A BioSecure (DS2) Report on the Technological Evaluation of Score-level Quality-dependent and Cost-sensitive Multimodal Biometric Performance

Norman Poh, Thirimachos Bourlai, Josef Kittler, Albert Salah, Fernando Alonso John Baker Harald Ganster, Lorene. Allano, Omalara Fatukasi, Onkar Ambekar and Tobias Scheidat.

Abstract

This report summarizes the result of the BioSEcure DS2 (Desktop) evaluation campaign. This campaign aims at evaluating multimodal fusion algorithms involving face, fingerprint and iris biometrics for person authentication, targeting at the application of physical access control in a medium-sized establishment with some 500 persons. While multimodal biometrics is a well investigated subject in the literature, there exists no benchmark results on which basis fusion algorithms can be compared. Working towards this goal, we designed two sets of experiments: quality-dependent and cost-sensitive evaluations. The quality-dependent evaluation aims at evaluating how well fusion algorithms can perform under changing quality of raw biometric images principally due to change of devices. The cost-sensitive evaluation, on the other hand, aims at how well a fusion algorithm can perform given restricted computation and in the presence of software and hardware failures, resulting in errors such as failure to acquire and failure to match. Since multiple capturing devices are available, a fusion algorithm should be able to handle this non-ideal but nevertheless realistic scenario. It is on this ground that this evaluation is proposed. In both evaluations, a fusion algorithm is supplied with scores from each biometric comparison subsystems as well as the quality measures of both the template and the quality measures. The evaluation campaign is very encouraging, receiving 15 fusion systems. To the best of our knowledge, the BioSecure DS2 evaluation campaign is the first attempt to benchmark both quality-dependent and cost-sensitive fusion algorithms. Our evaluation suggests that while using all the available biometric sensors can definitely increase the fusion performance, one has to trade-off with the increased cost in terms of acquisition time, computation time and the physical cost of hardware maintenance. A promising solution which does not increase this composite factors of cost is dynamic fusion, as demonstrated in our experiments. In the presence of changing image quality which may due to change of acquisition devices and/or device capturing configurations, we observe that the top performing fusion algorithms are those that exploit the automatically derived quality measures in order to recover the most probable biometric device from which the associated biometric data was scanned.

Index Terms

multimodal biometric authentication, benchmark, database, fusion

I. Introduction

In order to improve confidence in verifying the identity of individuals seeking access to physical or virtual locations both government and commercial organizations are implementing more secure personal identification (ID) systems. The challenge of creating a well-designed, highly secure and accurate personal identification system has always been a central goal in security business. This challenge can be responded to by the use of multimodal biometric systems [1], [2], [3] where both the security and performance levels can be further increased.

Research in multimodal biometrics has entailed an enormous effort on data collection, e.g., XM2VTS [4], VidTIMIT [5], BANCA [6], BIOMET [7], FRGC [8] and the recent M3 corpus [9]. While the existence of these databases should enable one to benchmark multimodal as well as multi-algorithmic (utilizing the

The organizers of the evaluation are with CVSSP, SEPS, University of Surrey, Guildford, Surrey, GU2 7XH, U.K. E-mail: normanpoh@ieee.org, t.bourlai@surrey.ac.uk, j.kittler@surrey.ac.uk

same biometric data but different matching software) fusion algorithms, it is a necessary prerequisite but not sufficient. For instance, it is not straight forward to compare two fusion algorithms in the case where each algorithm relies on its own set of comparison subsystems. This is because an observed improvement due to a particular fusion algorithm may be due to the superior performance of its comparison subsystems rather than the merits of the fusion process itself. This shows the importance of having *common* comparison subsystems when benchmarking score-level fusion algorithms.

The fusion paradigm we are interested in is where multibiometrics is treated as a two-stage problem where in the first stage one designs a comparison system. The scores produced by the comparison subsystems are then used as inputs to a fusion classifier. This paradigm, also known as the score-level fusion, is the mainstream research pursued in the literature on multibiometrics, e.g.,[1], [2], [3]. Another fusion paradigm treats these two stages as a single process by jointly training a single model. This paradigm is more complex because it involves combining information at the raw signal or feature levels. This often results in a learning problem in the spaces of increased dimensionality. This approach is appropriate when the data types are compatible. When the data types are incompatible, e.g., when we are combining fingerprint minutiae (containing location information) with speech features of varying length, it is not obvious how to introduce a single matching function or distance measure in the resulting joint feature space. For this reason, the first fusion paradigm, i.e., score-level fusion, is considered a more practical solution to the multibiometric information fusion problem.

One way to categorize score-level fusion algorithms is by whether or not quality measures are used. These are auxiliary measurements not necessarily discriminative in distinguishing the genuine user (client) from his/her impostors but are important as descriptors of the signal quality which is affected by the way a biometric is presented, the characteristics of acquisition device and the environment in which the acquisition takes place. We refer to the fusion algorithms that use quality measures as *quality-dependent fusion algorithms* and those that do not as *conventional fusion algorithms*.

The goal of this paper is to propose two evaluation schemes to benchmark score-level fusion algorithms rigorously. These schemes are quality-dependent and cost-sensitive evaluation schemes. The first scheme is designed to evaluate quality-dependent fusion algorithms whereas the second is designed to evaluate conventional fusion algorithms. In both cases, we consider also the possibility of having two further sub-problems. The first one involves client-specific or user-dependent fusion where one can train a fusion classifier that is tailored to each identity claim. The second one involves handling the fusion problem with missing information. For instance, when one or more baseline systems are not operational due to failure to acquire or failure to match a biometric sample, we require the fusion system to be able to output a combined score. This is sensible and indeed practical because the match scores to be combined contain redundant information (i.e., each can be seen as a support for a hypothesis of the same type although their actual values may be in different ranges).

An obvious disadvantage of score-level fusion is that, by using only scores, a lot of precious non-class discriminatory information is lost, e.g., the quality of raw biometric signal. Here are two examples: a person's face can change drastically with illness, diet, or age, as well as with the application of cosmetics, a change in hair color or style, or a sun tan; and a person's voice can differ significantly with congestion caused by a cold. Quality measures are expected to provide measurements designed to capture these changes in ways that could usefully be exploited in the fusion process. In this sense, quality measures should in theory compensate for the loss of information without sacrificing the practical advantage offered by score-level fusion. In practice, however, tapping the quality information – which is non-discriminatory in nature – in order to improve the classification performance is not a trivial problem.

In a cost-sensitive evaluation scheme, one considers a fusion task as an optimization problem whose goal is to achieve the highest performance with a minimal cost. We refer to "cost" as the price paid for acquiring and processing more information, e.g., requesting more samples from the same device or using more biometric devices (which entails longer processing time), and as a result of making wrong false acceptance and false rejection decisions. In this situation, a decision can be made even if not all the sub-system match scores are observed. Such an approach is different from the conventional multimodal or multi-algorithmic fusion algorithms which require all the scores to be available. Therefore, the proposed cost-based evaluation effectively considers the redundancy of multimodal or multi-algorithmic information, a subject that has not been adequately addressed in the literature on multimodal biometrics.

