) = p(y|k, Q)p(q|Q)p(Q)$, with respect to the distribution parameters, for each Q. A complete roll-out of an EM algorithm needs to specify the densities: $p(y|k, Q)$, $p(q|Q)$ and $p(Q)$. Typically, $p(y|k, Q)$ and $p(q|Q)$ are multivariate Gaussian and $p(Q)$ is a (discrete) probability table.

We shall propose a modular learning solution here that exploits the conditional independence, and it makes no assumption about the form of $p(y|k, Q)$. The basic idea consists of identifying the natural clusters Q given a set of q using a clustering algorithm, e.g., a Gaussian Mixture Model (GMM) [26]. GMM estimates the density $p(q) = \sum_Q p(q|Q)p(Q)$, and in this process, outputs $p(q|Q)$ as the (Gaussian) component that we need. Once the clusters are known, in the second step, for each $Q = Q_*$, $p(y|k, Q_*)$ can then be learnt using the training samples (y, q) belonging to component $Q_* = \arg\max_Q P(Q|q)$. In our implementation, $p(y|k, Q)$ is modelled using a GMM for each class label k separately and for all Q.

In summary, the class conditional match score distributions $p(y|k)$ (for the baseline comparison), $p(y|k, d)$, $p(y|k, Q)$, $p(y|k, Q, d)$ and $p(q|d)$ are estimated using GMM. When using a GMM, the number of Gaussian components is tuned by cross-validation on the development set.

G. Architectural considerations

This section deals with quality-based fusion involving several biometric traits. In the previous sections, y was treated as a score vector whose elements are individual subsystem outputs observing the *same* biometric trait. In multimodal fusion, one often incorporates the quality measures one modality at a time. This is done for two reasons: first, the quality measures of one biometric modality do not give any information about the other biometric modality – hence implying that there is no need to model the joint density of y belonging to one biometric modality and q of another biometric modality. Second, each of the output y^{norm} of different modalities, e.g., for any of (3), (4), (10), (11), (12) and (15), can be combined using the product rule thanks to the independence of the biometric modality. Such a Naive Bayes solution is appropriate when the biometric modalities are independent. It also has the advantage that fewer parameters need to be estimated than for the joint modelling of all variables[6].

For the above reasons, we shall introduce the modality dependent notation here. Let y_i be the output of the i-th biometric modality; q_i be its vector of quality measures extracted from a biometric sample. Let there be $i = \{1, \ldots, \mathtt{N}_m\}$ biometric modalities. Note that the number of elements in y_i corresponds to the number of classifiers, all observing the same biometric modality, to combine. This is done in order to capture the

[6]Note that modelling the joint distribution of variables y^{norm} across modalities is still possible provided that the data used to train the normalising parameters is different from the one that will be used to train the second stage fusion (taking y^{norm} across modalities as input). This can be done using cross validation, for instance.

inherent dependency among the system outputs. This problem is known as *intramodal* fusion.

In this context, the final output, realised using the Neyman Pearson lemma, for model (a-i) is:

$$y_i^{norm} = \log \frac{p(y_i|\mathtt{C}, q_i)}{p(y_i|\mathtt{I}, q_i)} \quad (17)$$

Thanks to the independence assumption, the final combined score can be written as:

$$\begin{aligned} y_{gen}^{com} &= \log \frac{\prod_i p(y_i|\mathtt{C}, q_i)}{\prod_i p(y_i|\mathtt{I}, q_i)} \\ &= \sum_i \log \frac{p(y_i|\mathtt{C}, q_i)}{p(y_i|\mathtt{I}, q_i)} = \sum_i y_i^{norm}. \end{aligned} \quad (18)$$

Similar formulation can also be written for the remaining models, i.e., by augmenting subscript i in (4), (10), (11), (12) and (15).

This means that one can implement any one of the four graphical models for each biometric modality in parallel and then combine all $\mathtt{N}_m$ biometric modalities using the sum rule when each pre-processed output is a log-likelihood ratio test.

H. Summary

In summary, we have presented four different approaches for quality-based score normalization. In the presence matching involving several devices, models (c) and (d) should be used, because both of them take into consideration the device information explicitly. Between the two, model (d) is more flexible in that it can handle possible variation in terms of quality of the data, thanks to the hidden variable Q. If there is only a single state, model (d) converges to model (c). Hence, in this sense, model (d) is a generalization of model (c). If only a single device is involved, model (d) converges to model (b). In this sense, model (d) is a generalization of model (b). Finally, by integrating the hidden variable Q, a necessary operation during inference, model (b) converges to model (a). In this sense, model (b) provides a means to estimate (the density of) model (a). In each of the above cases, we observe that model (d) is a complete generalization of all other models.

In the next section, we shall introduce the database that will be used to compare the performance of these models.

III. DATABASE

A. Face-Fingerprint Experimental Protocol

The Biosecure database used in this study contains as many as six biometric modalities, i.e., face, speech, signature, fingerprint, hand and iris. It was collected from 6 participating European sites each contributing biometric samples of between 20 and 110 persons. The database was captured using different devices under varying conditions. In this paper, we use a database subset designed for the purpose of access control, called "the DS2 (Desktop) evaluation".

The DS2 subset database was collected over two sessions[7], separated by about one month interval. In each session, two biometric samples were acquired for each device. This results in four samples per device collected over two sessions. The first sample of the first session is used to build a biometric template. The second sample of the first session is used to generate a genuine user match score of session 1 whereas the two samples of the second session are used in a similar way to generate two genuine user match scores.

[7]Downloadable at http://face.ee.surrey.ac.uk/qfusion

TABLE I
THE EXPERIMENTAL PROTOCOL OF THE BIOSECURE DS2 DATABASE

Data sets		No. of match scores per person	
		dev. set (51 persons)	eva. set (156 persons)
Session 1	Genuine	1	1
	Impostor	103×4	126×4
Session 2	Genuine	2	2
	Impostor	103×4	126×4

$\cdot \times \cdot$ are persons $\times$ samples. This number should be multiplied by the number of persons in the above set to obtain the total number of accesses for the genuine or the impostor classes.

