Dealing with sensor interoperability in multi-biometrics: The UPM experience at the Biosecure Multimodal Evaluation 2007

Fernando Alonso-Fernandez, Julian Fierrez, Daniel Ramos, Javier Ortega-Garcia

ATVS/Biometrics Research Lab., Escuela Politecnica Superior, Univ. Autonoma de Madrid, Avda. Francisco Tomas y Valiente 11, 28049 Madrid, SPAIN

ABSTRACT

Multimodal biometric systems allow to overcome some of the problems presented in unimodal systems, such as non-universality, lack of distinctiveness of the unimodal trait, noise in the acquired data, etc. Integration at the matching score level is the most common approach used due to the ease in combining the scores generated by different unimodal systems. Unfortunately, scores usually lie in application-dependent domains. In this work, we use linear logistic regression fusion, in which fused scores tend to be calibrated log-likelihood-ratios and thus, independent of the application. We use for our experiments the development set of scores of the *DS2 Evaluation* (*Access Control Scenario*) of the BioSecure Multimodal Evaluation Campaign, whose objective is to compare the performance of fusion algorithms when query biometric signals are originated from heterogeneous biometric devices. We compare a fusion scheme that uses linear logistic regression with a set of simple fusion rules. It is observed that the proposed fusion scheme outperforms all the simple fusion rules, with the additional advantage of the application-independent nature of the resulting fused scores.

Keywords: Biometrics, fusion, calibration, BioSecure, linear logistic regression, quality, multisensor.

1. INTRODUCTION

Biometric systems make use of the physiological and/or behavioral traits of individuals for recognition purposes.[1] However, using a single trait for recognition (i.e. unimodal biometric systems) is often affected by practical problems like noisy sensor data, non-universality, lack of distinctiveness of the biometric trait, spoof attacks, etc.[2] Multibiometric systems integrate the evidence presented by multiple biometric sources, being more robust and overcoming some of these problems. The use of multiple biometric indicators for identifying individuals has been shown to increase accuracy and population coverage, while decreasing vulnerability to spoofing.[3]

Integration at the matching score level is the most common approach used in multibiometric systems due to the ease in accessing and combining the scores generated by different matchers.[4,5] However, since the matching scores output by the various modalities are heterogeneous, score normalization is needed to transform these scores into a common domain prior to the fusion process.[6] The acceptance/rejection is based on a decision threshold and this threshold depends on the priors and decision costs involved in the decision taking process. The priors and costs have been together called "application".[7,8] A number of studies have evaluated the performance of different normalization techniques and fusion rules.[3,6,9] In this paper, we use linear logistic regression fusion,[7,10] a trained classification fusion approach which does not need any prior normalization stage. In addition, with this fusion, the scores are fused in such a way as to encourage good calibration of the output score. Calibration means that output scores are mapped to *log-likelihood-ratios*, thus being in an application-independent domain.[8] Calibration and logistic regression have been recently used in the field of speaker recognition,[7,8,10–12] but they are applicable to any other score-based biometric system.

We use for our experiments the *development* set of scores of the *DS2 Evaluation (Access Control Scenario)* of the BioSecure Multimodal Evaluation Campaign.[13,14] In this evaluation campaign, the recently acquired

Further author information: (Send correspondence to F. A.-F.)
F. A.-F.: E-mail: fernando.alonso@uam.es, Telephone: + 34 91 497 33 63.
Authors were formerly at UPM - Universidad Politecnica de Madrid.

Biosecure Multimodal Database has been used, which is also described in this paper. This set of scores contains face still samples collected with two cameras of different resolution and fingerprint samples collected both with an optical and a thermal sensor. The aim of this evaluation is to compare the performance of multi-modal fusion algorithms when query biometric signals are originated from different biometric devices. We propose a fusion algorithm that first estimates the device used in the access and then, applies a different linear logistic regression classifier adapted to each device. Since linear logistic regression classifiers tends to produce calibrated log-likelihood-ratios as output, accesses with different devices will produce output scores that are in a comparable domain. We compare the proposed fusion scheme to a set of simple fusion rules. We demonstrate in our experiments that the proposed fusion scheme outperforms all the simple fusion rules, with the advantage that output scores asymptotically lead to optimal decisions if Bayes thresholds are used, independently of the application (priors, costs).[15]

The rest of the paper is organized as follows. In Section 2, the concept of calibration is introduced and linear logistic regression fusion is described. Section 3 provides a brief description of the Biosecure Multimodal Database. Section 4 describes the evaluation framework, including the dataset used in our experiments, the proposed fusion scheme and the results. Conclusions are finally drawn in Section 5.

