Quality Dependent Fusion of Intramodal and Multimodal Biometric Experts

J Kittler [a], N Poh[a], O Fatukasi[a], K Messer[a], K Kryszczuk[b], J Richiardi[b] and A Drygajlo[b]

[a] CVSSP, University of Surrey Guildford, GU2 7XH Surrey, UK.
[b] EPFL, Laboratory of IDIAP, Lausanne, Switzerland.

ABSTRACT

We address the problem of score level fusion of intramodal and multimodal experts in the context of biometric identity verification. We investigate the merits of confidence based weighting of component experts. In contrast to the conventional approach where confidence values are derived from scores, we use instead raw measures of biometric data quality to control the influence of each expert on the final fused score. We show that quality based fusion gives better performance than quality free fusion. The use of quality weighted scores as features in the definition of the fusion functions leads to further improvements. We demonstrate that the achievable performance gain is also affected by the choice of fusion architecture. The evaluation of the proposed methodology involves 6 face and one speech verification experts. It is carried out on the XM2VTS data base.

Keywords: Biometric authentication, multiple classifier system, intramodal fusion, multimodal fusion, quality dependent fusion

1. INTRODUCTION

Biometric authentication is a process that uses a person's physical and behavioural characteristics to verify a claim. Classifier fusion is a process of combining output from different experts to provide error reduction in verifying a claim. Fusion can occur at three different levels; feature level fusion, score level fusion, and decision level fusion. In this paper, we are concerned with fusion at the score level. Several studies including [1–3] have shown that fusing experts improves the efficiency and the accuracy of a system. The evidence available suggests that a gain in accuracy can be achieved either with intramodal fusion or multimodal fusion. However the latter also offers better authentication system robustness against the failure of individual biometric modalities to acquire reliable biometric data either as a result of adverse environmental factors, user imposed constraints, or undesirable sensor characteristics. Unfortunately, at the same time, various studies have also shown that poor quality biometric data leads to reduction in the accuracy of the system.[4,5] The biometric data quality degradation is particularly acute in the operational phase, due to the fact that in most cases the data capture for training purposes is controlled by someone supervising the acquisition process, but during authentication supervision is not always possible. The investigation of fusion controlled by Quality Measures (QM) has shown that quality dependent fusion can offer a significant performance gain over fusion without quality.[6–12]

The original motivation for using sample quality information in fusion emerged in relation to multimodal biometrics systems where it is easy to comprehend that the modalities performing poorly as a result of degraded quality of biometric information should influence the final decision to a lesser extent than those based on good quality data. This suggested a quality dependent weighting of modalities as the obvious solution to the fusion problem. However, in our experiment, there is an indication that quality measures can also be used to improve the performance of individual biometric modalities by effectively offering a dynamic decision rule with a quality dependent threshold (Section 4). These examples of a completely different ways of using biometric data quality information raises the question whether quality should be considered as a feature or as a confidence coefficient.

Prior studies in this direction include.[6,7,9,13] Karthil *et al.* proposed a likelihood ratio-based approach to achieve quality dependent score fusion.[9] This approach allows the designer to get as much as possible from

Further author information: (Send correspondence to N.P.)
E-mails: {J.Kittler, Norman.Poh, O.Fatukasi, k.messer}@surrey.ac.uk, Telephone: Tel: +44 1483 68 {9294, 6042, 6046}
E-mails: {Krzysztof.Kryszczuk, Jonas.Richiardi, Andrzej.Drygajlo}@epfl.ch, Telephone: +41 21 693 {4691, 4691, 4328}

Biometric Technology for Human Identification IV
edited by Salil Prabhakar, Arun A. Ross Proc. of SPIE Vol. 6539
653903, (2006) · 0277-786X/06/$18 · doi: 10.1117/12.724008

each modality for any level of quality of the input data. At the same time, the less informative modalities will produce likelihood ratio close to one and will therefore not influence the decision making process. Fierrez-Aguilar *et al.* proposed a method where the fusion procedure is dynamically adapted each time a claim is made based on the estimated quality.[6] This is achieved by using quality for controlling the penalty function of the SVM learning criterion. Moreover, quality measures also weigh the relative influence of the respective modalities and the joint decision making system. Intuitively, the approach enables the multimodal system to focus on the single modality of dominant quality or for comparable qualities on the joint decision making system. Unfortunately, as a result of the SVM training strategy the joint decision making system is optimised for good quality data. Bigun *et al.* proposed the Bayesian Conciliation method.[7] This method relies on two components known as a client and an impostor supervisor. The client supervisor estimates the expected true authenticity score of a claim based on its expertise in recognising client data (likewise for the impostor supervisor). The final decision is made by taking into account the different expertise of the two supervisors and choosing the one which comes closest to its goal, which is defined as zero for impostor supervisor and one for client supervisor. Effectively, the supervisor adapts to each identity claim as a function of the quality of the input data. Kryszczuk *et al.* proposed to use a *derived* quality measure[13] instead of raw quality measures as done in.[6,7,9,13] The derived quality measure, or confidence is defined as the posterior probability of making the correct decision given some observed evidences, which include both the system output and raw quality measures. In the context of bimodal fusion, this means that if the decision of two systems are conflicting (different), one takes the decision of the system which is more likely to be correct.

The aim of this paper is to investigate the merit of using as features not only quality measures but also the cross terms obtained by taking the *product* of score and quality. This generalises the fusion feature space. The study also looks at several architectures that may be appropriate in different circumstances, namely when score and quality data for each expert and modality is made available to the fusion stage, and the situation where each modality delivers quality dependent scores for integration in the fusion system. We show that using quality weighted scores as features in the definition of the fusion functions leads to improved performance. We also demonstrate that the achievable performance gain is also affected by the choice of fusion architecture.

The paper is organised as follows: Section 2 develops the methodology of quality dependent fusion. Section 3 presents the database, verification systems and evaluation methodology. Section 4 presents the experimental results. Finally, conclusions are drawn in Section 5.