It is important to distinguish the two data sets, i.e., the development and the evaluation sets. The development set is used for algorithm development, e.g., finding the optimal parameters of an algorithm, including setting the global decision threshold. For unbiased performance assessment, the population of users in these two data sets are disjoint.

There are in total 333 subjects in the database, among which 206 are considered "clients", and a template is created for all of them. The *development impostor score set* contains 103×4 samples, i.e., 103 subjects, with 4 samples per subject. The 206 client-set is divided in two equal subsets of 103x4 samples, the genuine and the impostor score set. When a reference subject is considered a genuine user it is associated with the genuine subset, all the subjects of which are used as impostors in Session 2. This ensures that the impostors used in Sessions 1 and 2 are not the same. Such a characteristic is important for algorithm development. All the 4 samples of the remaining half of the 206 subjects are considered impostors in the development set in Session 1.

The remaining 126 subjects constitute an *evaluation impostor score set* that is considered as an external population of users who serve as zero-effort impostors. In this way, a fusion algorithm will not make use of impostors *seen* during its training stage; hence, avoiding systematic and optimistic bias of performance.

B. Dealing with Failures

In this study, the data in Session 1 is *not* used; instead, only the data in Session 2 is used. Session 1 data is intended for evaluating *client-specific* algorithms, i.e., algorithms the parameters of which differ from one person to another according to the claimed identity. The exact number of accesses differs from that listed in Table I because of missing data due to the failure of the segmentation process or other stages of biometric authentication. The experimental protocol involves minimal manual intervention. In the event of *any* failure, a default score of "-999" is outputted. Similarly, failure to extract quality measures will result in a vector containing a series of "-999". For the purpose of this study, samples with missing

values are removed. This is not critical in the context of this study because the comparison gauges the merit of using or not using quality-dependent score normalisation. It should be noted that inference even with missing values is still possible. In fact any missing observation will simply not contribute to the final classifier output, i.e., not part of the indice i in the sum of equation (18).

For the purpose of this paper, only face and fingerprint modalities are considered. They are described in the following sections.

C. Face Systems

Low and high quality still frontal face images are collected by using two different sensors: a Phillips SPC 900 web camera (low quality) and a CANON EOS 30D digital camera (high quality:with/without flash). The images and quality measures are provided by the Omniperception SDK[8]. The low quality images are denoted as "fa" whereas the high quality ones (without flash) are denoted as "fnf". The latter ones with flash were not used here.

The reference system is an LDA-based face verifier [27]. In the context of this study we utilised 14 quality detectors, i.e., face detection reliability, brightness, contrast, focus, bits per pixel, spatial resolution between eyes, illumination, degree of uniform background, degree of background brightness, reflection, presence of glasses, rotation in plane, rotation in depth and degree of frontal face images. Although there are different scenarios in the way we can perform a quality-based evaluation (high/low quality templates vs. high/low quality queries),[9] the face mismatch problem that we considered in this paper is *high quality templates* vs. *high/low quality queries*. Although only one of these scenarios is tested here, this does not constitute a weakness as long as the models are concerned, because all the models presented in this paper can be extended naturally to different scenarios by estimating the appropriate densities.

D. Fingerprint Systems

The fingerprint data was collected by using both an optical (denoted as "fo") and a thermal sensor (denoted as "ft"). The optical sensor captures the entire fingerprint by direct contact whereas the thermal sensor requires the user to slide his/her fingers. Although the right/left fingerprints of the thumb, index and middle finger of the subjects have been acquired, in the context of this study we have considered only the right-side fingerprints. The reference system used is the *NIST Fingerprint system*[10]. The fingerprint quality measure is based on a weighted average of local gradients as proposed in [8].

IV. EXPERIMENTS

The experiments are divided into six parts. The first three parts perform an analysis of the data, independent of the classifiers used; whereas the remaining three parts are related to the use of the four statistical models presented in this paper. These experiments are:

1) *Performance of same-device versus cross-device matching*: In the case of the same-device matching, the template and query data are acquired using a common device, whereas in cross-device matching, both acquisition devices are different. For the latter case, we consider only the situation where the template data is acquired using a device giving higher verification rate whereas the query data is acquired using a device giving lower verification rate.
2) *Analysis of the change of the match score distribution in cross-device matching*: This experiment is designed to validate our conjecture that the class conditional distribution of match scores (belonging to either genuine or impostor classes) may behave differently for different devices in *cross-device* matching. This provides an explanation for the observed phenomenon in the Part 1 experiment.
3) *The discriminative power of quality measures in identifying devices*: The goal of this experiment is to test the effectiveness of various quality measures in distinguishing the devices and conditions under which a biometric sample is collected. This essentially tests the feasibility of estimating the posterior probability of devices given quality measures, $p(d|q)$. In our context, this classifier was built from the density $p(q|d)$ and probability table $p(d)$ using the Bayes rules.
4) *A Comparison of various quality-based normalisation schemes*: We have presented four different models in this paper. The effectiveness of these models is assessed using the Biosecure DS2 subset database presented in Section III. Each of the four classifiers relies on input from a single biometric trait. The four biometric traits tested are face, right thumb, right index and right middle fingers.
5) *Multimodal fusion exploiting quality-based normalisation*: The goal of this experiment is to test the effectiveness of *quality*-normalised scores in fusion. This experiment takes the output of each classifier across all the four available biometric traits in the Part 4 experiment and combines them using the product rule (or the sum rule in the logarithmic domain), i.e., (18). This realises the Naive Bayes solution.
6) *Analysis of the relationship between quality measures and match score distributions*: Since all the statistical models presented in this paper are designed to model the relationship between quality measures and match scores, it is instructive to visualise them.