2. CALIBRATION AND FUSION OF BIOMETRIC SYSTEMS

2.1 Score calibration

A biometric system can be defined as a pattern recognition machine that, by comparing two (or more) samples of input signals such as speech, face images, etc., is designed to recognize two different classes. These two classes are known as *target* or *client* class, if both samples were originated by the same subject, and *non-target* or *impostor* class, if both samples were not originated by the same subject. As a result of the comparison, the biometric system output a real number known as *score*. The sense of this score is that higher scores favor the target hypothesis and lower scores favor the non-target hypothesis.

Usually, each biometric system outputs scores which are in a range that is specific of the system. For instance, a particular system can output scores in the $[0, 1]$ range, whereas another system can output scores in the $[-1, 1]$ range. Therefore, an score value of 0 has different meaning depending on the system. Even if two systems output scores in the same range, the same output value might does not favor the target or non-target hypotheses with the same strength. In this context, outputs are dependent of the system and thus, the acceptance/rejection decision also depends on the system.

In contrast, we are interested in output scores that do not depend of a particular system, i.e. the same score value favors the target or non-target hypotheses always with the same strength, independently of the biometric system that produced the score. This can be achieved by using a *calibration* stage, which map the score to a *log-likelihood-ratio*. This is the logarithm of the ratio between the likelihood that input signals were originated by the same subject, and the likelihood that input signals were not originated by the same subject. The act of designing or optimizing this calibration mapping is also known as calibration.[8] This form of output is *application-independent* in the sense that this log-likelihood-ratio output can theoretically be used to make optimal (Bayes) decisions for any given target prior and any costs associated with making erroneous decisions.[15]

2.2 Linear logistic regression fusion

The scores of multiple sub-systems are fused together, primarily to improve discriminative ability, but in such a way as to encourage good calibration of the output scores. Given N matchers which output the scores $(s_{1j}, s_{2j}, ...s_{Nj})$ for an input trial j, a linear fusion of these scores is:

$$f_j = a_0 + a_1 \cdot s_{1j} + a_2 \cdot s_{2j} + ... + a_N \cdot s_{Nj}$$

The constant a_0 does not contribute to the discriminative ability of the fusion, but it can improve the calibration of the fused score. This constant is included here, because when these weights are trained via logistic regression, the fused score tends to be a well-calibrated log-likelihood-ratio.[7,8]

INSTITUTION	COUNTRY	DS1	DS2	DS3
Joanneum Research Graz	Austria	X	X	-
Groupe des Ecoles des Telecommunications	France	X	X	*
University of Sassari	Italy	X	X	-
Pompeu Fabra University	Spain	X	-	X
Universidad de Vigo	Spain	*	X	-
Universidad Politecnica de Madrid	Spain	X	*	X
Ecole Polytechnique Federale de Lausanne	Switzerland	X	-	X
University of Fribourg	Switzerland	X	-	X
Bogazici University	Turkey	X	-	X
University of Kent	United Kingdom	X	X	X
University of Surrey	United Kingdom	X	X	X

Table 1. Institutions participating in the acquisition of the Biosecure Multimodal Database, including involvement in the three acquired datasets. For each dataset, there is an institution in charge of coordinating its acquisition (marked with *).

Let $[s_{ij}]$ be an $N \times N_T$ matrix of scores that each of the N component systems calculated for each of N_T target trials, and let $[r_{ij}]$ be an $N \times N_{NT}$ matrix of scores that each of the N component systems calculated for each of N_{NT} non-target trials. We use a logistic regression objective[7,10] that is normalized with respect to the proportion of target trials to non-target trials (N_T and N_{NT}, respectively), and weighted with respect to a given prior probability $P = P(\text{target})$. The objective is stated in terms of a *cost*, which must be *minimized*:

$$C_{wlr} = \frac{P}{N_T} \sum_{j=1}^{N_T} \log\left(1 + e^{-f_j - \text{logitP}}\right) + \frac{1-P}{N_{NT}} \sum_{j=1}^{N_{NT}} \log\left(1 + e^{-g_j - \text{logitP}}\right)$$

where the fused target and non-target scores are respectively:

$$f_j = \alpha_0 + \sum_{i=1}^{N} \alpha_i s_{ij},$$

$$g_j = \alpha_0 + \sum_{i=1}^{N} \alpha_i r_{ij}$$

and where:

$$\text{logitP} = \log\left(\frac{\text{P}}{1-\text{P}}\right)$$

It can be demonstrated that minimizing the objective C_{wlr} tends to give good calibration of the fused scores.[7,8] In practice, it is observed that changing the value of P has a small effect. The default of 0.5 is a good choice for a general application and it will be used in this work. The optimization objective C_{wlr} is convex and therefore has a unique global minimum. To find this minimum, a conjugate gradient algorithm can be used.[16]

