BioSecure DS2: A Score-level Quality-dependent and Cost-sensitive Multimodal Biometric Test Bed

Norman Poh, Thirimachos Bourlai and Josef Kittler

Abstract

This paper presents a test bed for evaluating, comparing and benchmarking fusion algorithms for multimodal biometric authentication. It consists of a database designed to benchmark *quality-dependent* and *conventional fusion algorithms*. These two types of algorithms are different because only the former one uses quality measures to derive the fused score. These auxiliary measurements describe the quality of a biometric signal. To the best of our knowledge, the BioSecure DS2 evaluation campaign is the first attempt to benchmark these quality-dependent fusion algorithms. A second unique characteristic of the BioSecure DS2 evaluation campaign is its focus on cost-sensitive fusion of multiple biometrics. In this type of evaluation, one considers a fusion task as an optimization problem whose goal is to achieve the highest performance with a minimal cost of acquiring and processing biometric information. The BioSecure DS2 database contains 24 streams of match scores produced by using multiple biometrics and multiple samples of each biometric acquired using multiple devices. The database contains almost 1000 subjects. As a variant to benchmarking the quality-dependent and cost-based fusion schemes, the proposed experimental protocols allow one to benchmark person-dependent (or client-specific) fusion algorithm, hence, providing a test bed for cost-, person- and quality-dependent fusion algorithms.

Index Terms

multimodal biometric authentication, benchmark, database, fusion

ef

All the authors 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

I. Introduction

In order to improve confidence in verifying the identity of individuals seeking access to physical or virtual locations both government and commercial organisations 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 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 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 itself. This shows the importance of having *common* baseline systems 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 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 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 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[1]. 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 [11] 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

[1]This database is collected over 11 European sites; the BioSec database documented in [10] refers essentially to only 2 sites which, unfortunately, is not necessarily representative of the BioSecure database described here.

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

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). The volunteers are selected to have both sexes 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:

II. The BioSecure DS2 Data Set and Reference Systems

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’s 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 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 finterprint acquistion, 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 [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 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 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".

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.

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. 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 intra-device signal quality.

B. 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 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.

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.

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. 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 5(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 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. 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 some important characteristics of the proposed database:

- **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-based 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 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 algorithm that can adapt itself according to the claimed identity label.

At present, this database contains 333 subjects and is still growing (up to some 500 users are expected). 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 experiments 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 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. Moon1, 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. Fierrez, D. T. Toledanoa J. Ortega-Garciaa, and J. Gonzalez-Rodriguez, "Biosec baseline corpus: A multimodal biometric database," *Pattern Recognition*, vol. 40, no. 4, pp. 1389–1392, 2007.

[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.

TABLES

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'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 [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

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.

FIGURE CAPTIONS

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.

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

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.

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”.

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.