A. Observations

1) Experiment 1: The performance of the same-device and cross-device matching is shown in Figure 3 (in terms of EER). Comparing the performance of single modalities in these two scenarios, we observe that the results of cross-device matching are much worse.

[8]http://www.omniperception.com/products/affinity

[9]A template is the data sample used to represent the claimed identity whereas a query is the sample with which the template is compared.

[10]http://www.itl.nist.gov/iad/894.03/fing/fing.html

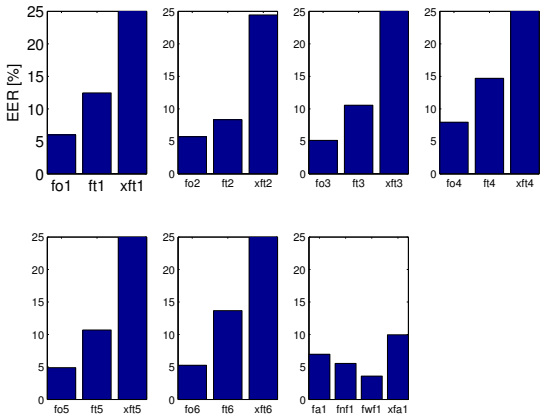
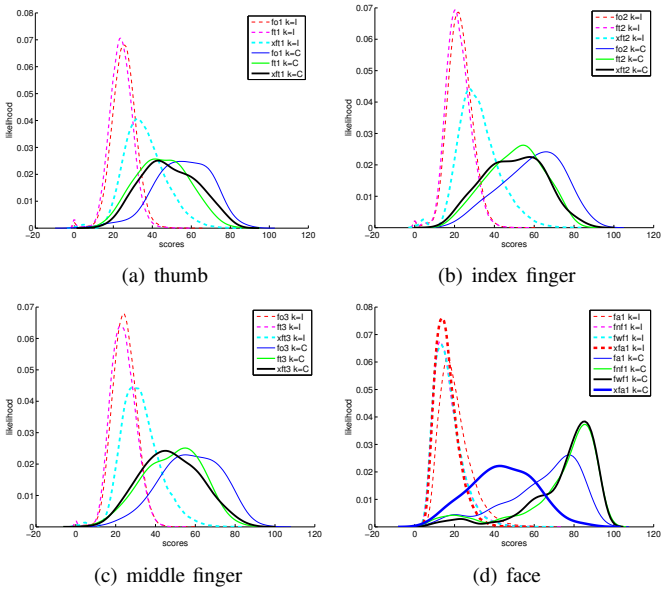


Fig. 3. The performance measured on the session-two development set (defined according to Table I) in terms of EER (%) of all the six fingerprint biometric traits (thumb, index and middle fingers of two hands) and face data (bottom right figure) obtained from the Biosecure DS2 data set. All labels in the x-axis prefixed with "x" denote cross-device performance. Fingerprints collected with the optical sensor are denoted by fo{n} and those collected with the thermal sensor are ft{n}, where $n \in \{1, 2, 3, 4, 5, 6\}$. xft{n} denotes cross-device matching between fo{n} as template and with ft{n} as query. Face images captured with the web cam are denoted by fa1; with high resolution camera by fnf1. The cross-device matching between fnf1 (as template) and fa1 (as query) is denoted by xfa1.

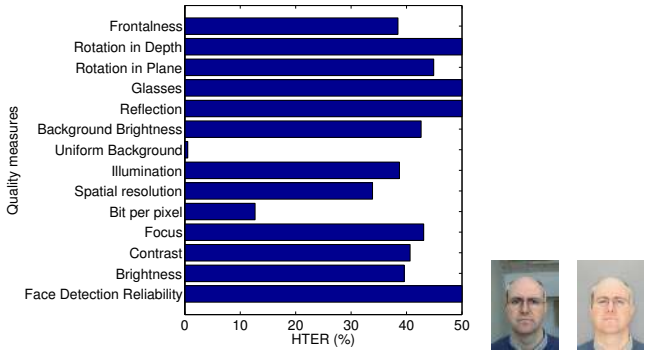


(a) thumb (b) index finger

(c) middle finger (d) face

Fig. 4. The genuine user match score distribution (blue continuous line) and impostor match score distribution (red dashed line) of four biometric traits under same- and cross-device matching. The *thickest lines* denote the distribution of cross-device matching. fnf1 and fwf1 are face images taken by a digital camera without and with flash, respectively.

2) Experiment 2: The class-conditional match score distributions for each biometric trait and each channel of data are shown in Figure 4. Both this figure and Figure 3 are plotted for the same set of match scores. We note that the match scores arising from cross-device matching overlap more heavily, hence explaining the degraded performance. The actual EERs are quoted in the legend. It is interesting to observe that the face and fingerprint systems respond differently to cross-device matching. For the face biometric, impostor match scores shift towards genuine match scores. This implies that with degraded image quality, the image of any impostor looks more similar to a genuine user, hence, producing *lower* dissimilarity match scores. On the other hand, for the fingerprint modality, under cross-device matching, genuine match scores shift towards impostor match scores, i.e., any genuine user fingerprint template becomes indistinguishable from an impostor trial. This produces *higher* dissimilarity genuine match scores.
3) Experiment 3: Each of the 14 quality measures have a varying degree of discriminative power in disambiguating the composite effect of change in acquisition device and environmental conditions. The uniform background turns out to be able to distinguish the two face images very well. Sample images on the right (Figures 5(b) and (c)) reveal that the background is essentially the most discriminative feature in this case. Because the two cameras have different image resolutions, the bit-per-pixel measure is also able to distinguish them well. Following this observation, only these two features are used in the subsequent experiments.
4) Experiment 4: The DET curves of the four classifiers corresponding to models (a-i), (b), (c) and (d), as well as that of the original match scores, are shown in Figure 6[11]. Each sub-figure shows the performance of the four classifiers, applied to a biometric modality. Note that no fusion is involved. This experiment thus shows the effect of quality-based score normalization using four different approaches. Table II lists the Half Total Error Rate (HTER) and NIST operating cost of the normalised as well as the fusion systems.
HTER is defined as the average of false acceptance rate and false rejection rate. The NIST operating point is