2. METHODOLOGY OF QUALITY DEPENDENT FUSION

Let $x \in \mathbb{R}^R$ be the vector of output scores of R experts, $q \in \mathbb{R}^P$ be the vector of P quality measures and $k \in \{C, I\}$ be one of the two possible classes of users, i.e., genuine users or clients and impostors. From the Bayesian point of view, the *generative* and *discriminative* approaches which incorporate the quality information directly can be written as follows:

$$y_{com}^{llr} \equiv f^{llr}(x,q) = \log \frac{p(x,q|C)}{p(x,q|I)} \tag{1}$$

$$y_{com}^{prob} \equiv f^{prob}(x,q) = P(C|x,q) \tag{2}$$

In practice (2) is approximated by:

$$P(C|x,q) \approx \mathrm{sigmoid}\big(f^{disc}(x,q)\big) = \frac{1}{1+\exp\left(-f^{disc}(x,q)\right)} \tag{3}$$

where the output $f^{disc}(x,q) \in [-\infty, \infty]$ does not have to be associated with probability. $f^{disc}(x,q)$ is known as a discriminative function and very often, based on the sign of its output, one classifies x as either belonging to a client or an impostor. In this study, the function $f^{llr}(x,q)$ is estimated by a Gaussian Mixture Model;[14] the function that outputs posterior probability, $f^{prob}(x,q)$, is estimated by logistic regression;[15] and the discriminative function $f^{disc}(x,q)$ is estimated by a support vector machine.[16] Note that the logistic regression function used here is different from the simplified version proposed by Pigeon *et al.* [17] which was the first reported attempt at biometric fusion by logistic regression. While this simplified version makes the class conditional Gaussian assumption over x, the one we use is purely data driven and its parameters are obtained by the "gradient ascent" algorithm[15] which is in fact a realisation of the maximum likelihood principal.

The conventional approaches without using the quality information can also be divided into either generative or discriminative. They can be written in the similar ways as in (1) and (2) except that q is not used, i.e.,:

$$y_{com}^{llr} \equiv f^{llr}(x) = \log \frac{p(x|C)}{p(x|I)} \tag{4}$$

$$y_{com}^{prob} \equiv f^{prob}(x) = P(C|x). \tag{5}$$

While in theory, i.e., from the Bayesian view point, both the generative and discriminative approaches of combining the quality information are equivalent, i.e.,

$$f^{prob}(x,q) = \text{sigmoid}\left(f^{llr}(x,q) + \log \frac{P(C)}{P(I)}\right)$$

and

$$f^{disc}(x,q) \propto f^{llr}(x,q)$$

in practice, we argue that the discriminative approach may be better than the generative one. This is because a discriminative classifier aims to optimise the decision boundary and as a result, is more robust to the drift in score distribution.

We consider here the case where the system outputs, x, can be obtained from the same biometric modality or from different modalities. For this reason, we introduce $x_{m,i}$ to denote the i-th classifier of the m-th biometric modality. There are I_m systems for the m-th modality and M biometric modalities are available. As a result, the number of systems available for fusion, P, is $\sum_m I_m$. In our experiments, we use only face and speech modalities, i.e, $m \in \{\mathtt{F}, \mathtt{S}\}$. The number of systems therefore is $F + S$ where $F = I_{\mathtt{F}}$ and $S = I_{\mathtt{S}}$.

In general, one expects higher dependency among the system outputs sharing the same biometric modality and, in contrast, independence when the system fuses different biometric modalities. By assuming different types of system output dependency, we identify the following three types of fusion architecture, in increasing level of complexity:

1. **Multi-stage single processing (MSSP)**: This architecture is a result of assuming independence among all the system outputs despite the fact that systems sharing the same biometric modality may be dependent. It can be written as:

$$y_{com}^{MSSP} = \prod_m \prod_i P(C|x_{m,i}, q) = \prod_m \prod_i f^{prob}(x_{m,i}, q) \tag{6}$$

   Note that since $f(x_{m,i}, q)$ operates on a single system at a time, it can be considered a quality-dependent score normalisation procedure. It is therefore not a deterministic one-to-one mapping function as studied in[18] but rather a function of $x_{m,i}$ and q jointly. Note that discriminative functions $f^{disc}(x,q)$, e.g., a Support Vector Machine (SVM), do not output scores which satisfy the axiomatic properties of probabilities and cannot therefore be used in conjunction with a product fusion rule. For this reason we investigated fusion by the sum equivalent of (9), given as

$$y_{com}^{MSSP} = \sum_m \sum_i f^{disc}(x_{m,i}, q) \tag{7}$$

   Note that by using the sum rule, one implicitly assumes that the class-conditional distributions of the outputs $f^{disc}(x_{m,i}, q)$ across all m and i are comparable. This is, in general, not the case, thus implying the need for normalising the outputs. Fortunately, this can be avoided by normalising the input to the function $f^{disc} : \mathbb{R}^{R+P} \to \mathbb{R}$. Suppose that each of the $\mathbb{R}^{R+P}$ input elements is normalised to having zero mean and unit variance (across all the training examples), and the same complexity of $f^{disc}(x_{m,i}, q)$ is used for all m and i, then the outputs $f^{disc}(x_{m,i}, q)$ will be comparable. For the generative approach, using the sum rule, i.e.,

$$y_{com}^{MSSP} = \sum_m \sum_i f^{llr}(x_{m,i}, q), \tag{8}$$

is a direct implication of assuming independence among the output of systems $x_{m,i}$ for all m and i. Whether with sum or product, we refer to the fusion assuming independent system outputs as Architecture 1.

2. **Multi-stage joint processing (MSJP)**: This architecture takes into consideration the dependency among system outputs sharing the same biometric modality yet ignores the dependency of the system outputs coming from different biometric modalities. It can be written as:

   $$y_{com}^{MSJP} = \prod_m P(C|x_m, q) = f^{prob}(x_m, q), \tag{9}$$

   where x_m denotes a vector the components of which are the system outputs sharing the m-th biometric modality, i.e., $x_m \equiv [x_{m,1}, \ldots, x_{m,P_m}]$. The practical implication of this architecture is that one designs a fusion classifier per biometric modality and then combines all M resulting fusion classifiers using a fixed rule, e.g., the product rule for $f^{prob}(x_m, q)$ and the sum rule for $f^{disc}(x_m, q)$ and $f^{llr}(x_m, q)$.

   This architecture is referred to as Architecture 2.

