A Multimodal Biometric Test Bed for Quality-dependent, Cost-sensitive and Client-specific Score-level Fusion Algorithms

Norman Poh, Thirimachos Bourlai and Josef Kittler

***Abstract*— This paper presents a test bed, called the Biosecure DS2 score-and-quality database, for evaluating, comparing and benchmarking score-level fusion algorithms for multimodal biometric authentication. It is designed to benchmark *quality-dependent*, *client-specific*, *cost-sensitive* fusion algorithms. A quality-dependent fusion algorithm is one which attempts to devise a fusion strategy that is dependent on the biometric sample quality. A client-specific fusion algorithm, on the other hand, exploits the specific score characteristics of each enrolled user in order to customize the fusion strategy. Finally, a cost-sensitive fusion algorithm attempts to select a subset of biometric modalities/systems (at a specified cost) in order to obtain the maximal generalization performance. To the best of our knowledge, the BioSecure DS2 data set is the first one designed to benchmark the above three aspects of fusion algorithms. This paper contains some baseline experimental results for evaluating the above three types of fusion scenarios.**

***Index Terms*— multimodal biometric authentication, benchmark, database, fusion**

I. Introduction

A. Motivations

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. Deploying a well-designed, highly secure and accurate personal identification system has always been a central goal in security business. This objective has posed a significant challenge that can be responded to by the use of multimodal biometric systems [1], [2], [3] which offers enhanced security and performance.

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]. Although the existence of these databases should enable one to develop and benchmark multimodal as well as multi-expert (utilizing the same biometric data but different matching software) fusion algorithms, they are 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 baseline systems. This is because an observed improvement due to a particular fusion algorithm may be due to the superior performance of its baseline systems rather than the merits of the fusion process. This shows the importance of having *common* baseline systems when benchmarking score-level fusion algorithms.

N. P. and J. K. are with CVSSP, FEPS, University of Surrey, Guildford, Surrey, GU2 7XH, UK. T. B. is with the Biometric Center, West Virginia University, Morgantown, WV 26506-6109, USA. E-mail: normanpoh@ieee.org, t.bourlai@surrey.ac.uk, j.kittler@surrey.ac.uk

B. Score-level Fusion and Signal Quality

The fusion paradigm we are interested in is where multibiometrics is treated as a two-stage problem where in the first stage we train the baseline systems. The scores produced by the baseline systems are then used as input 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 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.

An obvious disadvantage of score-level fusion is that, by using only scores, a lot of precious *non-class discriminatory* information is lost, for instance, 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. This information is non-class discriminatory, because it is not useful for distinguishing different individuals.

A recent trend in multimodal biometric fusion is to directly measure the signal quality using a set of criteria known to influence the system performance. These criteria are known as quality measures. Quality measures capture changes in signals that could directly impact on the system performance and can be used to weigh the participating classifiers in fusion appropriately. However, tapping the quality information – which is non-class discriminatory in nature – in order to improve the classification performance is not a trivial problem.

C. Three Types of Fusion Problems

In this study, we propose to evaluate fusion algorithms in three scenarios, namely, (i) quality-dependent fusion, (ii) client-specific fusion and (iii) cost-sensitive fusion. A quality-dependent fusion algorithm is one which attempts to devise a fusion strategy that is dependent on the biometric sample quality. The premise is that one should put more emphasis on the base classifier whose signal is of a better quality or deemed more reliable when computing the final score. A client-specific fusion algorithm further exploits the specific score characteristics of each enrolled user in order to customize the fusion strategy. Finally, a cost-sensitive fusion algorithm attempts to select a subset of biometric modalities/systems (at a specified cost) in order to obtain the maximal generalization performance. 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 often entails longer processing time).

As a variation of the problem of cost-sensitive fusion, one can also evaluate the capability of a fusion algorithm in dealing with missing modalities. This happens when one or more baseline systems are not operational due to failure to acquire or failure to match a biometric sample. Effectively, in this evaluation, one tests the robustness of a fusion algorithm with respect to missing scores (due to missing modalities).

To achieve our objective of evaluating the three mentioned fusion scenarios, we constructed a database with scores as well as quality measures obtained from the BioSecure DS2 database [10]. The following set of biometrics is used for the purpose of benchmarking: fingerprint, still face images and iris.

D. Contributions

This paper proposes a benchmark database for investigating a multi-expert biometric system in the following scenarios:

- To benchmark the performance of *quality-based* fusion algorithms, under changing conditions. The particular change of conditions considered here arises in cross-device matching (where template and query samples are acquired using two different devices).
- To benchmark *client-specific* fusion algorithms (where a fusion strategy is tailored to each enrolled user)
- To evaluate *cost-sensitive* evaluation of multimodal biometric system

To date, there simply exists no test bed for the above purposes. Without a benchmark data set, it is virtually impossible to measure real progress in multimodal biometric fusion. The availability of such database is one step towards achieving this objective. To our best knowledge, this is also the *first* multimodal biometric score-and-quality database released in the public domain. A similar work in this direction is [11] but quality measures are not available. The database described in this paper can be downloaded from: "http://face.ee.surrey.ac.uk/qfusion".

The paper is organized as follows: Section II describes the score-and-quality database. Section III provides experimental protocols for the evaluation. Section IV provides an initial experimental analysis. The baseline fusion experiments are reported in Section V. This is followed by conclusions in Section VI.

II. The BioSecure DS2 Data Set and Reference Systems

A. The Biosecure Database

The Biosecure database was collected with the aim 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 teleworking and Web or remote-banking services. As far 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. 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). Two genders are equally represented among the volunteers, whose ages have the following distribution: 2/3 in the range 18–40 of age and 1/3 above 40.