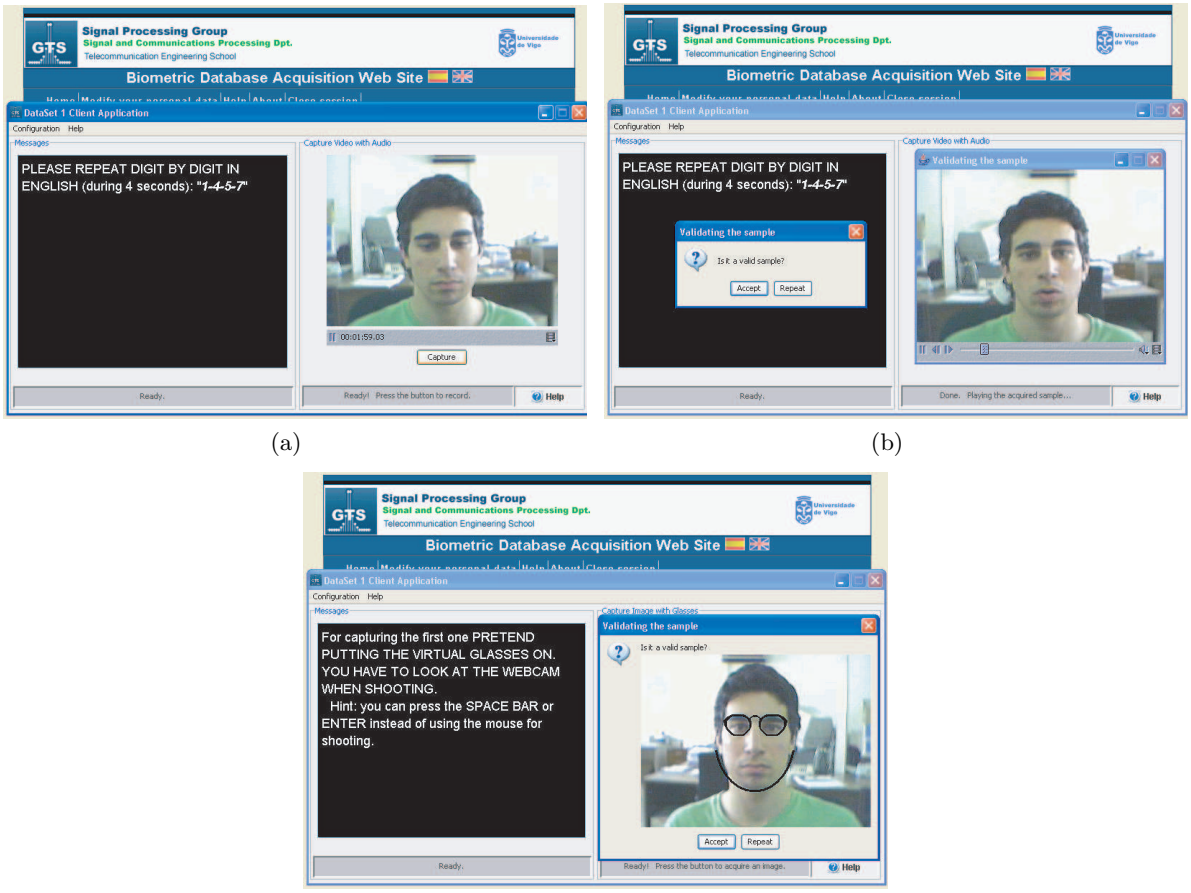

Figure 1. Graphical user interface of the application for the acquisition of DS1.

3. BIOSECURE MULTIMODAL DATABASE (MDB)

The Biosecure Multimodal Database (MDB) is a database whose acquisition has been recently conducted by 11 European institutions participating in the BioSecure Network of Excellence,[14] see Table 1. The MDB is comprised of three different datasets, namely: *i*) Data Set 1 (**DS1**), which is acquired over the Internet under unsupervised conditions (i.e. connecting to an URL and following the instructions provided on the screen); *ii*) Data Set 2 (**DS2**), which is acquired in an office room environment using a standard PC and a number of commercial sensors; and *iii*) Data Set 3 (**DS3**), which is acquired using a mobile handheld device under two acquisition conditions (controlled/indoor and uncontrolled/outdoor). Both DS2 and DS3 acquisition is managed by a human operator called supervisor..