3. **Single-stage joint processing (SSJP)**: This architecture does not assume system output independence. It can be written as:

   $$y_{com}^{SSJP} = P(C|x, q) = f^{prob}(x, q), \tag{10}$$

   where x is a vector containing all the system outputs, i.e., $x = \{x_{m,i}|\forall_{i,m}\}$. The function $f^{prob}(x, q)$ is simply replaced by $f^{llr}(x, q)$ when using a GMM classifier and by $f^{prob}(x, q)$ when using logistic regression. This architecture is referred to as Architecture 3.

In the discussion that follows, we will focus on training the discriminative function $f^{disc}(x, q)$. However, the discussion generalises to the functions $f^{llr}(x, q)$ and $f^{prob}(x, q)$. For this reason, we will use the generic term $f(x, q)$ and refer to one of the three particular fusion algorithms, i.e., $f^{llr}(x, q)$, $f^{prob}(x, q)$ or $f^{disc}(x, q)$, only when necessary.

For any function $f(x, q)$, the following decision function is used:

$$\text{decision}(f(x, q)) = \left\{ \begin{array}{ll} \text{accept} & \text{if } f(x, q) > \Delta \\ \text{reject} & \text{otherwise} \end{array} \right\}, \tag{11}$$

where Δ is a threshold tuned *a priori* to minimise a specific criterion on a separate development set. This will be discussed in Section 3. Note that the decision threshold is trained independently from the fusion classifier $f(x, q)$. This is sensible because the fusion algorithms we used, i.e., SVM, GMM and logistic regression, are not designed specifically to minimise biometric performance, e.g., equal error rate (EER). For instance, SVM maximises margin; and the optimisation algorithms for GMM and logistic regression follow the maximum likelihood principle.

In both the discriminative and generative approaches, jointly estimating x and q is a challenging problem. Suppose that one uses a linear function in $f(x, q)$ to distinguish the client class from the impostor one. In this case a weight will be associated with each element in x and q. The result after training is that magnitude of the weight associated with q will be comparatively small because q has no discriminative information. This suggests that using non-linear function of $f(x, q)$ may be more useful.

One way to introduce non-linearity is by using some kind of expansion between y and q, i.e., $x \otimes q$, where $\otimes$ is called a *tensor product*. Note that x and q are not vectors of the same length. If x has R elements and q has P elements, then $x \otimes q$ will result in $P \times R$ elements and each element is a product between a pair of the elements

Table 1. The complexity of the function $f(x,q)$ when implemented using a linear classifier, in increasing level of complexity due to different input arrangements.

no.	arrangement	the resulting function $f^{disc}(x,q)$	no. of parameters
1	$[x]$	$\sum_i x_i w_i$	R
2	$[x,q]$	$\sum_i x_i w_i + \sum_j q_j v_j$	$R+P$
3	$[x, x \otimes q]$	$\sum_i x_i \left(\sum_j q_j w_{i,j} + w_i\right)$	$R \times (P+1)$
4	$[x, q, x \otimes q]$	$\sum_i x_i \left(\sum_j q_j w_{i,j} + w_i\right) + \sum_j v_j q_j$	$R + P + R \times P$

in x and q. Therefore, when training $f(x,q)$, we need to feed the fusion classifier with inputs $[x, q, x \otimes q]$ instead of $[x,q]$. When one uses $[x, q, x \otimes q]$, the linear function can be written as:

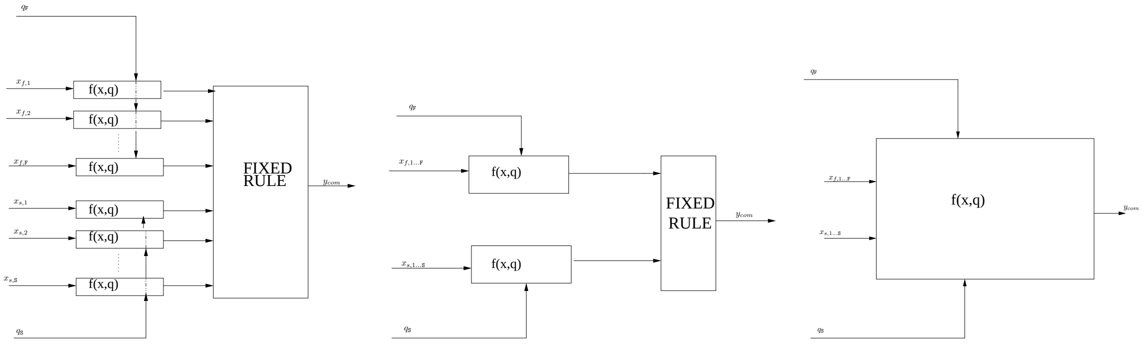
$$\begin{aligned} f(x,q) &= \sum_i \sum_j w_{i,j} x_i q_j + \sum_i w_i x_i + \sum_j v_j q_j \\ &= \sum_i x_i \left(\underbrace{\sum_j q_j w_{i,j}} + w_i \right) + \underbrace{\sum_j v_j q_j}, \end{aligned} \tag{12}$$

where the weight $w_{i,j}$ is associated with $x_i q_j$, the weight w_i is associated with x_i, and v_j is associated with q_j. In this notation, x_i is an element of vector x and q_j is an element of vector q. (12) clearly shows that the resulting classifier is *linear* except that the weight is modified *dynamically* by the quality measures via the first underbraced term. The second underbraced term shows that q *dynamically* adjusts the decision threshold in addition to that controlled by Δ shown in (11). In theory, due to the weights $w_{i,j}$ and v_j, one does not need to normalise the quality measures. The same argument also applies to w_i. In practice, however, normalising the elements in x and q can accelerate convergence of a particular chosen algorithm. More over, for some fusion rules such as product, it may be absolutely essential. To this end, we use the z-norm such that after the normalisation, a variable has zero mean and unit variance over all the examples in the training set and a min-max normalisation.