Table I presents the 17 streams of data available. A *stream* 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 as one stream of data. Using the notation presented in Table I, this stream of data is referred to as "fo5". The 17 streams of data are: fa1, fnf1, fwf1, ir1, ir2, fo1, fo2, fo3, fo4, fo5, fo6, ft1, ft2, ft3, ft4, ft5 and ft6.

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

While there are 17 streams, 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

TABLE I

THE DATA STREAMS 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 means the data stream of the right index fingerprint. 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 fingerprint as one sweeps a finger over it. The optical sensor acquires a fingerprint impression by direct contact (no movement required). This table results in 17 streams 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 data streams are not used for evaluation. For each data stream, two sessions of data acquisition were conducted. In each session, two biometric samples were collected.

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 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 Affinity SDK)
Fingerprint	NIST Fingerprint system	texture richness [12] (based on local gradient)
Iris	A variant of Libor Masek's iris system	texture richness [13], difference between iris and pupil diameters and proportion of iris used for matching

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.

There is only a fingerprint quality measure and it is based on the implementation found in [12]. 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, such as the NIST fingerprint system used in our experiments, is likely to under perform with this type of image.

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 [13]. 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. 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 [14].

III. THE EVALUATION PROTOCOLS

The current release of the desktop scenario contains data acquired from 333 persons. For each person, four samples per data stream 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 207 users available in the database. The remaining 126 subjects $(333 - 207)$ 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. In relationship to the template of a reference subject, all the 4 samples of the remaining half of the 207 subjects are considered impostors in the development set in Session 1. The other half of 207 subjects are used as impostors in Session 2. This ensures that the impostors used in Sessions 1 and 2 are

TABLE III
THE EXPERIMENTAL PROTOCOL FOR 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.

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".

Although the desktop scenario involves supervised data acquisition, the level of supervision differs from one collection site to another. As a result, there may be site-dependent bias in terms of performance.

In the following sub-sections, we shall explain 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 study, we adopt the following rules of thumb:

- If a device is used at least once, a fusion algorithm will be charged a unit cost, although we are aware that in reality, different devices may have different cost. This choice is clearly device and task dependent.
- The subsequent use of the same device will be charged 0.3 of a unit in view of 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 the recognition performance *while* minimizing the cost associated to the device usage. In this respect, there exists two strategies to solve this problem, which can be termed as a *fixed parallel* and a *sequential* approach. A fixed parallel solution pre-selects a set of channels and use them for all access requests. A sequential solution, on the other hand, may use different channels for different access requests. The sequence of systems used is determined dynamically.

For the cost-sensitive evaluation, the following streams of data are used: $\{$fa1, ft1–6, ir1$\}$, hence a total of eight expert outputs are considered. These subset of streams are chosen so that the fusion of all the systems will not give empirically observed zero error rate; otherwise, fusion algorithms cannot be compared. The total combination of eight expert outputs in this case is $2^8 - 1 = 255$. This is the result of choosing one out of 8 expert outputs to combine, two out of 8, etc, up to 8 out of 8, leading to 255 combinations. Note that it is not possible to choose none out of the 8 experts (hence explaining the minus one in $2^8 - 1$).

B. Cross-device Quality-dependent Evaluation

The goal of this evaluation experiment is to assess the ability of a fusion algorithm to select more reliable channels 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 acquired 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 scores and quality measures corresponding to all 8 channels were distributed to the participants. The (sequestered) evaluation set, on the other hand, contains only four channels of data as a result of mixing fnf/xfa (face taken with a digital camera/webcam) and fo$\{n\}$/xft$\{n\}$ for all $n \in \{1, 2, 3\}$ (optical/thermal fingerprint sensor for three fingers; see description in Table I). These four channels of data can be any of the following combinations:

(a) [fnf, fo1, fo2, fo3] – no device mismatch
(b) [fnf, xft1, xft2, xft3] – device mismatch for the fingerprint sensor
(c) [xfa, fo1, fo2, fo3] – device mismatch for the face sensor
(d) [xfa, xft1, xft2, xft3] – device mismatch for both the face and fingerprint sensors

In our experiment setting, the identity of the acquisition device is assumed to be unknown. This is a realistic scenario because as a biometric technology is deployed, it may be replaced by a newer device. Furthermore, its configuration may change, resulting in its acquired query biometric data being significantly different from the previously stored template data. This fusion problem is challenging because each of the four combinations shown above require a different fusion strategy in order to achieve the optimal result.

C. Simulation of Failure-to-acquire and Failure-to-match Scenarios

For each of the above mentioned two evaluation schemes, we also introduce a variation of the problem in order to simulate failure-to-acquire and failure-to-match scenarios. The motivation is to evaluate the robustness of a multimodal biometric system with respect to both types of failures. In principal, a multimodal system contains redundant subsystems, each of which produces a hypothesis regarding the authenticity of an identity claim. However, to our knowledge, such redundancy has never been formally evaluated.

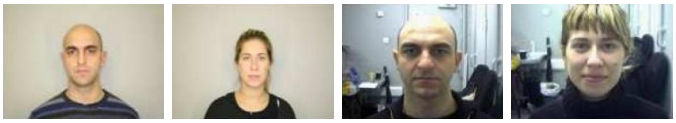
In order to simulate the failures, one can assume that they are device- and subject-dependent; device- and subject-independent; device-dependent but subject-independent; and, device-independent but subject-dependent. Among these four cases, we opted for the one that are both device- and subject-independent, i.e., the failures can happen randomly and spontaneously. This is actually a more difficult scenario among the four, as the failures are completely unpredictable. If they were, one could devise the following solutions: replace a particular device that is malfunctioning in the device-dependent case, or recommend a user to use a different biometric modality in the subject-dependent case. If a fusion algorithm can withstand our chosen scenario, the remaining three scenarios can therefore be solved easily. Based on this rationale, we shall focus on the device- and subject-independent case.