(a) Discriminative power of quality measures in identifying devices (b) (c)

Fig. 5. (a) The discriminative power of 14 face quality measures for high/low quality face images, measured on the development set. These detectors are: face detection reliability, brightness, contrast, focus, bits per pixel, spatial resolution between eyes, illumination, degree of uniform background, degree of background brightness, reflection, presence of glasses, rotation in plane, rotation in depth and degree of frontal face images. (b) An example of face image taken with the web camera. Note the cluttered background. (c) An example of face image of the same person taken with a digital camera

[11]Note that model (a-ii) was not tested because of the scalability issue relating to the fact that q cannot be extended to multiple dimensions. In this respect, model (b) can be seen as a way to implement model (a-ii).

defined as:

$$C_{DET}(C_{FR}, C_{FA}) = \underbrace{C_{FR} \times P(C)}_{} \times \text{FRR}(\Delta) + \underbrace{C_{FA} \times P(I)}_{} \times \text{FAR}(\Delta),$$

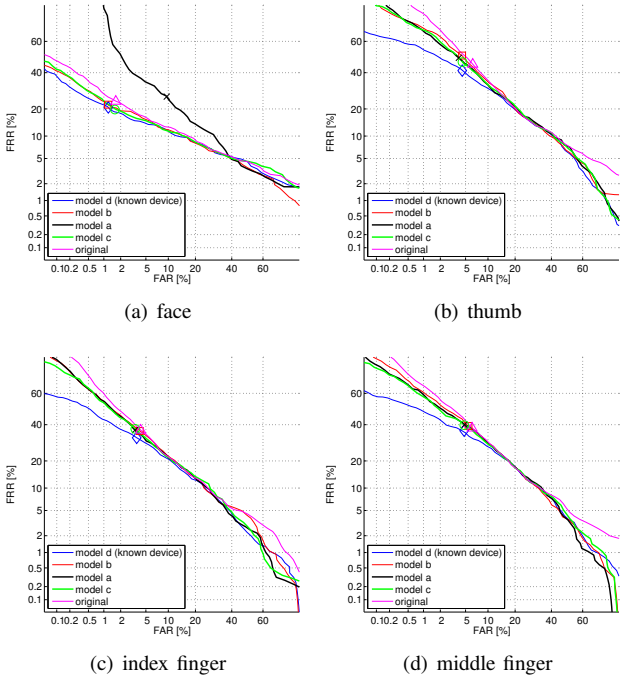
where C_{FA} and C_{FR} are respectively the costs of FA and FR, and $P(k)$ is the prior probability of class $k \in \{C, I\}$. Following the NIST evaluation, we use the following constants:

$$C_{FR} = 10 \text{ , } C_{FA} = 1 \text{ , } P(C) = 0.01 \text{ and } P(I) = 0.99.$$

The following observations can be made:

- For the face modality, the performance of model (a-i) is worse than that of the original system (without normalising the match scores). We have run the algorithm (based on GMM) with different numbers of Gaussian components and each time with several random seeds (since the algorithm is non-deterministic), but to no avail. For the fingerprint modality, the performance of model (a-i) is better than that of the baseline system. An important difference between the face and the fingerprint modalities is that $\mathtt{N}_q = 1$ for fingerprint and $\mathtt{N}_q = 2$ for face. The effectiveness of model (a-i) with $\mathtt{N}_q = 1$ is consistent with the result reported in [24] in the context of fusion. However, our results here further show that model (a-i) cannot be easily extended to multiple dimensions of quality measures.
- Although model (c) was not given the device qualitative information, it was able to improve over the baseline system, albeit only marginally. With the knowledge of device identity, (in which case $(p(d_*|q) = 1)$ for the actual device used), model (c) is slightly better in performance.
- Models (b) and (d) are effectively the same system, except that model (b) does not have access to the device information. The consequence is that model (b) is systematically worse in performance than model (d). Nevertheless, the absolute performance of model (b) is very close to that of the baseline system.
- With the device knowledge, model (d) performs the best.

5) Experiment 5: The fusion performance of the four systems (using the Naive Bayes principle), as well as that of the baseline system, are shown in Figure 7. In essence, the quality-based fusion classifiers take the sum of the already normalized scores of the four biometric modalities reported in Experiment 4. However, in order to fuse the baseline systems (without applying any quality-based score normalization method), logistic regression was used. It approximates the posterior probability of being a client given the observed (input) match scores (a vector of four elements due to the four biometric traits), without using any quality information. The HTER and the NIST operating points of these curves are shown in Table II. Referring to Figure 7, as well as Table II, the following observations can be made:



(a) face (b) thumb

(c) index finger (d) middle finger

Fig. 6. DET curves of the four classifiers as well as that of the baseline system assessed on (a) face, (b) thumb, (c) index and (d) middle fingers of the Biosecure DS2 session two data set. The dot on each DET curve is the NIST operating point. Note that no fusion is involved here.

- Model (a-i) is the worst. The principal cause of its poor performance is the suboptimal behaviour of its underlying face system component (See the black curve in Figure 6(a)).
- Model (c) outperformed the baseline system by a small but nevertheless consistent margin, with a reduction of 7% of HTER but 23% of the NIST operating cost.
- With the device knowledge, model (d) outperformed the baseline system by a large margin: a relative reduction of 56.8% of HTER and 60.8% of the NIST operating cost. It also attains the best generalisation performance among all the classifiers tested.