We outline here several possible "arrangements" in Table 1, presented in the order of increasing complexity, i.e., the number of parameters. We will write $f([x,q])$ to explicitly refer to the second arrangement, $f([x, x \otimes q])$ to refer to the third arrangement, etc. The second column shows the four possible arrangements, i.e., the way the features are used as input to a fusion algorithm. The third column shows the resulting discriminative linear function $f^{disc}(x,q)$. While similar analyses cannot be done for the linear discriminative function $f^{prob}(x,q)$ (due to the sigmoid function) and for the generative function $f^{llk}(x,q)$, our purpose of showing the elements in the expanded input vector along with their associated weight parameters is to illustrate the complexity of each arrangement. For instance, the first arrangement, i.e., $f([x])$, does not use any quality information. The second arrangement, i.e., $f([x,q])$ does not contain any interaction between x and q. However, it considers the case where the decision threshold may be modified by q. In the third arrangement, one creates a linear classifier whose weights can change dynamically as a function of q. The last arrangement, i.e., $f([x, q, x \otimes q])$ or (12), is the most general one since it contains all possible interactions between x and q of the first three arrangements.

The quality-enhanced discriminative fusion classifier with the input $[x, x \otimes q]$ (the third arrangement) is structurally very similar to the one proposed in[19] where a reduced polynomial discriminative function was used. In our case, one can use *any* discriminative classifier to implement it. This is an elegant solution because one no longer needs to design a dedicated fusion algorithm such as those proposed in[19] and[6,7] to achieve the same goal.

When considering multimodal fusion, e.g., using face and speech in our case ($m \in \{\mathtt{S}, \mathtt{F}\}$), the $\otimes$ operator is applied to x and q of the *same* biometric modality, i.e., $[x_m \otimes q_m]$ for each m. For instance, for arrangement four, the resulting vector is $[x_\mathtt{S}, q_\mathtt{S}, x_\mathtt{S} \otimes q_\mathtt{S}, x_\mathtt{F}, q_\mathtt{F}, x_\mathtt{F} \otimes q_\mathtt{F}]$. By doing so we assume that there is no interaction between the quality measures of one modality with the system outputs of another modality, hence, resulting in less number of parameters needed to be estimated as compared to the full expansion $\big[[x_\mathtt{F} x_\mathtt{S}], [q_\mathtt{F} q_\mathtt{S}], [x_\mathtt{F} x_\mathtt{S}] \otimes [q_\mathtt{F} q_\mathtt{S}]\big]$.



(a) Architecture 1 (b) Architecture 2 (c) Architecture 3

Figure 1. The proposed three architectures. In each figure, the function $f(x, q)$ can be implemented by any of the four arrangements shown in Table 1. In our study, the function $f(x, q)$ is implemented using logistic regression, SVM and GMM. In Architecture 1, $f(x_{m,i}, q)$ for i-th system of the m-th modality can be considered a quality-dependent score normalisation procedure since $x_{m,i}$ is a scalar. In Architectures 2 and 3, the fixed rule is chosen to be a product for $f^{prob}(x, q)$ (approximated by logistic regression) and sum for $f^{llr}(x, q)$ (GMM) and $f^{disc}(x, q)$ (SVM).

To summarise, firstly, we have proposed three possible types of architecture for the task of multimodal fusion involving several systems per biometric modality, i.e., (6), (9), and (10) defining the discriminative function $f^{prob}(x, q)$ (based on logistic regression). They differ mainly in the different underlying assumptions made for the system outputs. Similar functions, with the exception of the sum rule, can be applied to the discriminative function $f^{disc}(x, q)$ (based on SVM) and the generative function $f^{llr}(x, q)$ (based on GMM). A graphical diagram of each of the three types of architecture is shown in Figure 2. Secondly, we have proposed four training strategies to estimate the functions $f(x, q)$. Note that the input $[x, q, x \otimes q]$ is a generalisation of the conventional fusion classifier and of the work reported in.[19] We argue that such an arrangement is potentially better than using $[x, q]$ in conjunction with a more complex non-linear classifier because it guards against overtraining, yet realises a *linear* classifier whose weights are modified *dynamically* by a weighted sum of quality measures.

3. DATABASE, SYSTEMS AND EVALUATION

3.1. Database

In the current study, we used the original XM2VTS database[20] and its degraded version[21] in both the training and the testing phase of the fusion methods. The original database contains mugshot images with well controlled illumination. The degraded ones, on the other hand, are images taken under strong side illumination, which has been shown to degrade significantly face verification performance.[21] This database contains 295 subjects, which includes 200 subjects selected to be clients, 25 to be impostors for the algorithm development (training), 70 to be impostors for algorithm evaluation (testing). For each subject, face and speech biometric modalities are acquired in four sessions; the first three are used for training the classifiers and the last one for testing. For the face modality we consider the dark dataset with left illumination as the ”fifth session” and the one with right illumination as the ”sixth” session. There is unfortunately no equivalent of degraded speech data that can be paired with the degraded face images. We did so by first introducing additive white noise with a uniform random distribution between 0 and 20dB signal-to-noise ratio on the clean speech database, hence resulting in a degraded speech database with exactly the same size as the clean database. We then paired the degraded face images with the degraded speech data according to Table 3.1. For instance, the first row shows that the first shot of degraded face image in the fifth session is matched with the second shot of the degraded speech recorded in session one, and so on. Building the database this way will give us two types of data sets: good and degraded quality data sets for both modalities. A shortcoming is that there is no scenario where one modality is of good quality and the other one is degraded. Although this is more realistic, there is no obvious solution to introducing this scenario. However, it will be investigated in the future.

Degraded face		Degraded speech	
session	shot	session	shot
5	1	1	2
5	2	2	2
6	1	3	2
6	2	4	2

Table 2. Matching of degraded face and speech data

Sessions	Shots	180 Clients	20 Clients	25 Imposter	70 Imposter
S1	1	Training	Training	Evaluation	Test
	2	Evaluation	Evaluation		
S2	1	Training	Training		
	2	Evaluation	Evaluation		
S3	1	Training	Training		
	2	Evaluation	Evalaution		
S4	1	Test	Test		
	2				
Degraded	L1,R1	Test degraded	Evalaution degarded	Evaluation degraded	Test degraded
	L2,R2				

Table 3. The XM2VTS clean and degraded protocol.