We shall introduce missing values only on the evaluation data set, and *not* the development data set. The reason is that the development data set is often better controlled. The missing values are introduced for each of the genuine or impostor match scores *separately* as follows: Let M be a matrix of scores of N samples by d dimensions (corresponding to all the d columns of match scores from d devices: face, 6 fingers and 1 iris). The total number of elements in M is $d \times N$. Missing values were gradually introduced by replacing T observed values with "-999" (the dummy value denoting missing value) in such a way that all the elements in the matrix M have equal probability of being deleted. We varied T such that the ratio of $T/(dN)$ was 10%, 20%, 30% and 40% and that the subsequent subset always contained missing values of its precedent subset.

For this evaluation, the same eight expert outputs as those defined for the cost-sensitive evaluation are used.

IV. Preliminary Analysis

We divided the preliminary analysis into two groups: subjective and objective analyzes. The subjective analysis involves direct examination of the raw biometric images whereas the objective one involves computation or visualization of scores and quality measures.



(a) Canon (b) Canon (c) Web cam (d) Web cam

Fig. 2. Images acquired using a Canon digital camera (a and b) and that acquired using a web cam (c and d).

(a) good quality (b) small iris area (c) displaced contact lens (d) blurred

Fig. 3. (a) Iris of good quality versus the degraded ones, e.g., (b) small iris area , (c) displaced contact lens and (d) blurred iris images.

A. Subjective Analysis

- **Cross-site diversity**: Because the data is collected on several sites, and each site may have one or more supervisors (to guide users during data acquisition), it is reasonable to expect some cross-site diversities. We show the existence of this site-dependent diversity in Figure 1.
- **Inter-device signal quality**: By using different devices, it is reasonable to expect different levels of signal quality. We compared face images acquired using a Canon digital camera with that acquired using a web cam. As can be observed, the images acquired by two different devices have different statistical properties which are dependent not just on the device but also the acquisition environment. The web cam images are taken in a relative uncontrolled environment and thus will have highly non-uniform background. Its images are also not as sharp as those taken by a digital camera.
- **Intra-device signal quality**: By examining several images taken by a single device, it is possible also to observe variations in signal quality that may potentially affect the resulting matching performance. We did so for the iris images and they are shown in Figure 3. The various degrading factors that can be observed here are small iris area, displaced contact lens and blurred iris images (due to miss-adjusted focus or movement). These degrading factors occur quite naturally even in a controlled environment. This highlights the challenging task of dealing with varying intra-device signal quality.

B. Performance on Session Mismatch and Device Mismatch

In this section, we test for a 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

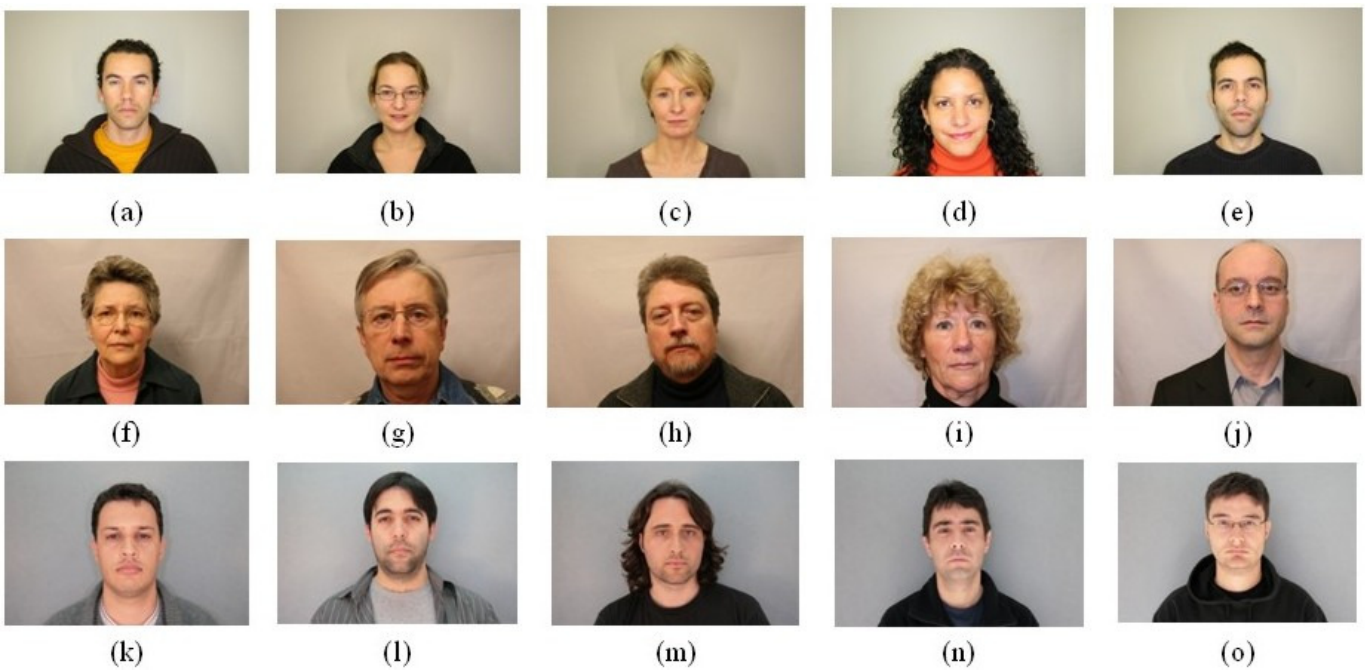


Fig. 1. Cross-site data diversity. Each of the three rows represents the data collected from three different sites. Even though the acquisition system setup was the same the quality of the data collected varies.