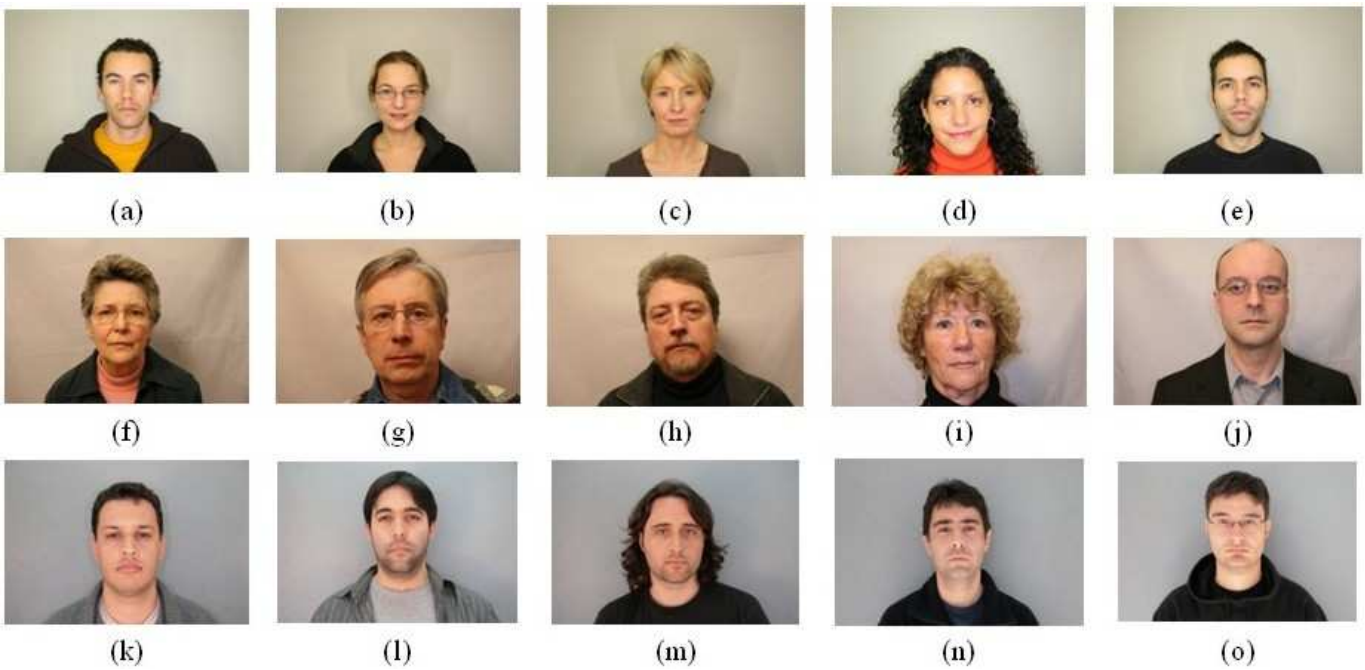
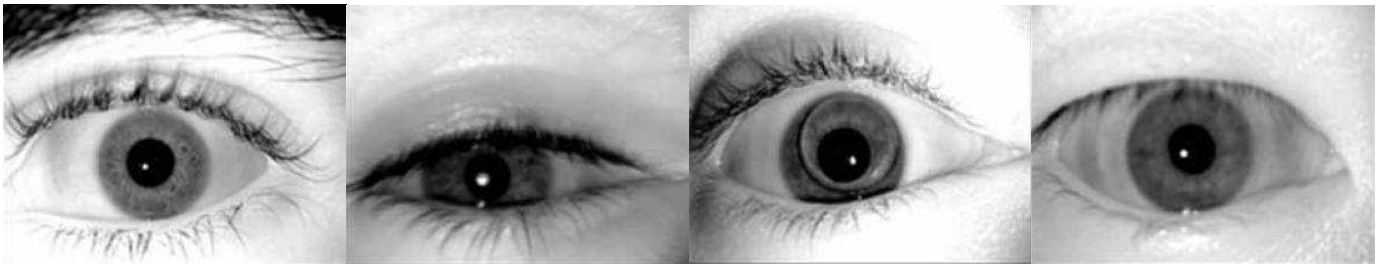


Fig. 1. Cross-site data diversity. Each row represents the data collected from three different sites. Even though the acquisition system setup was the same the quality of the data collected varies.

(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.