6) Experiment 6: In an attempt to explain what has been gauged by the models, the following densities are of interest:

- $p(q) = \sum_d \sum_{Q|d} p(q|Q,d)p(Q|d)p(d)$
- $p(y|k,Q,d)$, taken from model (d). It includes model (c) as a special case when the device is known, since:

$$p(y|k,d) = \sum_{Q|d} p(y|k,Q,d)p(Q|d)$$

  or model (b) when the device is not known, as shown in (16).
- $p(y,q|k)$ taken from model (a-i)

The three densities are shown in Figure 8, one for each row, respectively. We could only plot the density of fingerprint quality measure as it is one dimensional, i.e., $\mathtt{N}_q = 1$; the face has $\mathtt{N}_q = 2$ (recalling that two out of the 14 measures were used) and so could not be shown. For

TABLE II

NIST OPERATING POINT AND HTER

(a) NIST Operating points

subsystem	NIST operating point (%)				
	model d (known d)	model b (or d) (unknown d)	model a (no Q, d)	model c (no Q)	original
face	$*$ 3.159	3.274	11.208	3.402	3.842
thumb	$*$ 8.149	9.100	8.530	8.725	10.214
index	$*$ 6.482	7.261	6.735	6.859	7.428
middle	$*$ 7.919	8.875	8.340	8.271	9.351
fusion	$*$ 1.715	2.121	5.459	1.985	† 2.758

(b) HTER

subsystem	HTER (%)				
	model d (known d)	model b (or d) (unknown d)	model a (no Q, d)	model c (no Q)	origina
face	$*$ 10.211	10.295	16.297	10.301	10.77
thumb	$*$ 19.795	20.167	20.543	20.747	20.19(
index	$*$ 15.150	15.817	15.594	16.029	15.83
middle	18.709	18.674	18.569	$*$ 18.511	18.60
fusion	$*$ 4.479	6.068	9.884	5.333	† 6.52

†: Fusion with logistic regression; $*$: smallest value in a row

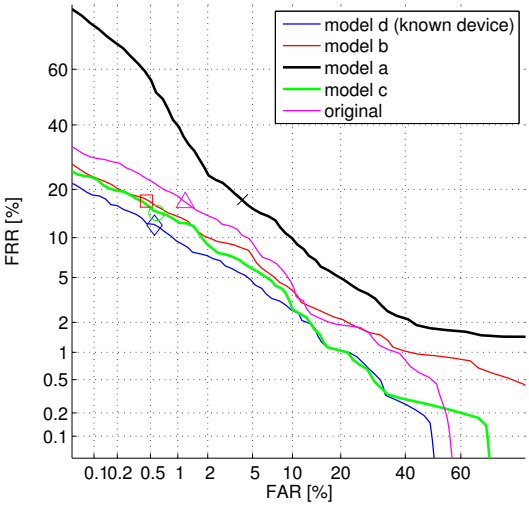


Fig. 7. DET curves of the fusion of four classifiers evaluated on the Biosecure DS2 session two data set. The dot on each DET curve is the NIST operating point.