The three datasets of the MDB include a common part of audio and video data (Common AV) which comprises still images of frontal face, and video with audio uttering of PINs and short sentences. Additionally, DS2 includes signature, fingerprint, hand and iris data, and DS3 includes signature and fingerprint data. Data of the three datasets have been acquired in two different sessions. The MDB has been collected in the period between November 2006 and May 2007, and currently is being built by its three acquisition coordinators (marked with * in Table 1). Still to be confirmed after the building process, the database will comprise around 950 users in DS1, 650 users in DS2, and 700 users in DS3. The MDB is expected to be built by fall 2007, although distribution policy and distribution dates still have to be defined.

Data Set 1 (DS1) : The purpose of DS1 is to acquire material over the Internet under unsupervised conditions. For DS1, the acquisition protocol consists of the common audiovisual part mentioned above. Therefore, the

Figure 2. Hardware devices used in the acquisition of DS2 together with acquisition samples.

modalities acquired in DS1 are: voice and face. The acquisition of DS1 is performed by connecting to an URL using a standard Internet browser and following the instructions provided on the screen. Acquisition is done using an standard webcam with microphone. In order to achieve realistic conditions, no specific webcam is imposed. The appearance of the graphical user interface of the application for the acquisition is shown in Figure 1. Figure 1a represents the state of the user interface prepared for the acquisition of an audiovisual sample. The left-side panel includes the instructions to the donor while the right-side panel shows the webcam stream. Figure 1b shows the graphical user interface once an audiovisual sample has just been acquired. The sample is presented to the donor in order to be validated before sending it to the

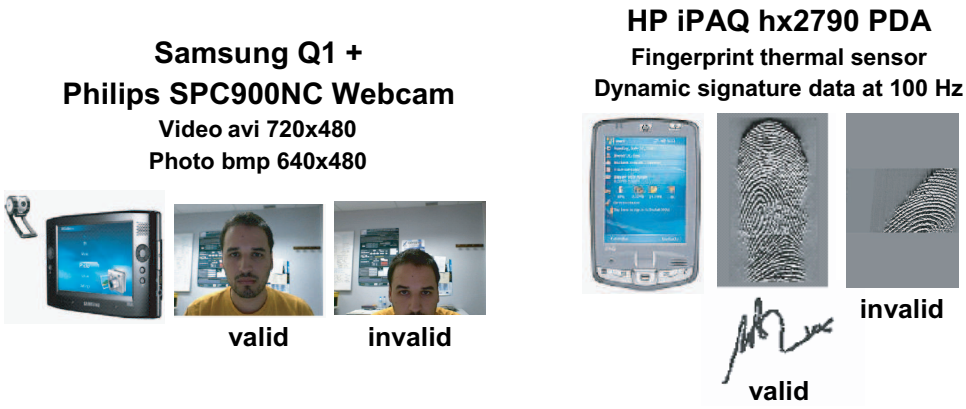

Figure 3. Hardware devices used in the acquisition of DS3 together with acquisition samples.

DS2	
Data type	**Sensor**
Signature	Tablet
Common AV	Webcam+headset
Iris image	Iris camera
Fingerprint	Optical, thermal
Hand	Digital camera
Face Still	Digital camera

DS3		
Data type	**Sensor**	**Condition**
Signature	IPAQ	indoor
Fingerprint	IPAQ	indoor
Common AV	Q1+Webcam	indoor
Common AV	Q1+Webcam	outdoor

Table 2. Data acquired in DS2 (left) and DS3 (right). AV = "Audio Video".

server. In Figure 1c, a frontal still image has just been acquired. The donor has to adjust the position of his face attending the overlaid "virtual glasses and chin" in order to normalize the pose.

Data Set 2 (DS2) : The scenario considered for the acquisition of DS2 is an office room environment. The acquisition is carried out using a standard desktop PC machine and a number of sensors connected to the PC via USB or Bluetooth interface.

The modalities acquired in DS2 are: voice, face, signature, fingerprint, hand and iris. Hardware devices used in the acquisition include a Windows-based PC with a USB hub, and the biometric sensors specified in Figure 2. For DS2, the data acquired is described in Table 2 (left).

Data Set 3 (DS3) : The objective of DS3 is to have a multimodal dataset with several modalities acquired on mobile platforms. The modalities acquired in DS3 are: face, voice, fingerprint and signature. For audio-video recordings each session comprises 2 acquisition conditions, indoor and outdoor, performed during

MODE	DATA TYPE	SENSOR	CONTENTS
fnf1	Face still	Digital camera (high resolution)	Frontal face images
fa1		Webcam (low resolution)	
fo1, fo2, fo3	Fingerprint	Optical	1 right thumb, 2 right index
ft1, ft2, ft3		Thermal	3 right middle finger

Table 3. Biometric traits and biometric devices considered for the experiments.