Towards these two objectives, we constructed a database with scores as well as quality measures for each access using the BioSecure database. To date, there exists no test bed for *quality-dependent fusion algorithms* nor *cost-sensitive* evaluation. They are therefore the novelties advanced by this paper. A similar work in this direction is [10] but the quality measures are not available.

The data used in our evaluation scheme is taken from the BioSecure database. *BioSecure*[1] is a European project whose aim is to integrate multi-disciplinary research efforts in biometric-based identity authentication. Application examples are a building access system using a desktop-based or a mobile-based platform, as well as applications over the Internet such as tele-working and Web or remote-banking services. As long as the data collection is concerned, three scenarios have been identified, each simulating the use of biometrics in remote-access authentication via the Internet (termed the “Internet” scenario), physical access control (the “desktop” scenario), and authentication via mobile devices (the “mobile” scenario). While the desktop scenario is used here, the proposed two evaluation schemes can equally be applied to the remaining two data sets.

The desktop scenario data set contains the following biometric modalities: signature, face, audio-video (PINs, digits, phrases), still face, iris, hand and fingerprint. However, only still face, iris and fingerprint are used for the evaluation schemes proposed here. This data set is collected from six European sites (only four are being used at the writing of this report). Although the data acquisition process is supervised, the level of supervision is extremely different from site to site (see Figure 1). This database contains two sessions of data separated by about one month interval. In each session, two biometric samples are acquired per modality per device, hence resulting in 4 samples per modality per device over the two sessions. There are several devices for the same biometric modality. The forgery data collected simulate PIN-reply attacks and imitation of dynamic signature (with several minutes of practice and with the knowledge of the signature dynamics). The volunteers are selected to have both genders in somewhat equal proportions and of ages with the following distribution: 2/3 in the range 18–40 of age and 1/3 above 40.

The paper is organized as follows: xxx

II. The BioSecure DS2 Data Set and Reference Systems

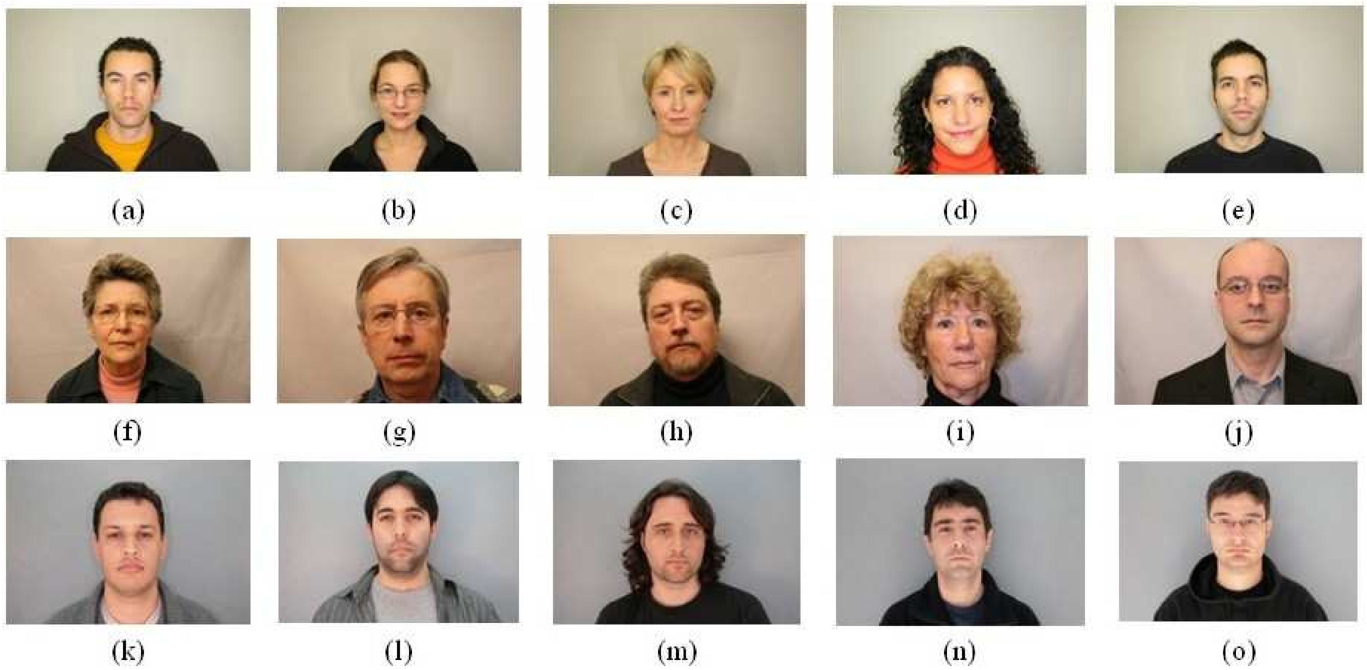
Table I presents the 17 channels of data available. A *channel* of data is composed of a biometric modality acquired by a biometric device in a particular configuration. For example, a left index fingerprint acquired using an optical fingerprint sensor is considered a channel of data. Using the notation presented in Table I, this channel of data is referred to as “fo5”. The 17 channels of data are: fa1, fnf1, fwf1, ir1, ir2, fo1, fo2, fo3, fo4, fo5, fo6, ft1, ft2, ft3, ft4, ft5 and ft6.

Each channel of data was collected in two sessions, separated by about one month’s interval. In each session, two biometric samples were acquired for each channel of data. Therefore, for each person, four biometric samples are available per channel of data.

While there are 17 channels, we need only three reference systems, corresponding to the three chosen biometric modalities, i.e., face, fingerprint and iris. We also need three pieces of software to extract their respective quality measures directly from the acquired images. Table II lists the reference systems of the three biometric modalities as well as their respective quality measures.

Among the 14 quality measures, six are face-related quality measures (hence relying on a face detector), i.e., face detection reliability, spatial resolution between eyes, presence of glasses, rotation in plane, rotation in depth and degree of frontal face. The remaining eight measures are general purpose image quality measures as defined by the MPEG standards. These quality measures were obtained using Omniperception’s proprietary Affinity SDK.

[1] http://www.biosecure.info


(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

Fig. 1. Cross-site data diversity. Each row represents the data collected from three different sites. Even though the acquisition system setup was the same the quality of the data collected varies.

TABLE I

A LIST OF CHANNELS OF DATA FOR EACH BIOMETRIC MODALITY CAPTURED USING A GIVEN DEVICE.