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* performance. We did so for all the 24 streams of data (see Figure 4). The first 17 streams of data involve matching using the same device. The remaining 7 streams 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 data streams, "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.

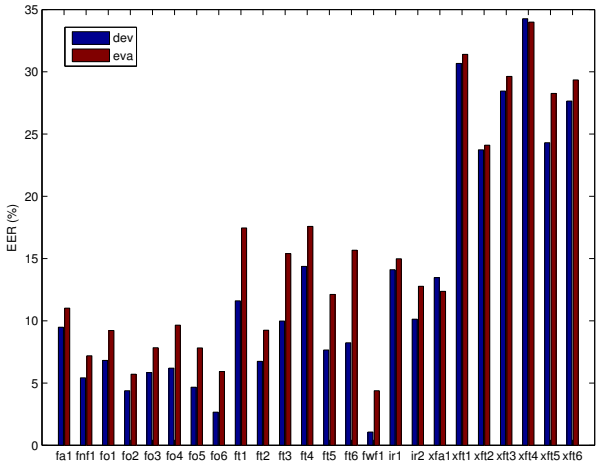


Fig. 4. Performance of Session 1 (blue; left bar) vs. Session 2 (red; right bar) in terms of EER for the 24 streams of data. The first 17 streams of data involve matching using the same device. The remaining 7 streams involve matching templates and query images acquired by different devices. The performance of Session 1 is labeled as "dev" whereas that of Session 2 as "eva".

The experiments here also allow us to compare the *cross-device matching* scenario, 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 under cross-device matching is always worse than that with the common device, even if the reference models/templates used may be of a higher quality.

C. 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 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 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 experiment, we used the quality measures of the development set and measured the performance of $P(d|q)$ on an evaluation set. 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 5(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 are taken in a more cluttered background whereas the digital camera images are taken in conditions conforming to passport standard with plain background. As a result, the images taken with a digital camera have lower 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 5(b). A classifier trained on this device classification problem (with the degree of uniform background and bits per pixel as features) gives an EER (assuming equal prior) of 0.5%. This result is shown in the first bar of Figure 5(c). The remaining four bars are EER of the fingerprint images. Among them, the first three are the performance of $p(d|q)$ where the devices can be either thermal or optical and the quality measure q is texture richness [12]. 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 5(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 5(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. Baseline Fusion Results

A. Cost-sensitive Fusion Evaluation

Let $y_i \in \mathbb{R}$ be the output of the i-th biometric subsystem and let there be N biometric subsystem outputs, i.e., $i \in \{1, \ldots, N\}$. For simplicity, we denote $\mathbf{y} = [y_i, \ldots, y_N]'$, where the symbol"$'$" is the matrix transpose operator. The most commonly used fusion classifier in the literature takes the following form:

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

where $y_{com} \in \mathbb{R}$ is a combined score. We shall refer to this classifier as *score-level* classifier.

The function f can be a *generative* or a *discriminative* classifier. In the former case, class-dependent densities are first estimated and decisions are taken using the Bayes rule or the Dempster-Shafer theory. In the latter, the decision boundary is directly estimated. A common characteristic of both types of classifiers is that the dependency among observations (scores or quality measures) is considered.

There exists also another approach that we will refer to as the *transformation-based approach* [15] which constructs a fusion classifier in two stages. In the first stage, the match scores of each biometric subsystem are independently transformed into a comparable range, e.g., in the range $[0, 1]$. In the second stage, the resulting normalized scores of all biometric subsystems are combined using a fixed rule such as sum or product [16].

As a baseline score-level fusion system, we used a GMM-based Bayesian classifier, i.e., a generative classifier constructed via the Bayes rules, using the Gaussian Mixture Model (GMM) as a density estimator. The output of this fusion classifier can be written as:

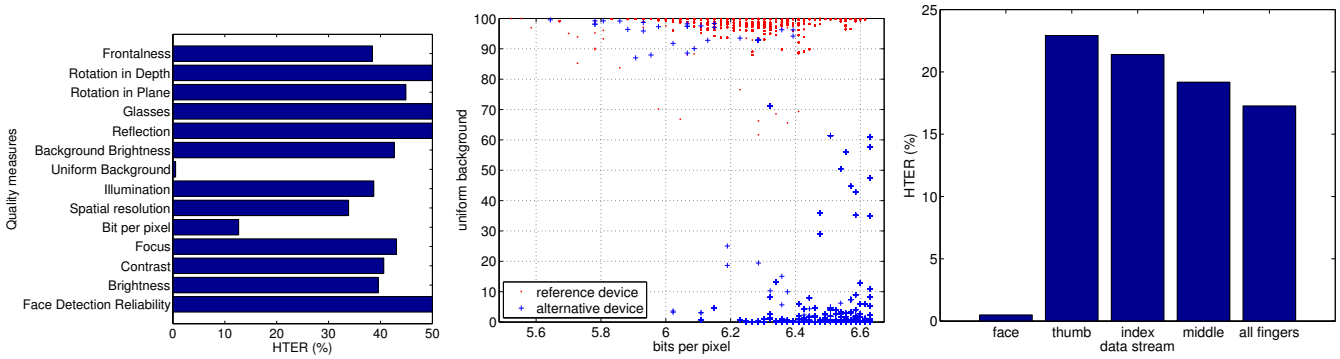
$$y_{com} = \log \frac{p(\mathbf{y}|\mathtt{C})}{p(\mathbf{y}|\mathtt{I})}. \tag{2}$$