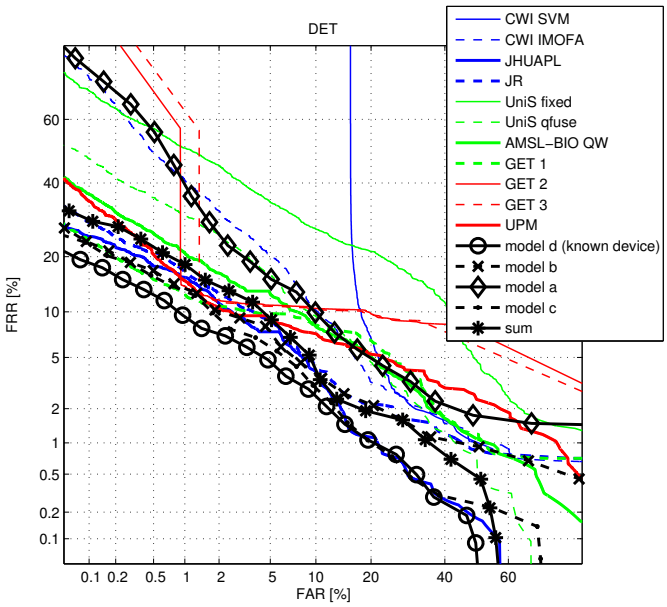


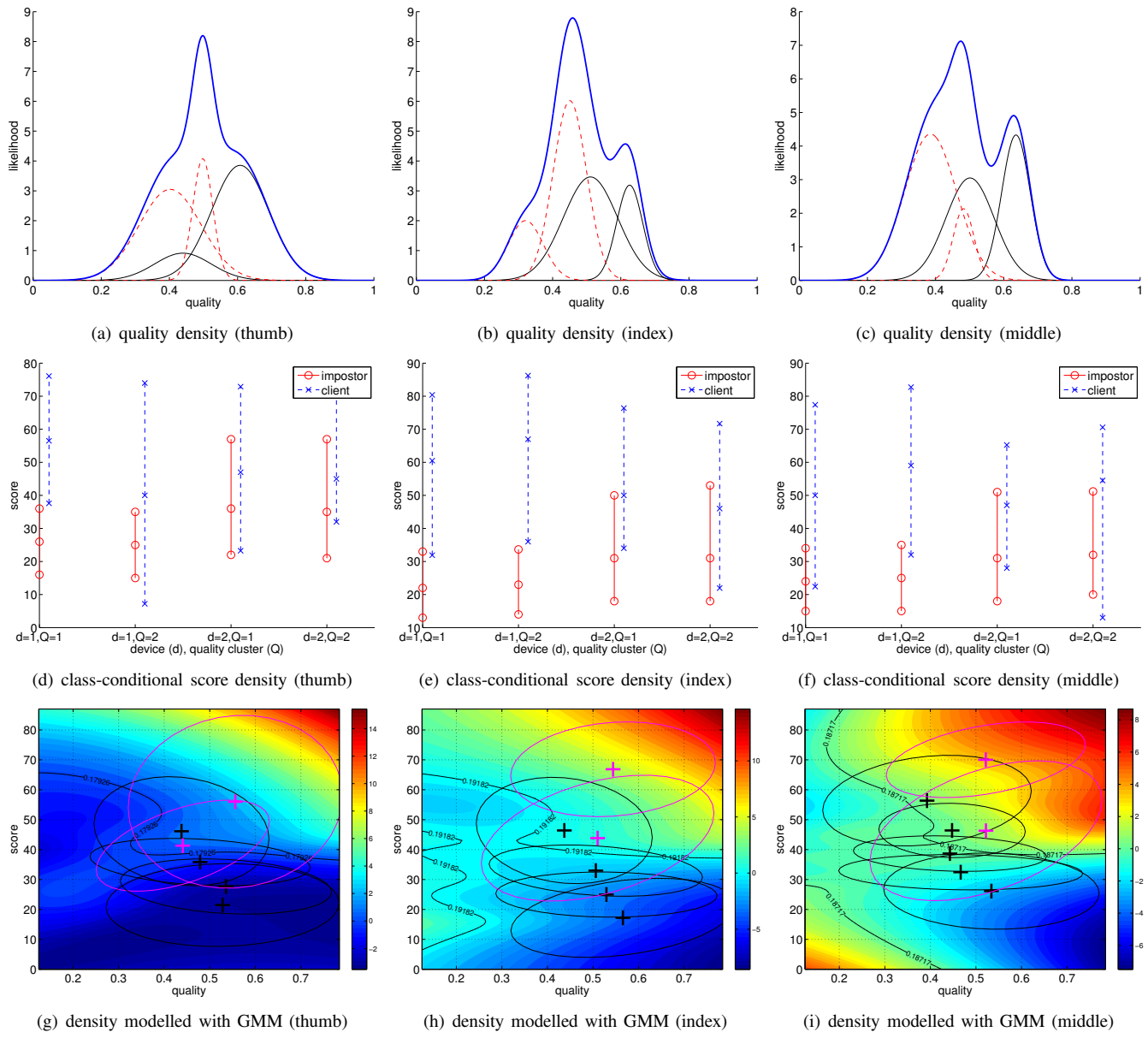
Fig. 9. Comparison of the performance of the four classifiers with the past Biosecure multimodal biometric fusion evaluation.

each finger, two Gaussian components are found to be attributed to each device. Since there are two devices (thermal or optical), this results in four components of $p(q|Q,d)$, enumerated as follow: $\{p(q|Q=1,d=1), p(q|Q=2,d=1), p(q|Q=1,d=2), p(q|Q=2,d=2)\}$. The pairs of class-conditional densities for each of these four components, i.e., $p(y|k,Q,d)$, are depicted in the second row of Figure 8. Let us examine figure (d) more closely. Note that for $d=1$, we see that the pair $\{p(y|k,Q=1,d=1)|\forall_k\}$ has less overlap whereas $\{p(y|k,Q=2,d=1)|\forall_k\}$ has a considerable overlap. This means that for the same device, there exists two distinctive latent components, discovered by clustering q. For the same problem, model (a-i) seems to be able to capture this information (see Figure 8(g). The background colour, which corresponds to the log of likelihood ratios, (3), shows that the quality-based problem is non-linear in (y,q). This observation is consistent with the finding in [14] (although different classifiers were used).

In order to compare how well the four models perform with the the rest of the fusion systems in the past Biosecure benchmarking effort [19], we replotted the DET curves of the four models (exactly as shown in Figure 7) with those of the participanting systems in Figure 9. In this figure, the following systems are used:

- CWI-SVM is a score-level SVM classifier;
- CWI-IMOFA is a Bayesian classifier with Infinite Mixture of Factor Analysis as a density estimator;
- JHUAPL is the quality-based fusion system reported in [15];
- JR is a generative fusion classifier based on the Dempster-Shafer theory of evidence, UniS fixed is a fixed-rule quality-based fusion [28];
- AMSL-BIO QW is quality-controlled weighted sum fusion [29];
- GET systems (1, 2 and 3) are Bayes classifiers with GMM as a density estimator, which can be viewed as a different implementation of model (a); and,
- UPM is device-dependent logistic regression classifier, a system very similar to the model (c) but implemented using a discriminative classifier.

Note that model (c) was also one of the submitted candidate systems in the compeition (previously labelled as "UniS qfuse"). In the competition, model (c) was the top performing system in the low FAR region. JHUAPL was the top performing system in the low FRR region. For all the systems in the competition, they were not given the knowledge of the device information. When this information is given, as can be observed, model (d) was able to take advantage of this and



(a) quality density (thumb) (b) quality density (index) (c) quality density (middle)

(d) class-conditional score density (thumb) (e) class-conditional score density (index) (f) class-conditional score density (middle)

(g) density modelled with GMM (thumb) (h) density modelled with GMM (index) (i) density modelled with GMM (middle)

Fig. 8. Each column in this figure is generated from the same data set. Column one, two and threes are generated from thumb, index and middle fingers, respectively. Each row presents different estimated densities from the data. Row one shows the density of fingerprint quality measure $p(q)$ (blue continuous line) fitted using the following mixture of Gaussian components: two Gaussian components (continuous black) for device one (giving high quality values), i.e., $p(q|Q, d = 1)$; and two Gaussian components (red dashed lines) for device two (giving low quality values). Row two shows the pair of distributions of class-conditional match scores $p(y|k, Q, d)$ for $k = \{\mathtt{C}, \mathtt{I}\}$. There are four pairs of $p(y|k, Q, d)$ since $Q = \{1, 2\}$ and $d = \{1, 2\}$, each corresponding to the four $p(q|Q, d)$ densities. Blue dashed vertical lines denote $p(y|\mathtt{C}, Q, d)$ whereas red vertical lines denote $p(y|\mathtt{I}, Q, d)$. These densities are obtained from model (d). The last row plots the density $p(q, y|k)$ as a mixture of Gaussian distributions as captured by model (a-i). A black ellipse denotes an impostor Gaussian component whereas a red ellipse denotes a genuine user Gaussian component (centred on their respective "+" sign as their mean). The colour in the background corresponds to $\log\{p(q, y|\mathtt{C})/p(q, y|\mathtt{I})\}$.

attained the best performance, with 4.48% EER.