Label	template ID {n}	Modality	Sensor	Remarks
fa	1	Still Face	web cam	Frontal face images (low resolution)
fnf	1	Still Face	CANON	Frontal face images without flash (high resolution)
fwf	1	Still Face	CANON	Frontal face images with flash (high resolution)
ir	1–2	Iris image	LG	1 is left eye; 2 is right eye
fo	1–6	Fingerprint	Optical	1/4 is right/left thumb; 2/5 is right/left index; 3/6 is right/left middle finger
ft	1–6	Fingerprint	Thermal	1/4 is right/left thumb; 2/5 is right/left index; 3/6 is right/left middle finger

For example, fo2 refers to the channel of data due to acquiring right index fingerprints using an optical fingerprint sensor. The web cam model is Phillips SPC 900. The model of CANON digital camera is EOS 30D. The iris capturing device is LG3000. The thermal sensor acquires a fingerprint as one sweeps his/her finger over it. The optical sensor acquires a fingerprint impression by direct contact (no movement required). This table results in 17 channels of scores. The actual data collected under the desktop scenario contains also audio-visual web cam (hence talking faces), signature and hand images but these channels of data are not used for evaluation. For each channel of data, two sessions of data acquisition were conducted. In each session, two biometric samples were collected.

There is only a fingerprint quality measure and it is based on the implementation found in [11]. It is an average of local image patches of fingerprint gradient. When too much pressure is applied during fingerprint acquisition, the resulting fingerprint image usually has low contrast. Consequently, a minutia-based fingerprint matcher (which is the case for the NIST fingerprint system) is likely to under perform with this type of image. The converse is also true for high contrast and clear fingerprint images.

Three iris quality measures are used. The first one, i.e., texture richness measure, is obtained by a weighted sum of the magnitudes of Mexican hat Wavelet coefficients as implemented in [12]. The other two quality measures are functions of estimated iris and pupil circles. The first one is the difference between iris diameter and pupil diameter. If this difference is small, the iris area to be matched will be small, hence implying that the match scores may not be reliable. The second measure is the proportion of iris used for matching which is one minus the proportion of a mask with respect to the entire iris area.

TABLE II

REFERENCE SYSTEMS AND QUALITY MEASURES ASSOCIATED TO EACH TO BIOMETRIC MODALITY CAPTURED BY A GIVEN SENSOR

Modality	Reference systems	Quality measures
Still Face	Omniperception's Affinity SDK face detector; LDA-based face verifier	face detection reliability, brightness, contrast, focus, bits per pixel, spatial resolution (between eyes), illumination, degree of uniform background, background brightness, reflection, glasses, rotation in plane, rotation in depth and degree of frontal face (from Omniperception's Affinity SDK)
Fingerprint	NIST Fingerprint system	texture richness [11] (based on local gradient)
Iris	A variant of Libor Masek's iris system	texture richness [12], difference between iris and pupil diameters and proportion of iris used for matching

A mask is needed to prevent matching on areas containing eyelashes and specular lights, for instance. Unfortunately, due to bad iris segmentation, and possibly suboptimal threshold to distinguish eyelashes from iris, our iris baseline system is far from the performance claimed by Daugman's implementation [13].

III. THE EVALUATION PROTOCOLS

The current release of the desktop scenario contains 333 persons. A newer version, yet to be released, contains some 500 persons. For each person, four samples per channel of data are available. The first sample of the first session is used to build a biometric template. The second sample of the first session is used as a query 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. A *template* is the data sample used to represent the claimed identity whereas a *query* is the sample with which the template is compared. The impostor scores are produced by comparing all four samples originating from another population of persons excluding the reference users.

It is important to distinguish 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. An important distinction between the two is that the population of users in these two data sets are *disjoint*. This ensures that the performance assessment is unbiased. There are 51 genuine users in the development set and 156 in the evaluation set. These two sets of users constitute the 206 users available in the database. The remaining 126 subjects are considered as an external population of users who serve as zero-effort impostors. The next two paragraphs explain the development and evaluation impostor score sets.

The *development impostor score set* contains 103×4 samples, i.e., 103 persons and each contributes 4 samples. When the reference subject is considered a template, all the 4 samples of the remaining half of the 206 subjects are considered impostors in the development set in Session 1. The other half of 206 subjects 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.

Note that the *evaluation impostor score set* contains 126 subjects set apart 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.

Table III summarizes the explanation of the genuine user and impostor score sets of the development and evaluation data sets. The exact number of accesses differs from that listed in this table because of missing observation as a result of 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, a failure to extract quality measures will result in a vector containing a series of "-999".

TABLE III

THE EXPERIMENTAL PROTOCOL OF THE BIOSECURE DS2 DATABASE. S1/S2=SESSION 1 AND 2.

Data sets		No. of match scores per person	
		dev. set (51 persons)	eva. set (156 persons)
S1	Gen	1	1
	Imp	103×4	126×4
S2	Gen	2	2
	Imp	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.

Although the desktop scenario involves supervised data acquisition, the level of supervision differs from one collection site to another. As a result, there exists a site-dependent bias in terms of performance and this bias is readily observable from the captured images for face and fingerprint biometrics (not shown here).

IV. PRELIMINARY ANALYSIS

A. Performance on Session Mismatch and Device Mismatch

In this section, we compare the possible existence of bias between the same-session versus different-session performance. Very often, data in a single session exhibit low intra-device signal quality variation but as soon as data is collected in different sessions (i.e., different visits separated by several days, weeks or months), high intra-device signal quality variation may be observed. This will affect the resulting system performance. We compared the performance of the Session 1 data versus that of Session 2 on the development set (with 51 users). Recall that the template of each user is constructed from data in Session 1. Hence, the performance calculated on Session 1 represents an *intra-session* performance whereas that of Session 2 represents an *inter-session*q performance. We did so for all the 24 channels of data (see Figure 2). The first 17 channels of data involve matching using the same device. The remaining 7 channels involve matching templates and query images acquired from different devices. “xfa1” means the device-mismatched version of “fa1”, i.e., the templates are acquired using “fnf1” (the Canon digital camera) and the queries are acquired using “fa1” (a web cam). Similarly, for the fingerprint channels, “xft$\{n\}$” refers to the templates acquired using “fo$\{n\}$” (optical device) but the queries are acquired using “ft$\{n\}$” (thermal device by sweeping a fingerprint over it), for $n \in [1, 2, 3]$ denoting the following three right fingers: thumb, index and middle fingers, respectively. An important observation is that the intra-session performance is almost always better than the inter-session performance. This shows that the intra-session performance is likely to be biased and should not be used for performance evaluation.