where $p(\mathbf{y}|k)$, for both classes $k \in \{\mathtt{C}, \mathtt{I}\}$ (client or impostor), is estimated using a GMM. Its associated decision threshold is optimal when

$$\Delta = -\log \frac{P(\mathtt{C})}{P(\mathtt{I})}.$$

The GMM-based Bayesian fusion classifier was trained on the entire score feature space (a total of 8 dimensions). It was then tested on all the $2^8 - 1 = 255$ combinations of the score feature space (as described in Section III-A) by means of Gaussian marginalization [17], [18]. Missing values were handled in the same way. For instance, if features 1, 2, and 4 are chosen, and 4 is missing, then the GMM-bayes fusion classifier will calculate the final combined score using only features 1 and 2; the remaining features, i.e., $\{3, 4 - -8\}$, are thus marginalized or integrated out.

In order to estimate the fusion performance using only the development set (recalling that the evaluation scores were sequestered), we employed a two-fold cross-validation. The resultant performance, measured in terms of averaged EER of the two folds, across all 255 combinations, is shown in Figure 6(a). Plotted in this figure are the median (red), the upper and lower quantiles (cyan and green lines resp.), and,

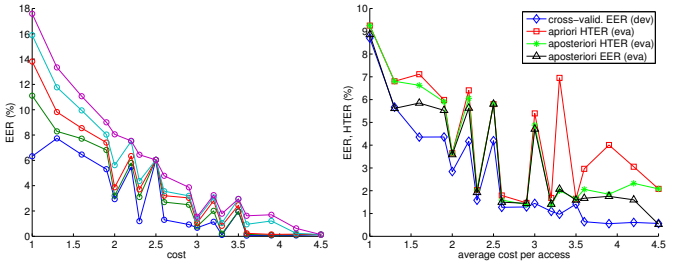


(a) face device classification given individual quality measures

(b) Scatter plot of bits per pixel vs. uniform background

(c) device classification for all streams of data

Fig. 5. 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.



(a) Optimization with GMM-bayes classifier

(b) Rank-one cost-performance curve

Fig. 6. (a) Optimization of UNIS GMM-bayes fusion classifier by a two-fold cross validation on the development set. (b) Rank-one performance vs average access cost. This GMM-bayes system was provided by the organizer.

the upper and lower range (in purple and blue lines) of performance in HTER for a given cost. Note that there is only one possible way to obtain a cost of 2.5, i.e., by combining all 6 fingers, hence explaining the convergence of performance to a single point. Among these curve, the lowest one (in blue) is the most important one because the goal here is find a fusion candidate that has the lowest error at a given cost.

The performance versus cost curves presented in Figure 6(b) are called "rank-one" cost-performance curve. This means that only the performance of the best fusion candidate (using the GMM-Bayes classifier) is reported. In a rank-two curve, one would choose the minimum of the top two performing candidates to plot the curve, etc. Three of the four curves were computed on the evaluation set and only one on the development set. The latter is plotted here (in blue) in order to show the actual performance optimized on the development set via the two-fold cross validation. The reported error is the average EER of the two folds. The EER measure (rather than HTER) is more suitable in this context so that the performance is independent of the choice of the decision threshold. The remaining three curves are explained below:

1) *a priori* **HTER**: This rank-one curve (plotted in red) shows the *achievable* generalization performance if one were to use the fusion system candidates minimizing a given cost, based on the development set, via cross-validation.
2) *a posteriori* **HTER**: This rank-one curve (plotted in green) shows the actual performance in terms of HTER of the fusion system candidate on the evaluation set. The assumption here is that the evaluation set is available but the optimal decision threshold is unknown.
3) *a posteriori* **EER**: Finally, this rank-one curve (plotted in black) is similar to the previous one, reporting the performance of the fusion system candidates optimizing a given cost on the evaluation set, except that it also assumes that the optimal threshold is known. This curve is hence reported in EER.

When optimizing a fusion classifier without any knowledge of the evaluation set (in the sequestered scenario), the best performance one can obtain is the first (*a priori*) curve.

The second and third (rank-one) curves are not achievable; they are shown here in order to show the oracle cases, where the evaluation set is available for the second curve; and on top of that, the optimal decision threshold is known for the third curve. As can be observed, by injecting more information, the error actually decreases from the first to the second curve; and, from the second to the third curve.

These curves show that the actual achievable fusion performance is dependent on two factors: the fusion candidate and the (correct) decision threshold. Choosing the correct candidate given only the development set requires a criterion yielding a solution that can generalize well across populations. In [19], the authors demonstrated that such a criterion can be effectively realized using parametric error bounds such as the Chernoff and Bhattacharyya bounds [20], rather than computing the EER of the fusion performance empirically, as commonly practised. Error bounds, however, do assume that the underlying scores are normally distributed and therefore, pre-processing is recommended to ensure the conformity of the data to this assumption. In practice, it was observed in [19] that even if the underlying multivariate distribution is not strictly Gaussian (as measured by the standard Gaussianity tests), the estimated bound is still better (in terms of rank-

one performance-cost curve) than the empirical estimates of error (via cross-validation on the development set) for fusion candidate selection.