MODALITY	REF. SYSTEM	QUALITY MEASURES
Face still	Omniperception SDK LDA-based face verifier[17]	Face detection reliability, Brightness, Contrast, Focus, Bits per pixel, Spatial resolution, Illumination, Uniform Background, Background Brightness, Reflection, Glasses, Rotation in plane, Rotation in Depth, Frontalness
Fingerprint	NIST fingerprint system[18]	Texture richness based on local gradient[19]

Table 4. Reference systems and quality measures used in the experiments.

DATASETS		Num. of match scores per subject	
		Development (51 subjects)	Evaluation (156 subjects)
Session 1	Genuine	1	1
	Impostor	103×4	206×4
Session 2	Genuine	2	2
	Impostor	103×4	126×4

Table 5. Experimental protocol.

the same day. Hardware devices to be used for the acquisition include the biometric sensors specified in Figure 3, and the data acquired in DS3 is described in Table 2 (right).

The two acquisition conditions considered for audio video recordings are intended to comprise different sources of variability. Indoor acquisitions are done in a quiet room, just changing the position between each audio video sequence. Outdoor acquisitions are done in noisy ambiences such as office corridors, the street, etc., allowing the donor to move and to change position during and between each audio video sequence. For signature and fingerprint, only a condition is considered, which is considered degraded condition with respect to DS2: signatures and fingerprints are acquired while standing with the PDA in the hand.

4. EXPERIMENTS

4.1 Dataset and experimental protocol

As dataset for our experiments, we use the *development* set of scores of the *DS2 Evaluation (Access Control Scenario)* of the BioSecure Multimodal Evaluation Campaign.[13,14] This evaluation campaign has been conducted during 2007 by the BioSecure Network of Excellence,[14] as a continuation of the acquisition campaign of the Biosecure Multimodal Database. The aim of the *DS2 Evaluation* is to compare the performance of multi-modal fusion algorithms, assuming that the environment is relatively well controlled and the users are supervised. We focus on the *quality-based evaluation*,[20] whose objective is to test the capability of a fusion algorithm to cope with query biometric signals originated from heterogeneous biometric devices. The biometric traits and biometric devices considered are shown in Table 3, which are extracted from the DS2 dataset of the Biosecure Multimodal Database (see Section 3). Several reference systems and quality measures have been used with the biometric modalities in order to compute the scores for the Evaluation, see Table 4.

As described above, DS2 has been collected over two sessions. Each session has two samples per trait and per biometric device. The first sample of session one is considered as the template, whereas the remaining three samples are considered as query data. The experimental protocol is summarized in Table 5. A total of 333 subjects extracted from the database have been used for the evaluation, among them 207 are considered "clients" for whom a template is reserved for each of them (51 "clients" are used for the *development* set and 156 for the *evaluation* set). The remaining 126 subjects are considered an external population of users who serve as zero-effort impostors.

<claimed ID>
<fnf1|xfa1 score> <quality measures of template of fnf1|xfa1> <quality measures of query of modality fnf1|xfa1>
<fo1|xft1 score> <quality measures of template of fo1|xft1> <quality measures of query of modality fo1|xft1>
<fo2|xft2 score> <quality measures of template of fo2|xft2> <quality measures of query of modality fo2|xft2>
<fo3|xft3 score> <quality measures of template of fo3|xft3> <quality measures of query of modality fo3|xft3>

Table 6. Data format of the set of development scores.

MIXTURE	MODALITIES	FACE	FINGERPRINT
1	(fnf1/fo1/fo2/fo3)	Good quality	Good quality
2	(fnf1/xft1/xft2/xft3)	Good quality	Bad quality
3	(xfa1/fo1/fo2/fo3)	Bad quality	Good quality
4	(xfa1/xft1/xft2/xft3)	Bad quality	Bad quality

Table 7. Possible mixtures for each access.

The *development* impostor set of scores contains 103×4 samples per subject, meaning that when the reference subject is considered a template, all the 4 samples of the half of the remaining 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. For the *evaluation* impostor set of scores of session 1, all the remaining 206 subjects are used. In session 2, the *evaluation* impostor set of scores contains 126 subjects due to the external population set apart as zero-effort impostors. In this way, a fusion algorithm will not have already seen the impostors during its training stage; hence, avoiding systematic and optimistic bias of performance.