The experiments here also allow us to compare device mismatch, i.e., “fa1” versus “xfa1” (“x” for cross device matching) and “ft$\{n\}$” versus “xft$\{n\}$” for $n \in [1, 2, 3]$, each corresponding to a right hand finger mentioned previously. In each of the experiments of “fa1” versus “xfa1” and “ft$\{n\}$” versus “xft$\{n\}$” for all n, the query images are taken with the same device but the templates used are acquired using a different device. For the face experiment, the template images are acquired using a digital camera (hence giving images of higher quality) whereas the query images are acquired using a web cam (of lower quality). Similarly, for the fingerprint experiments, the template images are acquired using an optical scanner (giving better performance) whereas the query images are acquired using a thermal sensor (giving slightly worse performance). As can be observed, the performance with device mismatch is always worse than that with common device, even if the templates used may be of a higher quality.

B. Analysis of Quality Measures

In order to verify that quality measures of our database are useful, we propose to evaluate their utility in distinguishing the acquisition device. For example, if one knows the device or can infer the device given

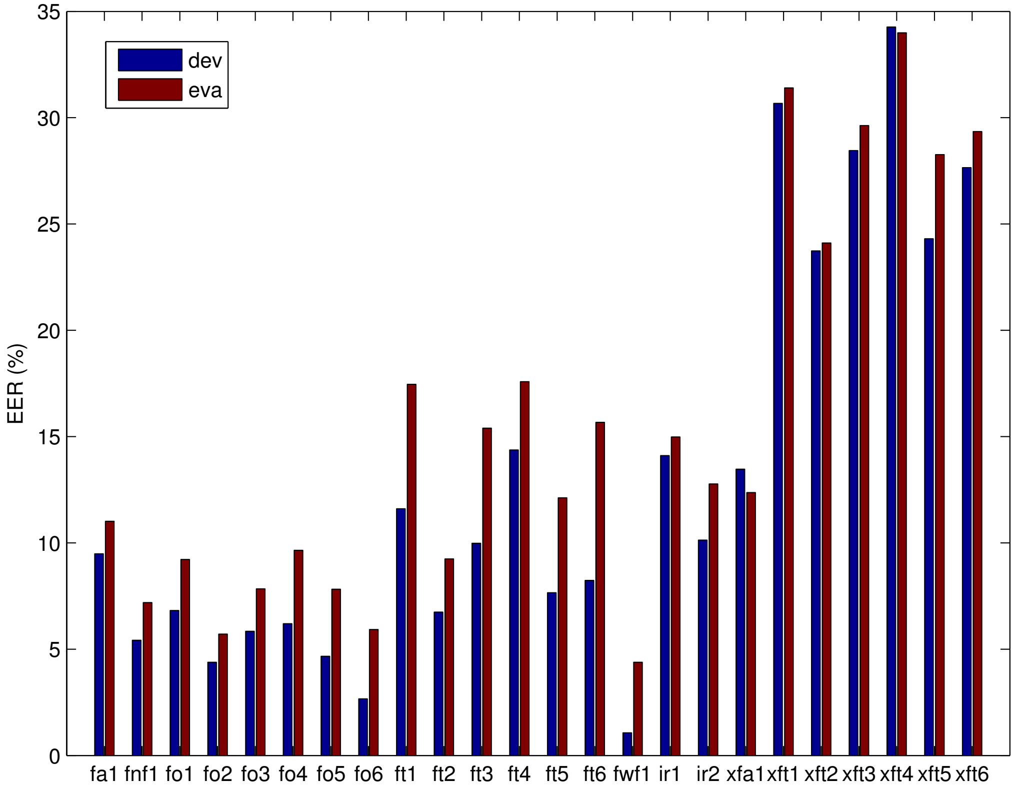


Fig. 2. Performance of Session 1 (blue; left bar) vs. Session 2 (red; right bar) in terms of EER for the 24 channels of data. The first 17 channels of data involve matching using the same device. The remaining 7 channels involve matching templates and query images acquired by different devices. The performance of Session 1 is labeled as "dev" whereas that of Session 2 as "eva".

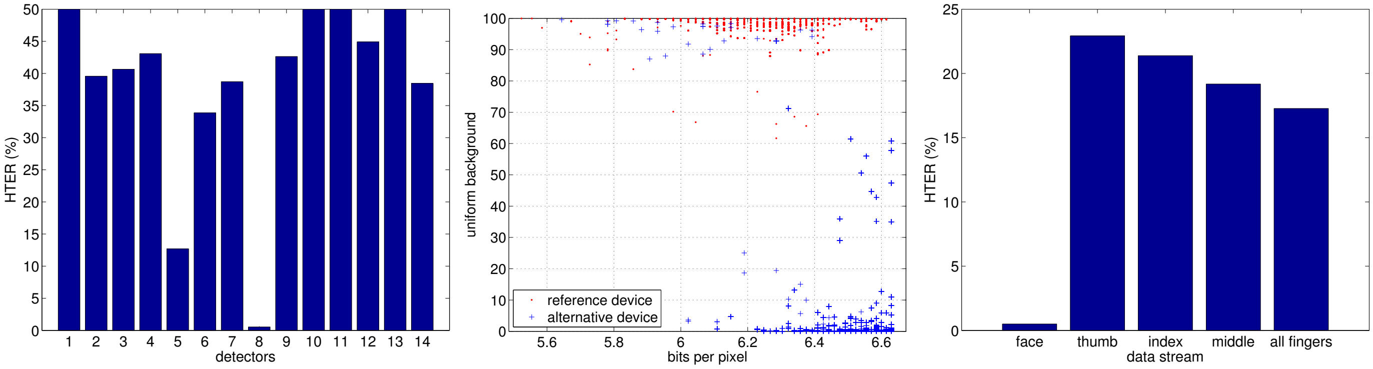
the observed quality measures, one can construct a device-specific fusion algorithm. We constructed a Bayes classifier to solve this problem by estimating the posterior probability of a device d given a vector of quality measures q, i.e.,

$$P(d|q) = \frac{p(q|d)P(d)}{\sum_{d_*} p(q|d_*)P(d_*)}$$

where $P(d)$ is the prior probability of a device, $p(q|d)$ is the density of q given d and the denominator is a normalizing term to ensure that the sum of $p(d_*|q)$ over all possible d_*'s equals to one. We use d_* as a variable that loops through all possible devices and d to denote a particular device whose posterior is being evaluated.

For this preliminary experiment, we used the quality measures of the development set and measured the performance of $P(d|q)$ on the evaluation set as specified in Table III. We did so for each of the 14 face quality measures in order to distinguish between images taken with a digital camera from those taken with a web cam. The results are shown in Figure 3(a). As can be observed, uniform background is the most discriminative quality measure and this is followed by bits per pixel. This is perfectly reasonable considering that the web cam images were taken in a more cluttered background whereas the digital camera images are mugshot images taken in a uniform background and under controlled lighting, thus conforming to most international passport standards. Due to higher color resolution, the images taken with a digital camera have higher average number of bits per pixel (over the entire image). We show a scatter plot of these two quality measures for each device in Figure 3(b). A classifier trained on this problem gives an EER (assuming equal prior) of 0.5%. This result is shown in the first bar of Figure 3(c).