B. Quality-based Fusion Evaluation

In quality-based fusion, one will have to consider a vector of quality measures in addition to the expert outputs. Let the signal quality of the i-th biometric subsystem be represented by a vector of $\mathtt{L}_i$ measurements, $\mathbf{q}_i \in \mathbb{R}^{\mathtt{L}_i}$. Note that different biometric subsystems may have different number of quality measures $\mathtt{L}_i$. For simplicity, we denote $\mathbf{q}$ as a concatenation of all $\mathbf{q}_i$'s, i.e., $\mathbf{q} = [\mathbf{q}_1', \ldots, \mathbf{q}_N']'$. The function f in this case takes the following form:

$$f : \mathbf{y}, \mathbf{q} \rightarrow y_{com} \tag{3}$$

Broadly speaking, there are two main categories of quality-based fusion: (i) feature-quality based and (ii) cluster-quality based fusion. In the former, quality measures are used directly as observation in a similar way as the expert outputs are. In the latter, quality measures are first clustered, and for each cluster, a fusion strategy is devised. The motivation for the cluster-based approach is that samples whose quality measures belong to the same category are subject to the same acquisition condition (e.g., a particular lighting condition for the face biometrics) as well as the human interaction (e.g., a particular head pose). Hence, it is reasonable to devise a fusion strategy for each cluster of quality measures [21].

To realize the two approaches, it is sensible to consider the fact that the noise affecting each modality is likely to be independent. For instance, the noise source affecting face does not affect the fingerprint modality, and vice versa. Therefore, in order to combine the quality information with the biometric system output, one should do so for each biometric modality. This implies that the quality-based fusion can be performed in two stages: first, perform quality-based normalization for each modality, and then combine the normalized output.

Using the likelihood ratio test, which is an optimal decision in Neyman-Perason sense [20], the feature-quality based approach computes a combined output score as follows:

$$y_i^{\mathtt{feature}} = \log \frac{p(y_i \mathbf{q}_i | \mathtt{C})}{p(y_i, \mathbf{q}_i | \mathtt{I})}, \tag{4}$$

whereas for the quality-cluster based approach, the output is computed as:

$$y_i^{\mathtt{cluster}} = \log \frac{\sum_Q p(y_i | \mathtt{C}, Q) p(Q | \mathbf{q}_i)}{\sum_Q p(y_i | \mathtt{I}, Q) p(Q | \mathbf{q}_i)}, \tag{5}$$

where Q denotes a cluster of quality measures. The sum over all the quality states of Q is necessary since the quality state is a hidden variable; only y_i and $\mathbf{q}_i$ are observed.

In both the above cases, each biometric subsystem output is processed independently. The resultant quality-normalized score, y_i^m, for $m \in \{\mathtt{feature}, \mathtt{cluster}\}$, is then combined using the sum rule:

$$y_{com} = \sum_i y_i^m$$

As a control fusion experiment, we also use a conventional fusion classifier that does not make use of any quality information, i.e., (2).

Note that in (5), the dimensionality involved in estimating $p(y_i|k, Q)$ is effectively one since y_i is one-dimensional and Q is a discrete variable. The partitioning function $P(Q|\mathbf{q}_i) : \mathbb{R}^{\mathtt{L}_i} \rightarrow \mathbb{R}$ refers to the posterior probability that $\mathbf{q}_i$ belongs to the cluster Q. This term is also known as *responsibility* in the GMM literature [22]; it is obtained via the Bayes rule:

$$P(Q|\mathbf{q}_i) = \frac{p(\mathbf{q}_i)|Q)P(Q)}{\sum_{Q'} p(\mathbf{q}_i)|Q')P(Q')}$$

In comparson, the feature-quality based approach, i.e., (4), involves the estimation of $p(y_i, \mathbf{q}_i|k)$ which has $\mathtt{L}_i + 1$ dimensions. This increased dimensionality may present a potential parameter estimation problem.

As an advanced quality-cluster based system, we shall report a variant of (5) by introducing the device information d:

$$y_i^{llr} = \log \frac{\sum_Q p(y_i | \mathtt{C}, Q, d) p(Q | \mathbf{q}_i, d)}{\sum_Q p(y_i | \mathtt{I}, Q, d) p(Q | \mathbf{q}_i, d)}, \tag{6}$$

This version is particularly suited for the problem at hand. For the quality-based fusion that we deal with, the variation in signal quality is mainly due to the fact that two devices are involved in collecting the query biometric samples. When the device used to acquire a query sample is different from that used to build the reference model (template), the comparison is called *cross-device* comparison. As a result, significant degradation is observed [23]. By using (6), one effectively normalizes against variation in quality induced by the biometric device, as well as latent factors (e.g., user interaction) causing the variability in performance even if a single device is used during testing.

(6) can also be used when the device information is unknown by treating d as unobserved, i.e., by marginalizing it during inference. Effectively, this requires the ability of quality measures to distinguish between the two devices. This is indeed the possible, as evidenced in Section IV-C.

The result of quality-based fusion using the feature-quality based approach, i.e., (4), and the cluster-device quality based approach, i.e., (6), as well as the baseline fusion (without quality measures), i.e., (4) are shown in Figure 7.

As can be observed, all fusion strategies outperform any of the unimodal systems. However, the cluster-device approach performs the best. The feature-quality based approach did not generalize as well as the conventional fusion (without quality measures) due to inaccurate estimation of the GMM distribution parameters. More explanation of this can be found in [23].