Prior to the evaluation, the *development* set (with both sessions 1 and 2) are released to the participants. It is recommended to use only session 2 of the development set as training data, since session 1 may be optimistically biased. In this work, we follow this recommendation, using only session 2 of the development set for our experiments, thus having $51 \times 2 = 102$ genuine score vectors and $51 \times 103 \times 4 = 21012$ impostor score vectors. The set of scores provided is a text file, with each line representing an access request. For the quality-based evaluation, each line has the structure shown in Table 6. Mode *xfa1* of Table 6 is the mismatched counterpart of *fnf1*, i.e. the template is captured using the high resolution camera (good quality) and the query image is captured using a webcam (low quality). Similarly, *xft1* (*xft2*, *xft3*) is the mismatched counterpart of *fo1* (*fo2*, *fo3*), i.e. the template is captured using the fingerprint optical sensor (good quality) and the query image is captured using the thermal sensor (bad quality). Notation "|" means "either ... or", so the two streams will be mixed during the *evaluation* and the fusion classifier will have to determine from which device the query is extracted. The mixture will have the combinations shown in Table 7 (it should be noted that for a given access all fingerprints will be acquired with the same device[20]). For our experiments, we separate each line of the set of scores (see Table 6) into the four possible combinations of Table 7, thus having four score subsets of equal size, one per combination.

It should be noted that there may be missing data in the set of scores and quality measures due to the fact that some matchings or quality estimates may not be computable by the algorithms used in the evaluation. It is not the target of this paper to study the effect of missing values. Therefore, prior to the experiments, we have corrected the missing values as follows. When a genuine (impostor) score of an specific sensor is missing, its value is set to the mean value of the remaining valid genuine (impostor) scores over the development set. Similarly, when a quality measure of an specific sensor is missing, its value is set to the mean value of the remaining valid measures.

Face quality feature	Global error	Error fnf1	Error xfa1
8	0.20%	0.04%	0.37%
6-8	0.20%	0.04%	0.37%
8-9	0.20%	0.04%	0.37%
6-8-9	0.08%	0.04%	0.12%
1-6-8	0.20%	0.28%	0.12%

Fingerprint quality feature	Global error	Error fo	Error xft
2	14.92%	21.81%	8.03%
1-2	16.68%	22.89%	10.47%
2-3-6	15.75%	22.08%	9.41%

Table 8. Quality feature combination for the estimation of the device used for the query acquisition (development set).

4.2 Fusion strategy: estimating the input device

The first step of the proposed fusion algorithm is to estimate from which device the query is extracted in each access, both for face and fingerprint. For this purpose, we use the quality measures provided together with the scores, supposing that:

- if the template and the query are from the same device (i.e. fnf1, fo1, fo2, fo3), both images should have similar quality values and they should be high,
- if the template and the query are from different devices (i.e. xfa1, xft1, xft2, xft3), the quality value of the template should be higher than the quality value of the query, and the quality value of the query should be low.

We estimate the device separately for face and fingerprint modality. We use a quadratic discriminant function with multivariate normal densities for each class.[15] For the face modality, we use the 14 quality measures of the query image (see Table 4). For the fingerprint modality, we derive the following 8 parameters from the quality of the templates (Q_{ti}) and queries (Q_{qi}) of the three scores corresponding to each access ($i = 1, 2, 3$): 1) Number of fingerprint scores such as $Q_{ti} > Q_{qi}$, 2) Max (Q_{qi}), 3) Max ($|Q_{ti} - Q_{qi}|$), 4) Min (Q_{qi}), 5) Min ($|Q_{ti} - Q_{qi}|$), 6) Mean (Q_{qi}), 7) Mean ($|Q_{ti} - Q_{qi}|$), and 8) Max ($Q_{ti} - Q_{qi}$).

We have tested all the combinations of one, two and three parameters in order to determine the device used for the query acquisition. Results of the best cases are shown in Table 8. We observe that for the face modality, estimation can be done at a remarkably low error rate, even with only one parameter. On the other hand, we observe high error in the estimation for the fingerprint modality. Interestingly enough, the estimation fails mostly with the optical sensor. This means that for the thermal sensor, the quality value of the template is higher than the quality value of the query, and that the quality value of the query is low. Such assumption is not true for the optical one.