V. Discussion and Future Directions

We have examined several quality-based fusion mechanisms from a Bayesian perspective, and demonstrated the benefit of adding the device information. This study is, however, a precursor directed to answering some of the issues below:

1) *Development of Discriminative strategies*: Although generative strategies have been thoroughly discussed in this paper, each of the four Bayesian models can be extended to using a discriminative classifier such as logistic regression. In essence, for model (b), this corresponds to a mixture of discriminative classifiers. For model (c), it is a device-dependent discriminative classifier, a special case of which was reported in [16]. Finally, for model (d), it is again a straightforward generalization of models (b) and (c) implemented using discriminative classifiers. The advantage of using the discriminative approach instead of the generative one is that much fewer number of parameters is needed than the generative approach (which requires explicit modeling

the class-conditional densities).

2) *The role of latent variable Q*: Our approach has shown that the latent variable Q plays a vital role. In particular, even with a single device, there are differences in the effect on performance of various states of Q. This suggests that with everything else being equal, there are factors affecting the system performance; and these factors can be parametrised by the latent variable Q. Investigating what exactly the role of latent variables is requires the knowledge of the baseline comparison subsystem. Unfortunately, this is beyond the scope of study. This issue demands further investigation.
3) *Identification of latent variable Q in unknown devices*: Our problem formulation was limited to identifying a device from a set of *known* devices. However, when the set of devices is *unknown*, the problem becomes more challenging. If a solution can be found, this will address the fundamental challenge of device interoperability. One potential solution is to further explore the role of latent variable Q. Given a sufficiently large number of devices, and a sufficiently large database, it may be possible to capture all sources of variation and factors affecting the performance of a biometric system. Finding a way to map Q to these factors offers another possible research avenue.
4) *Relevance of quality measures for a given matching algorithm*: Experiment 3 suggests that not all quality measures are useful to distinguish two devices. However, it is also required that these measures are able to distinguish among different states of the latent variable Q. In the first case, the device is known and hence this performance influencing factor is known. In the second case, the latent variable is by definition not observed. As a result, determining the usefulness or relevance of a quality measures for a matching algorithm, in the second case, demands further investigation.
5) *Semi-supervised learning from cross-device matching*: Our experimental results suggest that the same-device matching consistently outperforms cross-device matching. Hence, one can automatically update the template with a query sample from cross-device matching. Then, in theory, cross-device matching provides a mechanism for template adaptation which should ensure a long term stability of component classifiers.
6) *A theoretical model of quality-based fusion*: The Bayesian model we have presented can be potentially adapted as a theoretical model that helps to better understand the problem of quality-based fusion. The ultimate goal is to predict the performance given match scores and quality measures. A potential research direction is to extend the predictive model proposed in [30].

VI. Conclusions

If a biometric comparison subsystem has been designed to operate across different devices, its matching algorithm will be expected to compare the biometric template with a query image coming from the same or different biometric devices. The latter is called cross-device matching. This is an important issue in biometric device interoperability. This paper addresses such issue in the context of multimodal fusion.

We introduced two sub-problems of cross-device matching, depending on whether the device identity is given or not. In the latter case, we demonstrated the feasibility of probabilistically inferring the device identity using quality measures. This has the practical advantage of automatically configuring the system ("plug-and-play").

While there exist several quality-based fusion algorithms, e.g., [10], [11], [12], [13], [14], many are not suitable to handle the device information, which is not discriminative in distinguishing genuine users from impostors. We have proposed a family of generative models, capable of exploiting quality measures as well as the device identity, that can be used to address the above two sub-problems. The framework proposes a two-step strategy. In the first step, the output of each comparison subsystem is individually normalised (based on the respective modality-dependent quality measures). In the second step, the normalised outputs are combined using the Naive Bayes principal.

This study also compares (and generalises) several relevant state-of-the-art quality-based algorithms theoretically (from a Bayesian perspective) and empirically. Our novelty here lies in introducing the device qualitative information in this framework. The experimental results obtained on the publicly available Biosecure DS2 (score and quality measure) database have shown that the normalised match scores, in general (with the exception of model (a) due to non scalability to multi-dimensional quality measures), can give better performance than the original systems, especially at the extreme values of the DET curve (high FAR or high FRR), albeit insignificantly. However, when combining all the normalised outputs, the overall fusion performance is *statistically significantly* better than the baseline fusion (without quality normalisation), especially for the sub-problem where the device identity is given. In particular, the newly proposed quality-based normalisation with device qualitative information (model (d)) outperforms all existing quality-based algorithms.

VII. Acknowledgements

This work was supported partially by the advanced researcher fellowship PA0022_121477 of the Swiss National Science Foundation, and the two following EU-funded projects: Mobile Biometry (MOBIO, www.mobioproject.org) grant IST-214324 and the Biosecure Network of Excellence (www.biosecure.info).