C. Client-specific Fusion Evaluation

Let $j \in \{1, \ldots, J\}$ be a claimed identity and there are J enrollees in the database. A client-specific fusion strategy takes a claimed identity, j, as well as the biometric subsystem outputs $\mathbf{y}$ in order to produce a final combined output.

$$f : \mathbf{y}, j \rightarrow y_{com} \tag{7}$$

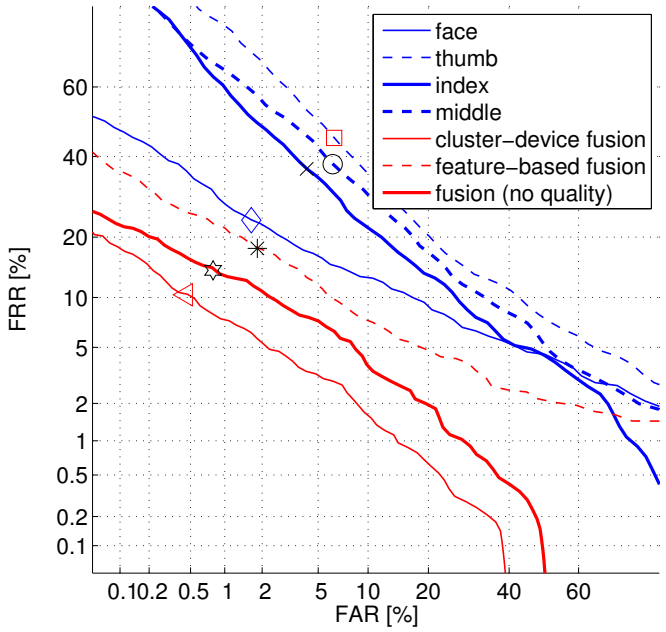


Fig. 7. A comparison of the unimodal systems and the following fusion systems: cluster-device based fusion, feature-quality fusion and fusion without any quality information.

By exploiting the specific characteristic of each enrollee, the resultant fusion classifier differs slightly in parameters from one user to another. This fusion problem is not entirely new, and has been examined in [24], [25], [26], [27], [28], [29], [30]. The main difficulty in designing a client-specific classifier is the scarcity of the genuine user scores *for each enrollee*. In our data set, there is only a genuine score (obtained from the second sample of session one). A second difficulty is that this sample was collected in session one, i.e., the same session as the sample which was used to build a reference model/template (recalling that two samples are collected in each session). As a results, due to the lack of cross-session variability, the resultant genuine scores are *biased*, i.e, resulting in higher performance than it should be under different sessions. The performance with and without cross-session variability has already been shown in Figure 4.

As a baseline client-specific fusion approach, we shall use the method proposed in [30]. The idea is to apply a client-specific score normalization procedure to each biometric subsystem output independently and then only designing a fusion classifier to combine the normalized score outputs. This two steps can be formally described in two steps, i.e., the following client-specific score normalization

$$\Psi_j : y_i \rightarrow y_{i,j}^{norm}, \text{ for } i \in \{1, \ldots, N\}.$$

followed by the following fusion:

$$f : y_{1,j}^{norm}, \ldots, y_{N,j}^{norm} \rightarrow y_{com}$$

According to [30], there are three families of client-specific score normalization procedures producing very distinctive effects, namely, the Z-, F- and EER-norm. Among them, the F-norm was empirically shown to be the most effective in exploiting the genuine score information. The reason is that it relies only on the first order of moment (and not the second order), hence providing some robustness to the paucity of the genuine scores. The F-norm is defined as:

$$\Psi_j(y) = \frac{y - \mu_j^{\mathtt{I}}}{\gamma\mu_j^{\mathtt{C}} + (1-\gamma)\mu^{\mathtt{C}} - \mu_j^{\mathtt{I}}}$$

where $\mu^{\mathtt{C}}$ is the client-independent genuine mean score, $\mu_j^{\mathtt{I}}$ is the client-specific impostor mean score, and $\mu_j^{\mathtt{C}}$ is the client-specific genuine mean score. The parameter $\gamma \in [0, 1]$ weighs the contribution between the client-specific genuine mean score and the client-independent one. The F-norm has the following effect: the resultant impostor match scores have zero mean whereas the genuine match score has an expected value of one.

It should be noted that the among the parameters of the F-norm, $\mu_j^{\mathtt{C}}$ is the one which cannot be estimated reliably. In our experiment setup, in essence, the parameter is estimated from a *single* score. Moreover, the sample from which this score was obtained was collected from the same session as the enrollment sample. Hence, the resultant score is *biased* as it does not contain inter-session variability (hence leading to optimistically better performance on the dev set compared to the eva set, as depicted in Figure 4). As a means to guard against the small sample size (leading to the estimation problem of $\mu_j^{\mathtt{C}}$) as well as the systematic biased of the estimated parameter, in [30], it was recommended that $\gamma = 0.5$ is used. Following this recommendation, we applied the F-norm to each of the 8 expert outputs in the cost-based evaluation experimental protocol, and then combined the resultant normalized scores using logistic regression [31]. For the fusion experiments, we exhaustively combined two and three out of the possible eight expert outputs. Figure 8 compares the fusion results of the client-specific fusion classifier as described here with a client-independent one realized using logistic regression (without applying any client-specific score normalization procedure before fusion). Also shown in the figure is the performance of each of the eight systems before and after normalization (shown with the legend "1"). As can be observed, consistent with the literature, the client-specific fusion classifier, in the majority of the cases, outperforms the client-independent fusion classifier.

VI. Conclusions

While score-level fusion has always been treated as a static problem, in particular, of the form $\sum_i w_i y_i$, where w_i is the weight associated with the i-th system output y_i, by using quality measures, one can realize a fusion rule of the form $\sum_i w_i(q) y_i$, where $w_i(q)$ is dependent on the signal quality, as characterized by the vector of quality measures, q. This is an example of a quality-dependent fusion algorithm.