4.3 Fusion strategy: architecture

For our fusion experiments, we have used the evaluation tools for C_{wlr} included in the toolkit FoCal.[16] Since the error rate in the estimation of the face device is low, we train one linear logistic regression classifier for each face modality (fnf1 or xfa1). However, as we are not able to reliably estimate the fingerprint sensor used in the access, we train a unique classifier using both fingerprint modalities (fo and xft). The architecture of the proposed fusion mechanism is shown in Figure 4. For each access, we compute one calibrated face score s_{face} and one calibrated fingerprint score s_{finger} which combines the three fingerprint scores provided. Supposing independence between them (since s_{face} and s_{finger} are computed from different biometric traits), their sum will also be a calibrated score:[15]

$$s_{fused-loglik-sum} = s_{face} + s_{finger}$$

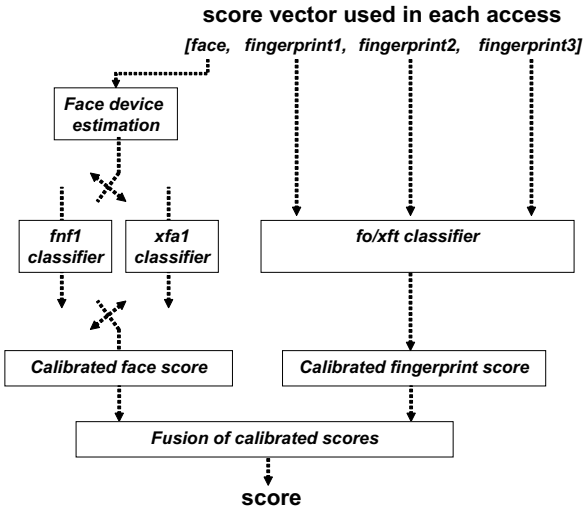

Figure 4. Architecture of the proposed fusion strategy.

Mixture	Modalities	loglik SUM	Arithm. mean	Minimum	Maximum	Geom. mean
1	(fnf1/fo1/fo2/fo3)	3.92%	2.94%	8.56%	1.82%	3.92%
2	(fnf1/xft1/xft2/xft3)	4.90%	5.88%	10.00%	12.88%	5.88%
3	(xfa1/fo1/fo2/fo3)	1.96%	1.95%	7.24%	1.96%	2.93%
4	(xfa1/xft1/xft2/xft3)	5.88%	8.82%	13.68%	19.45%	8.80%
	ALL	**4.17%**	**5.39%**	**9.73%**	**9.39%**	**4.90%**

Table 9. Verification results of the fusion in terms of EER (%) - development set.

4.4 Results

We compare the proposed fusion scheme to a set of simple fusion rules of the four scores of an access (arithmetic mean, minimum, maximum and geometric mean).[9] We first compute a normalization scheme using the MAX-MIN normalization rule[6] as follows:

$$MIN = \mu_I - 2 * \sigma_I$$
$$MAX = \mu_G + 2 * \sigma_G$$

where μ_I and σ_I (μ_G and σ_G) are the mean and the standard deviation values of the impostor (genuine) scores. Similarly, we use a normalization scheme for each face modality, and a single normalization scheme for both fingerprint modalities.

Table 9 summarizes the verification results of the evaluated fusion schemes (row named ALL). In Fig. 5, verification results of the proposed fusion rule $s_{fused-loglik-sum}$ are depicted, together with the best simple fusion rule (the geometric mean) and the individual modalities of the dataset. We also report in Table 9 the performance of the four possible combinations for each access. We observe that the overall performance of the proposed fusion scheme is better than the performance of all simple fusion rules. The best simple fusion rule (geometric mean) performs 18% worse than $s_{fused-loglik-sum}$ in terms of EER. This difference is even higher for low FRR and low FAR, as can be seen in Fig. 5. Interestingly enough, it can be observed that the simple maximum and arithmetic mean rules perform better for the single combination involving the two high quality sensors (i.e. the mixture 1). However, the proposed scheme performs better for the other combinations involving at least one low quality sensor and in any case, it is the best one to cope with query biometric signals originated from different devices.

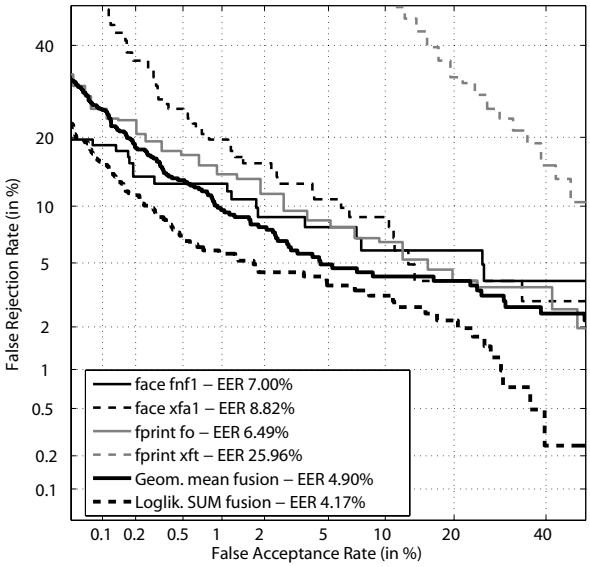

Figure 5. Verification results of the individual sensors and of the fusion (development set).