According to the original experimental protocol known as the Lausanne Protocols, the degraded data sets were not designed for algorithm development but for algorithm evaluation. In this paper, however, in order to show the advantage of having observed some degraded data, we used the 25-impostor data set in which good and degraded quality data is available. However, there is simply no degraded client data for algorithm development. Given the fact that these scores can only be found in the 200-client data set, we further divided this data set into 20- and 180-client data sets such that the 20-client data set is set aside uniquely for algorithm development and the 180-client for both algorithm development and evaluation. The resulting protocol for *the good and degraded quality scenarios* is summarised in Table 3.

3.2. Experts and Quality Measures

In this paper, we use a set of proprietary quality measures developed by Omniperception Ltd for the face image quality assessment. These measures are: "frontal quality", measuring the deviation from the frontal face; and "illumination quality", quantifying the uniformity of illumination of the face. It should be noted that none of these quality detectors were designed specifically to distinguish the three strong dominant quality states of the face images in the XM2VTS database: good illumination, left illumination and right illumination. Using the above quality measures makes the problem of quality-dependent fusion classifier more challenging.

The classifiers used for the face experts in this paper can be found in.[22] There are two classifiers with three types of pre-processing, hence resulting in a matrix of six classifiers. The two classifiers used are Linear Discriminant Analysis (LDA) with correlation as a measure of similarity[23] and Gaussian Mixture Model (GMM) with maximum a posteriori adaptation, described in.[24] The use of the GMM in face authentication can be found in.[25] The face pre-processing algorithms used include the photometric normalisation as proposed by Gross and Brajovic,[26] histogram equalisation and local binary pattern (LBP) as reported in.[22] The feature extraction and classification algorithms are implemented on the open-source Torch Vision Library[*].

Two quality measures are used for the speech system: signal-to-noise ratio (SNR) and "entropy quality". Both measures are used for voice activity detection, i.e., to separate speech from non-speech. According to the "Murphy algorithm",[27] an SNR is obtained by calculating the magnitude or energy of speech versus that of non-speech in decibels. The entropy quality[28] measures the degree of peakiness of the distribution of the power spectrum within an observed short-term window of speech frames. A speech signal has consistent energy and

[*] Available at "http://torch3vision.idiap.ch". See also a tutorial at "http://www.idiap.ch/ marcel/labs/faceverif.php".

thus its resulting distribution is likely to be peaky. This results in low entropy values. In contrast, a pause or random signal will have a wider distribution of energy, thus resulting in high entropy values. Both these measures can be found in.[29] The speech system used is based on the ALIZE toolkit.[30]

3.3. *A posteriori* versus *a priori* Performance Evaluation

In this paper, two types of performance are quoted: *a posteriori* and *a priori* performance. In the former, one assumes that the class-conditional score distributions of the test set are known. Consequently, the decision threshold can be tuned directly on the test score set. This assumption is somewhat unrealistic in a real application. However, the advantage is that this method allows for the evaluation to be carried out without the need to address the global threshold selection problem. In the latter case, one needs to devise such a strategy and this requires an additional development (combined) score data set. We use both type of evaluations. For the *a priori* evaluation, we obtain the threshold by minimising the weighted error rate, which is defined as:

$$\mathrm{WER}_\alpha(\Delta) = \alpha\mathrm{FAR}(\Delta) + (1-\alpha)\mathrm{FRR}(\Delta), \tag{13}$$

for $\alpha = \frac{1}{2}$. This corresponds to the optimal Equal Error Rate on the development set. The performance on the evaluation set is measured by Half Total Error Rate (HTER), i.e.,

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2} \tag{14}$$

4. EXPERIMENTS

We have designed fusion experiments so as to study the relative merits of

- using the conventional fusion algorithm without quality, i.e., using the input x; features $[x, q]$, and the proposed quality-weighted features, i.e., $[x, x \otimes q]$ and $[x, q, x \otimes q]$.
- different types of fusion algorithms, i.e., discriminative or generative. For the discriminative classifier, SVM and logistic regression are used whereas for the generative classifier, GMM is used.
- different architecture types, i.e., (6), (9), and (10).

We will first report the baseline system performance and then the intramodal expert fusion performance. Then, we present and analyse the multimodal biometric expert fusion performance.

4.1. Baseline System Performance

The performance of the six face and one speech baseline systems are shown in Table 4(a). As the test data includes samples of good and degraded qualities, the overall performance is not very high. In fact one can glean the impact of the degraded quality test data on the overall performance from the last two columns of the table where the overall error is split into the contributions of the expert performance obtained on good and degraded quality data respectively. We can see, for instance, that the second best face expert (f5) on good quality data becomes the worst face expert on degraded data as well as the individually worst face expert on the mixed data.

4.2. Performance of Intramodal Quality-Dependent Fusion/ Score Normalisation

The results of the intramodal fusion of the six face systems are shown in Table 4(b). Since we have only one speech system, we effectively performed quality-dependent score normalisation, thus fusing x and q where x is a scalar value and q has two components given by SNR and the entropy quality measure. The results are shown in Table 4(c).

The results of intramodal fusion of face modalities are shown in the second section from the top in Table 4(b). As expected, conventional intramodal fusion (ignoring quality information) leads to a 30% improvement in performance over the individually best expert (expert f2). However, the use of quality information allows a further reduction in HTER to some 45% improvement over the performance of the best expert. Note in the table that the generative method of fusion using the Gaussian mixture model (GMM) begins to struggle with the extra