Despite the importance of research in quality-dependent fusion, to the best of our knowledge, there existed no publicly available database to benchmark the algorithms. The BioSecure DS2 database (with the desktop scenario) is the first benchmark database designed for this purpose. We summarize here the multibiometric system fusion scenarios that can be investigated by the proposed database:

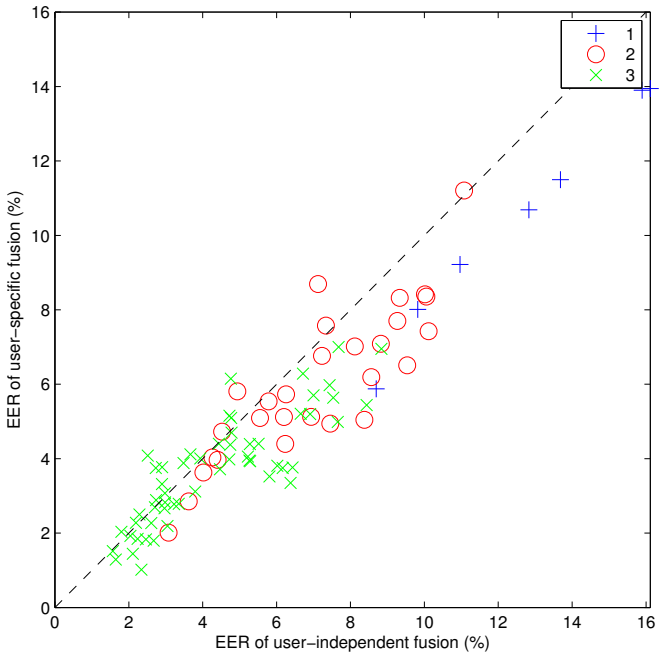


Fig. 8. Client-specific versus client-independent fusion. Shown in the legends are the number of systems to combine. The legend with "1" reports the system performance before and after normalization (no fusion was involved).

- **Quality-dependent evaluation**: This evaluation scheme allows matching with query images obtained from several sensors which may be different from the one used to build a biometric template/model. A matching algorithm often under-performs in the presence of a device mismatch between a template and a query image. In this situation, quality measures are made available so that the designer of a fusion algorithm can develop a fusion algorithm that adapts itself according to the quality of the raw biometric signal as captured by the quality measures. For the face biometrics, as many as 14 quality measures are made available for this purpose.
- **Cost-sensitive evaluation**: The evaluation is posed as an optimization problem where the objective is to minimize a cost-sensitive criterion while maximizing the overall system performance. Cost in this sense refers to 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.
- **Missing observation**: Scores and quality measures may not be available. Observations are missing because a biometric system fails to process or match a query sample with a template.
- **User-specific/person-dependent strategy**: The score/quality data set is designed to test fusion algorithms that can adapt themselves according to the claimed identity label.

At present, this database contains 333 subjects and is still growing (it is expected to contain in excess of 500 users).

Our analysis based on this database reveals the followings:

- **Biased intra-session performance**: The performance measured on data with intra-session data (where template and query images are taken in a single session or visit) is likely to be optimistically biased as opposed to performance on inter-session data (collected in two or more sessions or visits).
- **Degraded performance with device mismatch**: When the template and query images are taken with different devices, in a scenario referred to as a device mismatch, the resulting performance will be worse than what would be obtained when matching with the same device.
- **The discriminatory power of quality measures to distinguish acquisition devices**: The automatically derived quality measures from the raw biometric data can be used to suggest the identity of the acquisition device.

Our on-going work extends the possibility of using the inferred device identity to realize a device-specific score normalization procedure as well as using such information at the fusion level.

ACKNOWLEDGMENTS

This work was supported partially by the advanced researcher fellowship PA0022_121477 of the Swiss National Science Foundation, by the EU-funded Mobio project (www.mobioproject.org) grant IST-214324 and by the BioSecure project (www.biosecure.info).

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," M.S. 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. Garcia-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] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] A. Jain, K. Nandakumar, and A. Ross, "Score Normalisation in Multimodal Biometric Systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[16] 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.

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

[18] A. C. Morris, M. P. Cooke, and P. D. Green, "Some Solutions to the Missing Features Problem in Data Classification with Application to Noise Robust Automatic Speech Recognition," in *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, 1998, pp. 737–740.

[19] N. Poh and J. Kittler, "On Using Error Bounds to Optimize Cost-sensitive Multimodal Biometric Authentication," in *Proc. 19th Int'l Conf. Pattern Recognition (ICPR)*, 2008.

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

[21] 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.

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

[23] N. Poh, T. Bourlai, and J. Kittler, "Quality-based score normalisation with device qualitative information for multimodal biometric fusion," *IEEE Trans. on Systems, Man, and Cybernatics (part B)*, 2009, accepted for publication.

[24] A. Jain and A. Ross, "Learning User-Specific Parameters in Multibiometric System," in *Proc. Int'l Conf. of Image Processing (ICIP 2002)*, New York, 2002, pp. 57–70.

[25] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 450–455, 2005.

[26] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Exploiting General Knowledge in User-Dependent Fusion Strategies For Multimodal Biometric Verification," in *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, 2004, vol. 5, pp. 617–620.

[27] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Bayesian Adaptation for User-Dependent Multimodal Biometric Authentication," *Pattern Recognition*, vol. 38, pp. 1317–1319, 2005.

[28] A. Kumar and D. Zhang, "Integrating Palmprint with Face for User Authentication," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 107–112.

[29] K.-A. Toh, X. Jiang, and W.-Y. Yau, "Exploiting Global and Local Decision for Multimodal Biometrics Verification," *IEEE Trans. on Signal Processing*, vol. 52, no. 10, pp. 3059–3072, October 2004.

[30] N. Poh and J. Kittler, "Incorporating Variation of Model-specific Score Distribution in Speaker Verification Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 594–606, 2008.

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