5. CONCLUSIONS

Fusion at the matching score level is widely used in multibiometric systems due to the ease in accessing the scores of different matchers.[6] In this work, we use linear logistic regression fusion[7,10] to perform multibiometric fusion. This mechanism fuses the scores in such a way that output scores are mapped to log-likelihood-ratios, thus being in an application-independent domain. When output scores tend to be log-likelihood-ratios, it is said that they are *calibrated*.[8]

This paper compares a fusion scheme based on calibrated scores with simple fusion rules. We use for our experiments the *development* set of scores of the *DS2 Evaluation (Access Control Scenario)* of the BioSecure Multimodal Evaluation Campaign,[13,14] which contains face still samples collected with two cameras of different resolution and fingerprint samples collected both with an optical and a thermal sensor. The aim of this evaluation is to compare the performance of multi-modal fusion algorithms when query biometric signals are originated from different biometric devices. In the proposed fusion strategy, we first estimate the device used in the access and next, we apply a different linear logistic regression classifier adapted to each device. Since linear logistic regression classifiers produce log-likelihood-ratios, output scores produced by the different devices are then easily combined.

We demonstrate the effectiveness of the proposed approach by comparing it to a set of simple fusion rules with standard MAX-MIN normalization. Reported results show that the proposed fusion approach outperforms all the simple fusion rules, with the advantage that scores after logistic regression are mapped to an application-independent range. In addition, the proposed fusion approach is able to cope with missing values of one modality, which will be the source of future work.

ACKNOWLEDGMENTS

This work has been supported by Spanish project TEC2006-13141-C03-03, and by European Commission IST-2002-507634 Biosecure NoE. Author F. A.-F. thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his PhD studies. Author J. F. is supported by a Marie Curie Fellowship from the European Commission.

REFERENCES

1. A. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Transactions on Information Forensics and Security* **1**, pp. 125–143, 2006.

2. A. Jain and A. Ross, "Multibiometric systems," *Communications of the ACM, Special Issue on Multimodal Interfaces* **47**, pp. 34–40, 2004.
3. 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 Anal Mach Intell* **27**, pp. 450–455, 2005.
4. J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Adapted user-dependent multimodal biometric authentication exploiting general information," *Pattern Recognition Letters* **26**, pp. 2628–2639, 2005.
5. J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Discriminative multimodal biometric authentication based on quality measures," *Pattern Recognition* **38**(5), pp. 777–779, 2005.
6. A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition* **38**, pp. 2270–2285, 2005.
7. N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Scwartz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Signal Processing* , 2007. To appear.
8. N. Brummer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language* **20**, pp. 230–275, 2006.
9. J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans Pattern Anal Mach Intell* **20**, pp. 226–239, 1998.
10. S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions," *Digital Signal Processing* **10**, pp. 237–248, 2000.
11. J. Gonzalez-Rodriguez and D. Ramos, *to appear in Speaker Classification Collection, Volume 1: Fundamentals, Features and Methods. Springer LNAI-4343*, ch. Forensic Automatic Speaker Classification in the Coming Paradigm Shift. Springer, 2007.
12. J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *to appear in IEEE Trans. on Audio, Speech and Language Processing, Special Issue on Recent Advances in Speaker and Language Recognition* , 2007.
13. BMEC *The BioSecure Multimodal Evaluation Campaign - http://www.int-evry.fr/ biometrics/ BMEC2007/ index.php* , 2007.
14. BioSecure, "Biometrics for Secure authentication, FP6 NoE, IST - 2002-507634 - http://www.biosecure.info," 2004.
15. R. Duda, P. Hart, and D. Stork, *Pattern Classification - 2nd Edition*, 2004.
16. N. Brummer, "Focal toolkit," *Available in http://www.dsp.sun.ac.za/ nbrummer/focal/* .
17. A. Martinez and A. Kak, "PCA versus LDA," *IEEE Trans Pattern Analysis and Machine Intelligence* **23**(2), pp. 228–233, 2001.
18. C. Watson, M. Garris, E. Tabassi, C. Wilson, R. McCabe, and S. Janet, *User's Guide to Fingerprint Image Software 2 - NFIS2 (http://fingerprint.nist.gov/NFIS)*, NIST, 2004.
19. Y. Chen, S. Dass, and A. Jain, "Fingerprint quality indices for predicting authentication performance," *Proc. AVBPA* **LNCS-3546**, pp. 160–170, 2005.
20. N. Poh and T. Bourlai, "The BioSecure desktop DS2 evaluation documentation," *http://biosecure.ee.surrey.ac.uk/* .