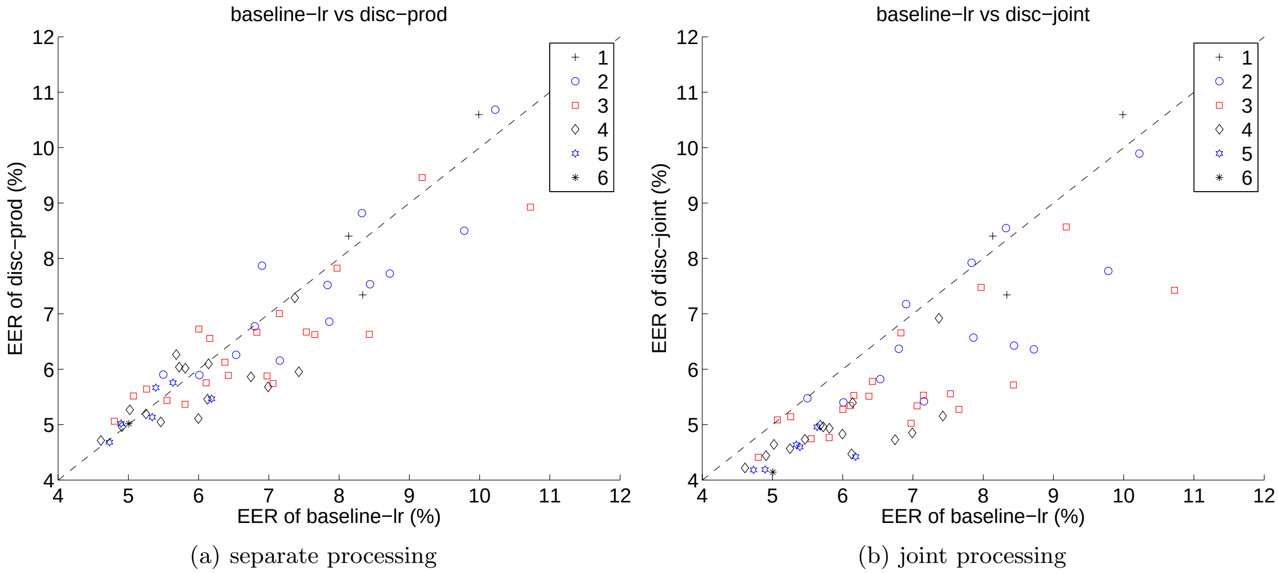
(a) Baseline systems				
		good + degraded	**good**	**degraded**
modality	no.	HTER (%)	HTER (%)	HTER (%)
face	1	11.06	6.66	13.50
face	2	7.67	3.48	9.78
face	3	8.29	5.86	9.57
face	4	10.39	2.13	17.17
face	5	24.56	2.97	39.28
face	6	16.96	5.51	23.42
speech	1	11.40	1.15	17.48

(b) Fusion of all six face systems				
arrangement	algo.	**good + degraded**	**good**	**degraded**
[$\mathbf{x}$]	lr	5.31	1.63	7.41
	svm	5.38	1.70	7.49
	gmm	5.61	2.22	7.61
[$\mathbf{x}$,$\mathbf{q}$]	lr	4.47	1.45	6.64
	svm	4.23	1.34	6.61
	gmm	5.28	1.63	7.70
[$\mathbf{x}$,$\mathbf{x}\otimes\mathbf{q}$]	lr	4.44	1.30	6.59
	svm	4.43	1.28	6.96
	gmm	12.44	3.56	18.04
[$\mathbf{x}$,$\mathbf{q}$,$\mathbf{x}\otimes\mathbf{q}$]	lr	4.52	1.51	6.38
	svm	4.23	1.34	6.26
	gmm	8.03	3.27	10.65

(c) Quality-dependent score normalisation of the speech system				
arrangement	algo.	**good + degraded**	**good**	**degraded**
[$\mathbf{x}$]	lr	11.44	1.15	17.54
	svm	11.40	1.15	17.48
	gmm	11.40	1.15	17.48
[$\mathbf{x}$,$\mathbf{q}$]	lr	11.22	1.07	17.37
	svm	11.59	1.19	17.88
	gmm	11.59	1.28	17.65
[$\mathbf{x}$,$\mathbf{x}\otimes\mathbf{q}$]	lr	11.35	1.14	17.41
	svm	11.35	1.14	17.42
	gmm	13.08	1.59	19.52
[$\mathbf{x}$,$\mathbf{q}$,$\mathbf{x}\otimes\mathbf{q}$]	lr	11.06	1.19	17.07
	svm	11.38	1.14	17.68
	gmm	10.77	1.06	16.79

Table 4. *A priori* HTER (%) of good + degraded test data, with the *a priori* HTER (%) of the good and degraded data sets recorded separately. The separate good and degraded data results were obtain by using the threshold (Δ) set on the good + degraded training data. (a) baseline systems of six face experts and one speech expert, (b) intramodal fusion of all six face experts using four different arrangements, and (b) quality-dependent score normalisation of the speech systems using four different arrangements.

degrees of freedom offered by the quality related features. When examined closely, we found that the distribution of the feature $x_i q_j$ for any system i and any quality measure j is not Gaussian. As a result, the generalisation performance based on GMM when using $x_i q_j$ is not expected to be better than the two discriminative classifiers. In fact, the score-quality product feature is designed for these classifiers and not for generative classifiers.



(a) separate processing

(b) joint processing

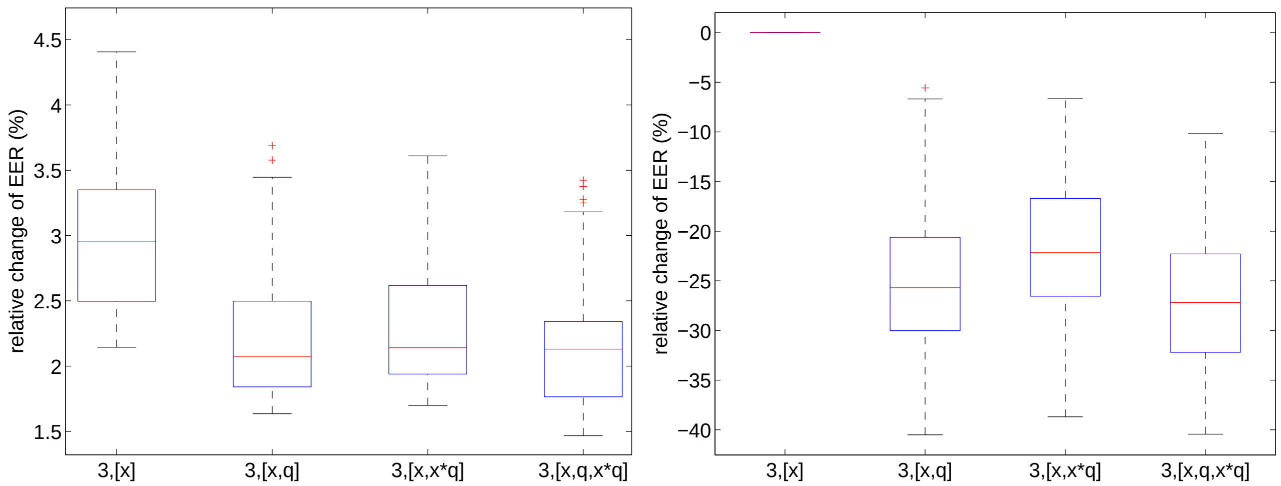
Figure 2. Comparison of (a) separate (Architecture 1) and (b) joint processing (Architecture 2 or 3 ; in the Y-axes) with respect to the baseline system (without using quality; in the X-axes). Each point in the figures are the *a posteriori* EER (%) of one of the possible 63 face and speech multimodal fusion tasks. In both figures, the numbers in the legend are the number of experts used in one of the 63 multimodal fusion tasks.