The remaining four bars in Figure 3(c) are the classification performance in distinguishing between a thermal and an optical sensor given the fingerprint quality measure, i.e., the sum of local gradients, derived from fingerprint images. Among them, the first three are the performance of $p(d|q)$ where the devices



(a) face device classification given individual quality measures (b) Scatter plot of bits per pixel vs. uniform background (c) device classification for face and fingers

Fig. 3. The performance in terms of HTER (with threshold chosen to minimize EER on the development set) of each of the 14 face quality measures in discriminating 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.

can be either thermal or optical and the quality measure q is texture richness [11]. The performance of $p(d|q)$ for each of the three fingers are in the range of 18–23%. If we had three fingerprint samples from these three respective fingers for each access and we assumed that the *same* acquisition device was used, we could take the product of $p(d|q)$ for each image, i.e., $\prod_{i=1}^{3} p(d|q_i)$ since each measurement is independent. This results in the last error bar of Figure 3(c) (denoted by "all fingers"), giving about 17% of EER. Obviously, more independent observations improve the estimate of $p(d|q)$. Our main message here is that *automatically derived quality measures can be potentially used to distinguish devices*. Note that in our experiments, the quality measures were not designed specifically to distinguish the devices for this database. While not all quality measures appear to be useful on their own (as illustrated in Figure 3(a)), given some development data, an *array* of quality measures used jointly would certainly be necessary to distinguish a multitude of possible devices in the framework of $p(d|q)$.

V. The Two Evaluation Schemes

A. *Cost-Sensitive Evaluation*

The cost-sensitive evaluation was designed with two goals:

1) to assess the robustness of a fusion algorithm when some match scores and/or quality measures are not present; this is typically due to failure to acquire and/or failure to match.
2) to test how well a fusion algorithm can perform with minimal computation and hardware cost.

Note that a "cost" can also be associated with the time to acquire/process a biometric sample. Hence, longer time implies higher cost, and vice versa.

Assigning a cost to a channel of data is a very subjective issue. In this matter, we follow the following rules:

- If a device is used at least once, a fusion algorithm will be charged a unit cost. While in reality, different devices may have different costs, we opted for a unit cost because it is simply difficult to arrive at a consensus otherwise.
- The subsequent use of the same device will be charged a 0.3 cost instead considering the fact that the same hardware is being reused.
- A device is considered used if a fusion algorithm acquires a sample for subsequent processing, i.e., to extract quality measures and/or to obtain a match score. This is regardless of whether the resulting match score will actually contribute to the final combined score.

Through the cost-sensitive evaluation, the design of a fusion algorithm becomes more challenging because the task now is to maximize recognition performance *while* minimizing the cost associated to the device usage. In this respect, there exists two strategies to solve this problem: static and dynamic. A static solution pre-selects a set of channels and use them for all access requests. A dynamic solution, on the other hand, may use different channels for different access requests.

B. Cross-device Quality-dependent Evaluation

The goal of this evaluation scheme is to assess the ability of a fusion algorithm in selecting the more reliable channel of data given quality measures derived from biometric data. The task is made more challenging with cross-device matching, i.e., a matching can occur between a biometric template acquired using one device and a query biometric data using another device. In our case, the template data is always acquired using a high quality device (giving better verification performance) and the query data may be acquired using a high or a low quality device. Note that cross device matching occurs only in the latter case. The channels of data considered are face and the three right fingerprints, denoted as fnf, fo1, fo2 and fo3. In case of cross device matching, these channels are denoted as xfa, xft1, xft2 and xft3. The development set consisting of all 8 channels was distributed to the participants. The (sequestered) evaluation set, on the other hand, contains only four channel of data as a result of mixing fnf/xfa and fo$\{n\}$/xft$\{n\}$ for all $n \in \{1, 2, 3\}$. These four channels of data are made up of the following combinations:

- fnf and fo$\{n\}$
- fnf and xft$\{n\}$
- xfa and fo$\{n\}$
- xfa and xft$\{n\}$

A fusion algorithm does not not know from which device a biometric sample was acquired since the identity of the device is unknown. This is a realistic scenario because as a biometric technology is deployed, it may be replaced by a newer device. Furthermore, it configuration may change, hence, resulting in its acquired query biometric data to be significantly different from the previously stored template data. This fusion problem is challenging because each of the four combinations require a different fusion strategy in order to achieve the optimal result.

VI. Brief Description of Submitted Systems

The call for participation has resulted in ten submissions, including the baseline fusion technique, i.e., a Naive Bayes classifier which is provided by the evaluation organizer, the University of Surrey (UniS).

A. Classifiers categorization and Notations

Let $y_i \in \mathbb{R}$ be the output of the i-th base system and there are N base system outputs, i.e., $i \in \{1, \dots, N\}$. For simplicity, we denote $\mathbf{y} = [y_i, \dots, y_N]'$. The most commonly used fusion classifier in the literature takes the following form:

$$f : \mathbf{y} \to y_{com} \tag{1}$$

where $y_{com} \in \mathbb{R}$ is a combined score. We shall refer to this classifier throughout this report as a *conventional* fusion classifier.

If the function f takes into consideration of the signal quality, then f is considered a *quality-dependent* fusion classifier. Let the signal quality be represented by the vector of Q measurements, $\mathbf{q}_i \in \mathbb{R}^Q$. For simplicity, we denote $\mathbf{q}$ as a concatenation of all $\mathbf{q}_i$'s, i.e., $\mathbf{q} = [\mathbf{q}_1, \dots, \mathbf{q}_N]'$. The function f in this case takes the following form:

$$f : \mathbf{y}, \mathbf{q} \to y_{com} \tag{2}$$

Any fusion classifier can be categorized into one of the two forms just mentioned. The function f can be a generative or a discriminative classifier. In the former, class-dependent densities are first estimated and decisions are taken using the Bayes rule or the Dempster-Shafer rule. In the latter, the decision boundary

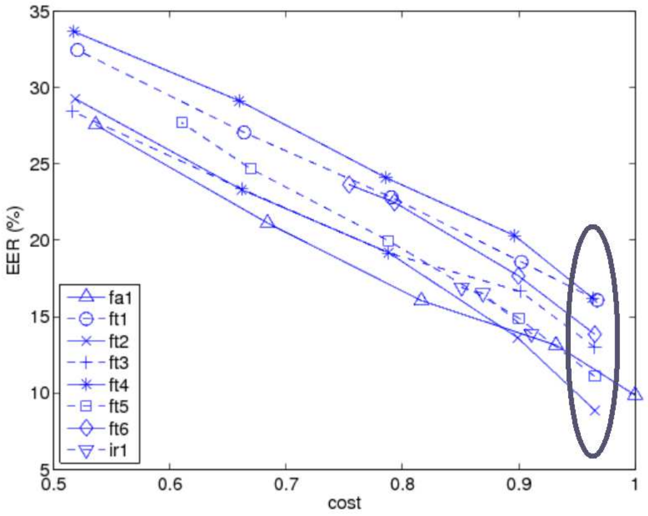


Fig. 4. The baseline performance of comparison subsystems in EER (%) versus the average cost per access.

is directly estimated. A common characteristic of both types of classifiers is that the dependency among observations (scores or quality measures) are considered.