References

[1] A. K. Jain, P. Flynn, and A. Ross, *Handbook of Biometrics*, Springer Verlag, 2008.
[2] A. Ross and A. K. Jain, "Biometric sensor interoperability: A case study in fingerprints," in *Proc. Biometric Authenticaion Workshop, LNCS 3087*, 2004, pp. 134–145.
[3] N. Malayath, *Data Driven Methods for Extracting Features from Speech*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, 2000.
[4] F. Alonso-Fernandez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Sensor interoperability and fusion in signature verification: A case study using tablet pc," in *Advances in Biometric Person Authentication*, 2005, pp. 180–187.

[5] A. Ross and R. Nadgir, "A thin-plate spline calibration model for fingerprint sensor interoperability," *IEEE Tran. on Knowledge and Data Engineering, Special Issue on Intelligence and Security Informatics*, vol. 20, no. 8, pp. 1097–1110, 2008.

[6] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, , no. 16, pp. 2639–2664, 2004.

[7] N. Poh, T. Bourlai, and J. Kittler, "Improving Biometric Device Interoperability by Likelihood Ratio-based Quality Dependent Score Normalization," in *accepted for publication in IEEE Conference on Biometrics: Theory, Applications and Systems*, Washington, D.C., 2007, pp. 1–5.

[8] Y. Chen, S.C. Dass, and A.K. Jain, "Fingerprint Quality Indices for Predicting Authentication Performance," in *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, New York, 2005, pp. 160–170.

[9] N. Poh, G. Heusch, and J. Kittler, "On Combination of Face Authentication Experts by a Mixture of Quality Dependent Fusion Classifiers," in *LNCS 4472, Multiple Classifiers System (MCS)*, Prague, 2007, pp. 344–356.

[10] K. Nandakumar, Y. Chen, S.C. Dass, and A.K. Jain, "Quality-based Score Level Fusion in Multibiometric Systems," in *Proc. 18th Int'l Conf. Pattern Recognition (ICPR)*, Hong Kong, 2006, pp. 473–476.

[11] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Kernel-Based Multimodal Biometric Verification Using Quality Signals," in *Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE*, 2004, vol. 5404, pp. 544–554.

[12] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communnications," in *12th Int'l Conf. on Image Analysis and Processing*, Mantova, 2003, pp. 2–13.

[13] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo, "Error Handling in Multimodal Biometric Systems using Reliability Measures," in *Proc. 12th European Conference on Signal Processing*, Antalya, Turkey, September 2005.

[14] J. Kittler, N. Poh, O. Fatukasi, K. Messer, K. Kryszczuk, J. Richiardi, and A. Drygajlo, "Quality Dependent Fusion of Intramodal and Multimodal Biometric Experts," in *Proc. of SPIE Defense and Security Symposium, Workshop on Biometric Technology for Human Identification*, 2007, vol. 6539.

[15] D. E. Maurer and J. P. Baker, "Fusing multimodal biometrics with quality estimates via a bayesian belief network," *Pattern Recognition*, vol. 41, no. 3, pp. 821–832, 2007.

[16] F Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Ortega-Garcia, "Dealing with sensor interoperability in multi-biometrics: The upm experience at the biosecure multimodal evaluation 2007," in *Proc. of SPIE Defense and Security Symposium, Workshop on Biometric Technology for Human Identification*, 2008.

[17] K. Kryszczuk and A. Drygajlo, "On combining evidence for reliability estimation in face verification," in *Proc. 13th European Conference on Signal Processing*, Florence, 2006.

[18] N. Poh and S. Bengio, "Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks," in *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, New York, 2005, pp. 474–483.

[19] N. Poh, T. Bourlai, J. Kittler, L. Allano, F. Alonso-Fernandez, O. Ambekar, J. Baker, B. Dorizzi, O. Fatukasi, J. Fierrez, H. Ganster, J. Ortega-Garcia, D. Maurer, A. A Salah, T. Scheidat, and C. Vielhauer, "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *IEEE Trans. on Information Forensics and Security*, 2009, accepted for publication.

[20] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. R. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J-L. Alba-Castro, E. Gonzalez-Agulla, E. Otero-Muras, S. Garcia-Salicetti, L. Allano, B. Ly-Van, B. Dorizzi, J. Kittler, T. Bourlai, N. Poh, F. Deravi, R. Ng, M. Fairhust, J. Hennebert, A. Humm, M. Tistarelli, L. Brodo, J. Richiardi, A. Drygajlo, H. Ganster, F. Sukno, S-K. Pavani, A. Frangi, L. Akarun, and A. Savran, "The multi-scenario multi-environment biosecure multimodal database (bmdb)," *IEEE Trans. on Pattern Analysis and Machine*, 2009, accepted for publication.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.

[22] F.V. Jensen, *An Introduction to Bayesian Networks*, Springer, 1996.

[23] A. Ross, K. Nandakumar, and A.K. Jain, *Handbook of Multibiometrics*, Springer Verlag, 2006.

[24] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio based biometric score fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 342–347, 2008.

[25] K. Kryszczuk and A. Drygajlo, "Improving classification with class-independent quality measures: Q-stack in face verification," in *LNCS 4472, Multiple Classifiers System (MCS)*, 2007, pp. 1124–1133.

[26] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1999.

[27] A. Martinez and A.C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[28] O. Fatukasi, J. Kittler, and N. Poh, "Quality Controlled Multimodal Fusion of Biometric Experts," in *12th Iberoamerican Congress on Pattern Recognition CIARP*, Via del Mar-Valparaiso, Chile, 2007, pp. 881–890.

[29] T. Scheidat, C. Vielhauer, and J. Dittmann, "Distance-level fusion strategies for online signature verification," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[30] N. Poh, *Multi-system Biometric Authentication: Optimal Fusion and User-Specific Information*, Ph.D. thesis, Swiss Federal Institute of Technology in Lausanne (Ecole Polytechnique Fédérale de Lausanne), 2006.