4.3. The Performance of Quality-Enhanced Multimodal Fusion

To demonstrate the advantage of the proposed feature space $[x, q, x \otimes q]$, we opted for one specific algorithm, i.e., logistic regression, but with different architectures (hence different dependency assumptions on the intramodal face fusion). We designed a set of $2^6 - 1 = 63$ intramodal fusion tasks corresponding to all the possible combinations of the six experts into distinct groups of increasing size, constituted by individual experts, all pairs, all triplets, etc. The two types of architectures considered are separate and joint processing. Note that in this experiment, (9) and (10) are the same. The results of using Architecture 1 and Architecture 2 (or Architecture 3) are shown in Figure 2. As can be observed, the proposed approach using $[x, q, x \otimes q]$ is almost always better than the baseline fusion approach using only $[x]$. We then repeated the same experiment except that the other two arrangements, i.e., $[x, q]$ and $[x, q, x \otimes q]$ were also used. The absolute and relative performance measures expressed in terms of *a posteriori* EER are shown in Figure 3. All three arrangements with quality measures show improvement over the baseline fusion $[x]$. The average observed relative improvement is about 25% but up to 40% can be attained. In particular, the arrangement $[x, q, x \otimes q]$ delivers the lowest absolute EER among all three arrangements considering the quality information.

The result for multimodal fusion are presented in Figure 4(b) for the different architectures. In Table 4, and Figures 3 and 4 we observe the tendency for the quality dependent fusion to yield better performance than the conventional fusion methods which ignore quality measurements. We also observe in Figure 4 that multimodal fusion performs better than intramodal fusion. Figure 4(b) also shows that the architectures can be listed in order of performance: Architecture 3, 2, and 1, with Architecture 3 being the most complex and Architecture 1 being the least complex.

In Architecture 1 all the expert outputs are effectively individually quality-normalised before being combined by averaging. When the quality measures are ignored, the architecture implements a fusion rule akin to the sum rule. Interestingly, the inclusion of the speech expert is beneficial. However, the results are not as good as those obtained by the quality free trained fusion methods realised by Architecture 3. For the logistic regression fusion the incorporation of the quality information in Architecture 1 did not enhance the fused system performance. The reason for that could be that combining quality measures with the scores of the respective experts individually, we create quality normalised scores that are correlated. Their capacity to enhance the performance by fusion is



(a) Absolute *a posteriori* EER (b) Relative *a posteriori* EER

Figure 3. Absolute and relative change of *a posteriori EER* (%) of Architecture 3 implemented using logistic regression for the four arrangements, evaluated on the good and degraded XM2VTS face database on modified Lausanne Protocol I. Each box plot contains 63 values corresponding to the 63 possible face and speech multimodal fusion tasks obtained by exhaustively matching all face and all speech systems in fusion. In arrangement $[x]$, the quality information is not used.

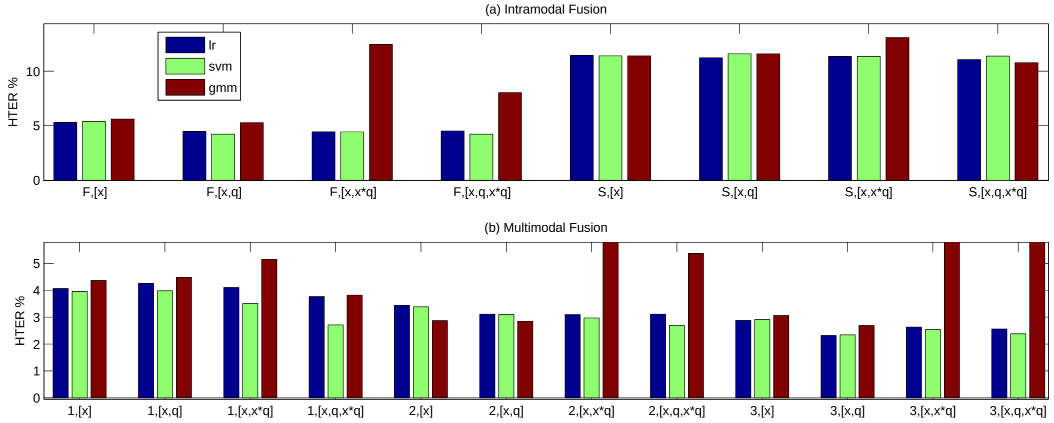


Figure 4. *A priori* HTER (%) of three types of classifiers (logistic regression, SVM and GMM), with four types of arrangements according to Table 1, evaluated on the good and degraded XM2VTS face images on: (a) Intramodal fusion on face and speech modality, (b) Multimodal fusion on three types of architectures. The bar plots in (a) are the good and dark HTER shown in Tables 4(b) and (c) for the face and the speech modalities, respectively.

then severely compromised. We can therefore conclude that, in all respects, Architecture 1 represents the least promising approach to fusion, whether quality information is used or not.

Architecture 3, on the other hand, appears to be consistently the best. The only exception is the GMM fusion classifier in $[x, x \otimes q]$ and $[x, q, x \otimes q]$. This supports the finding that in general, discriminative methods tend to be more stable than generative ones, especially when the number of degrees of freedom increases. Although we can get very good results by a trained fusion rule of multimodal experts even without using quality measures, we get up to 20% gain in performance for quality dependent fusion. In comparison with the individually best expert, the performance improves overall by a factor of more than three.