There exists also another approach that we will refer to as the *normalization-based approach* constructs a fusion classifiers in two stage. In the first stage, the match scores of each comparison subsystem are independently transformed into a common domain, e.g., in the range of $[0, 1]$. In the second stage, the resulting normalized scores of all comparison subsystems are combined using a fixed rule such as sum or product [14].

B. Submitted Systems

The submitted systems are very diverse in nature, covering from generative and discriminative classifiers to normalization-based classifiers. While most submitted generative classifiers make decisions based on the Bayes rule (hence resulting in Bayesian classifiers), the submission from Joanneum Research used Dempster-Shafer theory instead. The submitted discriminative classifiers include Support Vector Machines and logistic regression. The submitted classifiers along with their main characteristics are listed in Table IV. Since the focus of this report is on reporting the results, each algorithm will not be elaborated in great details. The interested reader should refer to the relevant citations.

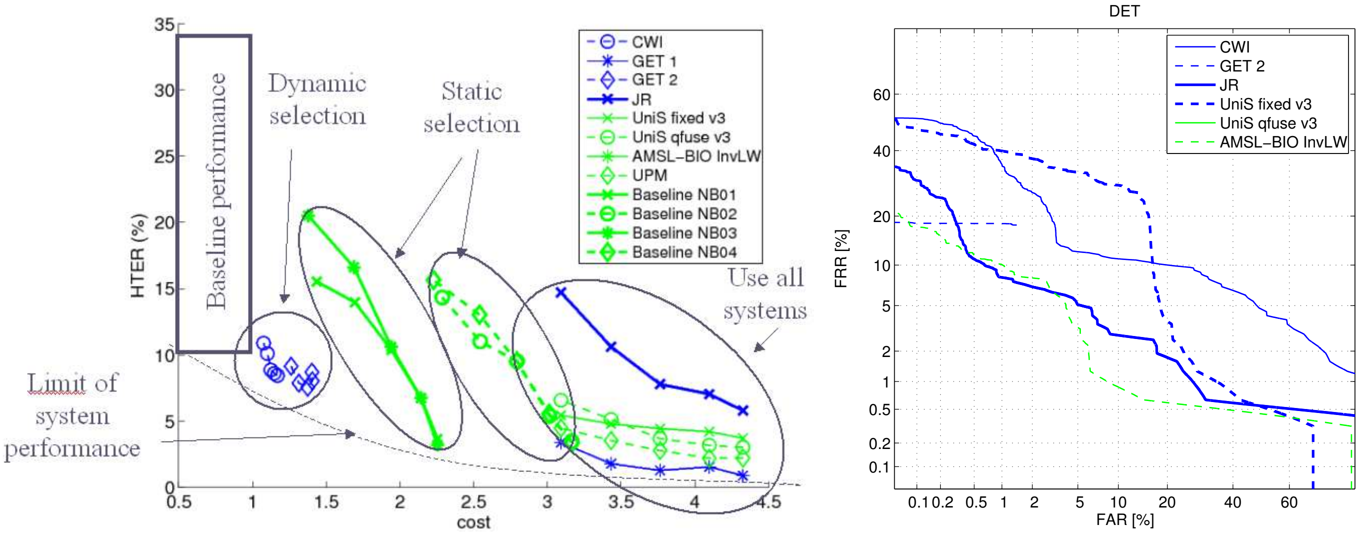
VII. Experimental Results

This section presents the results of the two evaluations, i.e., cost-sensitive and quality-based evaluations.

A. Cost-Sensitive Evaluation

Figure 4 shows the baseline performance of comparison subsystems in EER (%) versus the average cost per access. Each curve represents a channel gradually degraded with by deleting some entries, hence simulating missing scores and quality measures. A deleted entry is considered a failed access request, hence resulting in a reject decision. Since an entry is not observed, no unit cost is assigned to the entry, hence, resulting in reduced cost as more entries are deleted. The consequence is that lower average cost per access is followed by increased verification error mainly due to false rejections of genuine claims. The oval shows that all the fingerprint channels of data contain missing values even before deliberately deleting some entries/access claims; only the face channel has complete data.

The rectangle at the top left corner of in Figure 5(a) shows the range of performance of the baseline comparison subsystems. The submitted fusion systems naturally shows four clusters of data which are



(a) Fusion system performance versus average cost per access

(b) DET curves of the systems using all accesses

Fig. 5. Performance of all submitted fusion systems participating in the cost-based evaluation.

marked with ovals. The left most cluster contains algorithms with dynamic selection of channels. The two middle clusters are static fusion systems which use different number of channels. The left one uses one system less than the right one, hence, explaining the difference of one unit of cost between the two. The right cluster achieves the lowest HTER, hence, the best performance, but at the highest cost per access. Finally, the bottom dashed curve shows the general trend of the limit of fusion performance which decreases in HTER, hence better performance, as the cost per access increases. This is sensible because increased cost implies using more information. Adding more information that is complementary to what is already available will certainly improve the fusion performance, but at a higher cost.

Finally, in order to show the performance in DET, we selected only the systems in the last cluster to plot the DET curves because their average cost per access are comparable. Note that the DET curves of GET 2 and UniS qfuse could not be plotted because they contain too many repeated values of zeros and ones. It should be cautioned that these DET curves are not directly comparable because when computing these DET curves, the decision thresholds associated to each fusion system are optimized on the test scores. Consequently, these curves may be optimistically biased. Nevertheless, AMSL-BIO InvLW is probably the best system near EER. The HTER curve of AMSL-BIO InvLW was not shown in Figure 5(a) because during the submission, a suboptimal threshold was chosen, resulting in HTER near 50%. A correctly chosen threshold would have given HTER between 5% and 15% of HTER depending on the cost. This shows the importance of choosing the right threshold.

B. Cross-device Quality-dependent Evaluation

For the cross-device quality dependent evaluation, we first assessed the baseline performance of the four channels of data, i.e., fnf1/xfa1, fo1/xft1, fo2/xft2 and fo3/xft3, where fnf1/xfa1 means that in this channel of data, the query face images may be captured by a digital camera (fnf1) or a web-cam (xfa1). The fnf1 channel has a higher image quality than the xfa1 channel. The template images were captured using a digital camera. Matching the template with the xfa{n} is considered a cross-device matching. Recall that fo{n} means fingerprint images captured using an optical sensor whereas xft{n} means the same images but captured using a thermal sensor by sliding a finger over the sensor. An ideal fusion system should consider from which device the query images are captured and use the right fusion strategy given only the scores and quality measures.