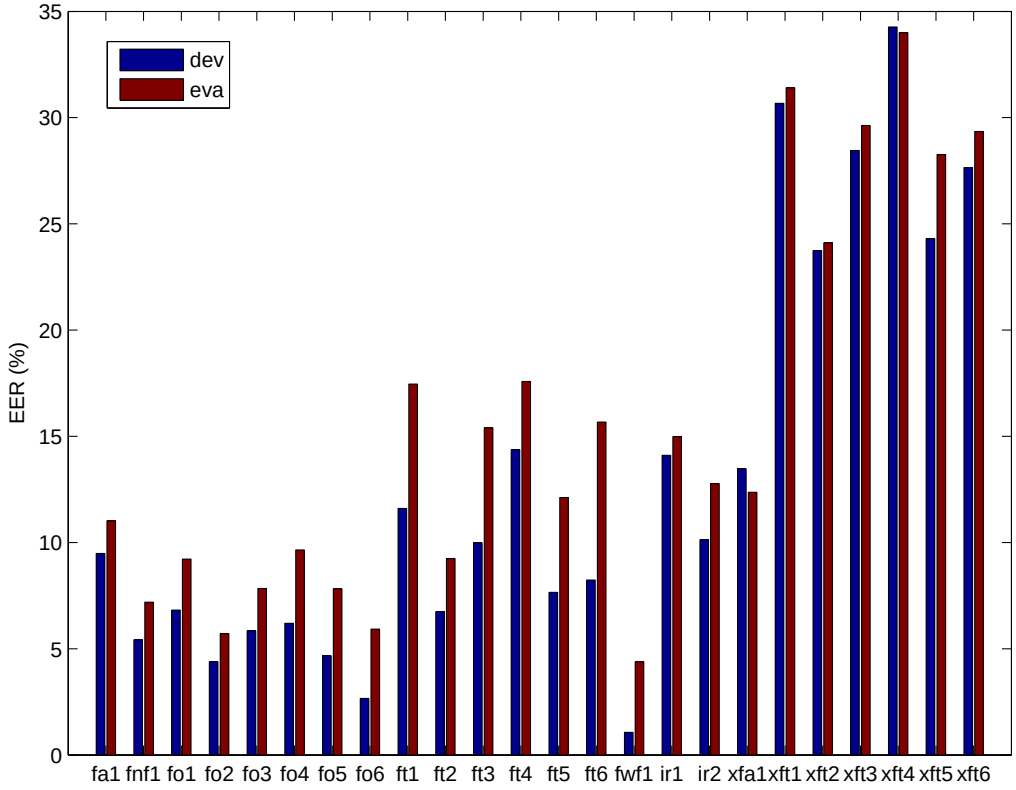
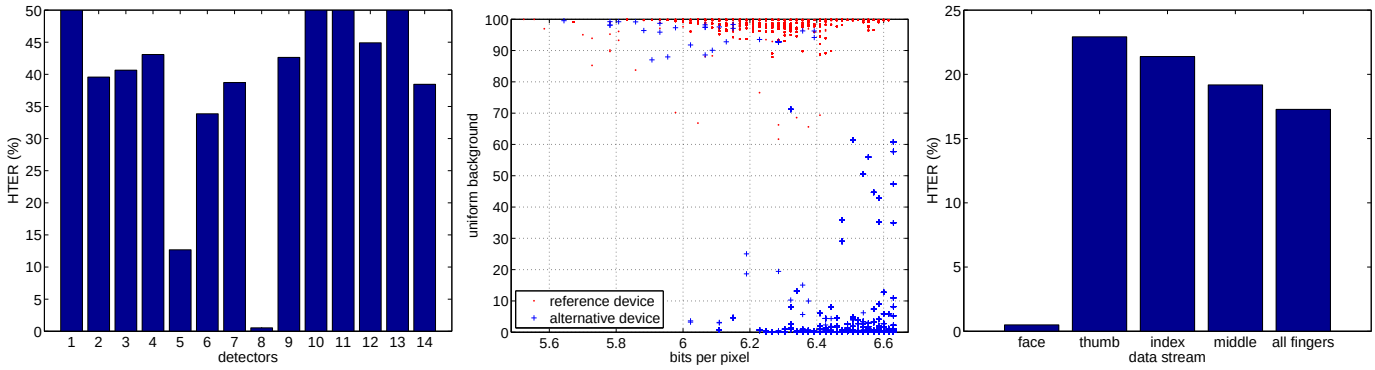


Fig. 4. Performance of Session 1 (blue; left bar) vs. Session 2 (red; right bar) in terms of EER for the 24 streams of data. The first 17 streams of data involve matching using the same device. The remaining 7 streams involve matching templates and query images acquired by different devices. The performance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



(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.