Figure 6 shows the baseline performance of the comparison subsystems (prior to fusion). Similar to the cost-sensitive evaluation, here, some entries of the data are gradually deleted, resulting in the proportion of missing data being increased from 0% to up to 40%. The access requests with missing observations of scores/quality measures are automatically rejected by the system. As a result, as the proportion of missing data is increased, more and more false rejections of genuine accesses occur, resulting in increased EER.

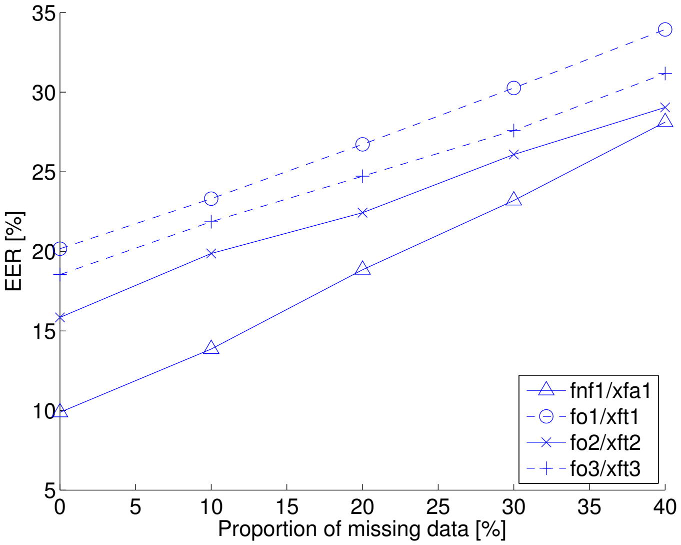


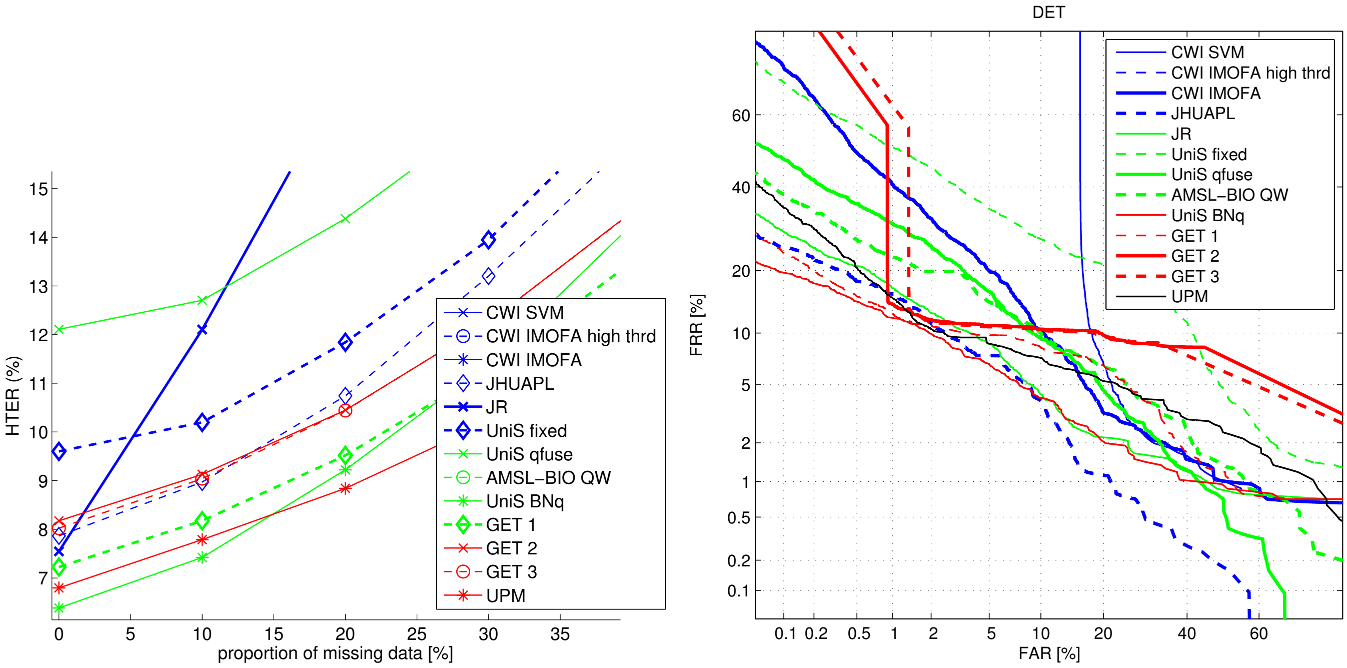
Fig. 6. Performance of comparison subsystems

The fusion performance of the submitted systems is shown in Figure 7(a). In this assessment, all the observable channels of data are used, contrary to the cost-sensitive evaluation. Our focus here is to assess how well a fusion system can perform with changing image quality and in the presence of gradual missing data. As can be observed, the top two systems are UniS BNq and UPM. These systems are device-specific, meaning that they first estimate how probable the channel of data is from the observed quality measures and then pick the right fusion function. The next best system is GET 1, which is a Bayesian classifier whose class-conditional densities are estimated using a mixture of Gaussian components. This system does not use the provided quality measures and so does not change its fusion strategy under cross-device matching. Some of the systems that actually use the quality measures are not among the best systems because they did not use the right threshold. Such is the case for JHUAPL. To assess the performance independently of the decision threshold, we plotted the DET curves of all the systems with 0% of missing data in Figure 7(b). Here, the JHUAPL fusion system performs well under high FAR whereas UniS-BNq dominates under high FRR. The winning systems are those that exploit the quality information. This experiment shows that quality measures can be used to mitigate the effect of cross-device mismatch in the context of multimodal biometrics.

VIII. Conclusions

The BioSecure DS2 evaluation campaign aims at assessing fusion algorihms under restricted computational resources, possible hardware failures and changing image quality. This campaign gathered some 15 fusion algorihtms and were evaluated using the BioSecure DS2 database constructed for the application of phyiscal access control of a medium-sized institute. The diverse nature of submitted fusion algorithms along with carefully designed experimental protocols makes this report a valuable reference. The submitted algorithms include generative classifiers, e.g., Bayesian belief network, Bayesian classifiers and Dempster-Shafer rules of combination; discriminative classifiers, e.g., support vector machines and logistic regression; and normalization-based classifier combiners (using sum or product rules).

Our findings suggest that while using all the available biometric sensors can definitely increase the fusion performance, one has to trade-off with the increased cost in terms of acquisition time, computation time and the physical cost of hardware maintenance. A promising solution which does not increase this



(a) Fusion system performance versus proportion of missing data

(b) Fusion system performance in DET

Fig. 7. While (a) shows the performance of fusion system in HTER (with *a priori* chosen threshold on the provided development set) when some data is missing, (b) shows only the DET curves when no data is deleted.

composite cost factor is dynamic fusion, as demonstrated in our experiments. A dynamic fusion strategy utilizes only the amount of information needed in order to arrive at the most confident decision. It can save computational costs because for most access requests, only a few comparison subsystems are needed. In the presence of changing image quality which may due to change of acquisition devices and/or device capturing configurations, we observe that the top performing fusion algorithms are those that exploit the automatically derived quality measures in order to recover the most probable biometric device from which the associated biometric data was scanned.

Acknowledgments

This work was supported partially by the prospective researcher fellowship PBEL2-114330 of the Swiss National Science Foundation, by the BioSecure project (www.biosecure.info) and by Engineering and Physical Sciences Research Council (EPSRC) Research Grant GR/S46543. This publication only reflects the authors' view.

References

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

[2] C. Sanderson, *Automatic Person Verification Using Speech and Face Information*, Ph.D. thesis, Griffith University, Queensland, Australia, 2002.

[3] K. Nandakumar, *Integration of M]ultiple Cues in Biometric Systems*, Ph.D. thesis, Michigan State University, 2005.

[4] J. Lüttin, "Evaluation Protocol for the XM2FDB Database (Lausanne Protocol)," Communication 98-05, IDIAP, Martigny, Switzerland, 1998.

[5] Conrad Sanderson, "The VidTIMIT Database," Communication 06, IDIAP, 2002.

[6] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *LNCS 2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA 2003*. 2003, Springer-Verlag.

[7] S. Carcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacrtaz, "BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities," in *LNCS 2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, Guildford, 2003, pp. 845–853.

[8] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the Face Recognition Grand Challenge,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 947–954.

[9] H. Meng, P.C. Ching1, T. Lee1, M. W. Mak, B.Mak, Y.S. Moon, X. Tang M.-H. Siu, H. P.S. Hui, A. Lee, W-K. Lo, B. Ma, and E. K.T. Sioe, “The multi-biometric, multi-device and multilingual (m3) corpus,” in *Workshop on Multimodal User Authentication (MMUA 2003)*, Toulouse, 2006.

[10] N. Poh and S. Bengio, “Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication,” *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, February 2005.

[11] 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.

[12] Y. Chen, S. Dass, and A. Jain, “Localized iris image quality using 2-d wavelets,” in *Proc. Int’l Conf. on Biometrics (ICB)*, Hong Kong, 2006, pp. 373–381.

[13] J. Daugman, *How Iris Recognition Works*, chapter 6, Kluwer Publishers, 1999.

[14] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas, “On Combining Classifiers,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[15] 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.

[16] A.A. Salah and E. Alpaydın, “Incremental Mixtures of Factor Analysers,” *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR’04) Volume 1-Volume 01*, pp. 276–279, 2004.

[17] 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.

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

[19] 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.

[20] 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.

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 2001.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.

[23] 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., (one of the six best papers), 2007.

TABLE IV

A LIST OF SUBMITTED SYSTEMS

System name	Q	C	Characteristics
AMSL-BIO InvLW	$\times$	$\surd$	normalization based, weighted sum rule is used with weights being inversely proportional to Equal Error Rate empirically calculated from the development set. See [15].
AMSL-BIO QW	$\surd$	$\times$	normalization based, weighted sum rule is used with weights being inversely proportional to the quadratic term of Equal Error Rate empirically calculated from the development set. See [15].
CWI SVM	$\surd$	$\surd$	discriminative classifier, quality-independent, train with high quality data only
CWI IMOFA	$\surd$	$\surd$	Bayesian classifier (generative) whose class-conditional densities are each independently estimated using an Incremental Mixtures of Factor Analyzers [16]. The classifier estimates the posterior probability of client given the observed scores. It is quality-independent and uses sequential strategy with double thresholding (for cost-sensitive evaluation).
JHUAPL	$\surd$	$\times$	Bayesian belief network (generative classifier), quality-dependent, use all available scores. See [17]
GET 1	$\surd$	$\surd$	Bayesian classifier with equal class priors, realized using $P(\mathtt{C}\vert y) = \frac{p(y\vert\mathtt{C})}{p(y\vert\mathtt{C}) + p(y\vert\mathtt{I})}.$ The density $p(y\vert k)$ is estimated using a mixture of Gaussian components [18]. In case of missing values, scores are independently normalized into $[0,1]$ using the empirically observed minimum and maximum values. The average rule is used to combine the normalized scores.
GET 2 (cost)	$\times$	$\surd$	Similar to GET 1 (cost) except that scores are taken into account sequentially until all channels of data are exhausted or the combined score is smaller than lower threshold or greater than a higher threshold.
GET 2 (quality)	$\surd$	$\times$	Similar to GET 1 except that each fusion classifier is designed specifically for each of the four cases as mentioned in Section V-B.
GET 3 (quality)	$\surd$	$\times$	This is the enhanced version of GET 2 (quality) with double thresholding.
UniS qfixed	$\surd$	$\surd$	normalization-based, quality-dependent, selectively switches between sum and product rule depending on the quality measures. See [19].
UniS qfused	$\surd$	$\surd$	logistic regression (discriminative) modeling the posterior probability of being a client given scores and quality measures as observation, i.e., $P(C\vert y,q)$. See [20].
UniS Naive Bayes	$\times$	$\surd$	Discriminative classifier, modeling $\prod_i P(C\vert y_i)$ where $P(C\vert y_i)$ is estimated using logistic regression. See [21] for Naive Bayes and [22] for logistic regression.
UniS BNQ	$\times$	$\surd$	Generative classifier that estimates the device identity directly, i.e., $y_i^{norm} = \log \frac{\sum_d p(y\vert\mathtt{C},d)p(d\vert q)}{\sum_d p(y\vert\mathtt{I},d)p(d\vert q)}.$ The sum rule is used to combine all y_i^{norm}'s that are observable. See [23].
UPM (cost)	$\times$	$\surd$	normalization based, with each normalization function (one for each channel of data) being a logistic regression, mapping scores into log-likelihood ratios. The sum rule is used to combine the scores.
UPM (quality)	$\surd$	$\times$	Similar to its cost-based counterpart, except that the normalization function is device-specific for the face and device-independent for the fingerprint. The sum rule is used to combine the scores.

Legend: Q = Quality-based evaluation; C = cost-sensitive evaluation
$\surd$ means used in quality-based and/or cost-sensitive evaluation; $\times$ means not used
AMSL-BIO is a submission from University of Magdeburg; GET from Institut National des Tlcommunications; UPM from Universidad Autonoma de Madrid; UniS from University of Surrey; JR from Joanneum Research; CWI from Centrum voor Wiskunde en Informatica; and JHUAPL from Johns Hopkins University Applied Physics Laboratory.