Figure 4 also shows that for $[x, x \otimes q]$ and $[x, q, x \otimes q]$ GMM performs better in Architecture one. The main reason for that is the low dimensionality of the feature spaces associated with the individual models of the architecture as compared with Architecture 2 and 3. The additional factor is that $x \otimes q$ is not Gaussian.

It is interesting that for Architecture 3, the performance of SVM and logistic regression methods is comparable, although the performance of SVM is slightly better due to the extra dimensionality. This is the consequence of having to make no simplifying assumptions regarding independence of experts. The system is easy to train and in the absence of quality measurements it achieves the best fusion results among all the fusion architectures. The only limitations of this approach are practical ones. First of all the suppliers of the individual biometric modalities in this case have to provide access to the quality information, and to training data. In this respect Architectures 1 and 2 have an immense advantage over Architecture 3.

The results of the experiments summarised in Figure 4 show that in the case of Architecture 3, for logistic regression, the arrangement $[x, q]$ works better than the other two arrangements which include quality, but overall when all possible fusion schemes are looked at, the $[x, q, x \otimes q]$ arrangement provides the best performance, albeit by a small margin. This is shown in Figure 3. Interestingly, the best fusion scheme does not require the inclusion of all six face experts, but only a subset of them.

5. CONCLUSIONS

We addressed the problem of score level fusion of intramodal and multimodal experts in the context of biometric identity verification. The focus was on confidence based fusion controlled by biometric data quality. We investigated the merit of using as features not only quality measures but also the cross terms obtained by taking the product of score and quality to generalise the fusion feature space. The study also explored several architectures that might be appropriate in different circumstances, namely when score and quality data for each expert and modality is made available to the fusion stage, and the situation where each modality delivers quality dependent scores for integration in the fusion system. We showed that the use of quality weighted scores as features in the definition of the fusion functions leads to improved performance. We also demonstrated that the achievable performance gain is also affected by the choice of fusion architecture. Whenever practicable, the best performance can be achieved when scores and quality features and the associated cross terms are considered jointly as a basis for decision making. However, when this design approach is not feasible, the conventional multiple classifier fusion architectures still offer considerable gain in performance when quality information is exploited.

The data available for experimentation somewhat limited the scientific scope of the investigation. Apart from improving the overall performance of the biometric system, the use of multimodal biometrics is motivated by the expectation that it makes the system more robust, as in principle one can switch between individual biometric modalities or dynamically control their influence as a function of their quality. Unfortunately, this scientific hypothesis could not be tested, as the degradation of quality of the two modalities used (face and speech) was strongly correlated. We had no evaluation data where the quality of one of the modalities was high while that of the other quality was low. However, such data is potentially available in the Banca database and we intend to study this particular aspect in the future.

ACKNOWLEDGEMENTS

This work was partially supported by the prospective researcher fellowship PBEL2-114330 of the Swiss National Science Foundation, by the BioSecure project (www.biosecure.info) and by EPSRC Research Grant GR/S46543. The authors thank Guillaume Heusch at IDIAP for having provided the scores of the face systems. This publication only reflects the authors' view.

REFERENCES

1. T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," in *IEE Trans. Pattern Anal. Machine Intell*, **16**, pp. 66–75, 1994.
2. J. Kittler, A. Ahmadyfard, and D. Windridge, "Serial multiple classifier systems exploiting a coarse to fine output coding," in *Multiple Classifier Systems*, pp. 106–114, 2003.
3. J. Kittler, R. P. W. D. M Hatef, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, pp. 226–239, Mar 1998.

4. P. Norman and B. Samy, "A score-level fusion benchmark database for biometric authentication," in *Audio and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science* **3546**, pp. 1059–1070, 2005.
5. E. Tabassi, C. Wilson, and C. Watson, "Fingerprint image quality: Nistir 7151," tech. rep., NIST, 2004.
6. J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Kernel-Based Multimodal Biometric Verification Using Quality Signals," in *Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE,* **5404**, pp. 544–554, 2004.
7. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communnications," in *12th Int'l Conf. on Image Analysis and Processing*, pp. 2–11, (Mantova), 2003.
8. K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo, "Error Handling in Multimodal Biometric Systems using Reliability Measures," in *Proc. 12th European Conference on Signal Processing*, (Antalya, Turkey), September 2005.
9. N. Karthil, C. Yi, and K. J. Anil, "Quality-based score level fusion in multibiometric systems," in *Internation conference on Pattern Recognition*, pp. 1059–1070, 2006.
10. K. Kryszczuk and A. Drygajlo, "On combining evidence for reliability estimation in face verification," in *Proc. 13th European Conference on Signal Processing*, (Florence, Italy), 2006.
11. J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia, and A. K. Jain, "Incorporating image quality in multi-algorithm fingerprint verification," in *ICB*, 2006.
12. P. Norman and B. Samy, "Improving fusion with margin-derived confidence in biometric authentication task," in *Audio and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science* **3546**, pp. 474–483, 2005.
13. K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo, "Error Handling in Multimodal Biometric Systems using Reliability Measures," in *Proc. 12th European Conference on Signal Processing*, (Antalya, Turkey), September 2005.
14. C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1999.
15. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.
16. V. N. Vapnik, *Statistical Learning Theory*, Springer, 1998.
17. S. Pigeon, P. Druyts, and P. Verlinde, "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions," *Digital Signal Processing* **10**(1–3), pp. 237–248, 2000.
18. A. Jain, K. Nandakumar, and A. Ross, "Score Normalisation in Multimodal Biometric Systems," *Pattern Recognition* **38**(12), pp. 2270–2285, 2005.
19. K.-A. Toh, W.-Y. Yau, E. Lim, L. Chen, and C.-H. Ng., "Fusion of Auxiliary Information for Multimodal Biometric Authentication," in *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pp. 678–685, (Hong Kong), 2004.
20. J. Matas, M. Hamouz, K.Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. .Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of face verification results on xm2vts database," in *Proc. 15th int'l Conf. Pattern Recognition*, pp. 858–863, (Barcelona), 2000.
21. K. Messer, J. Kittler, J. Short, G. Heusch, F. Cardinaux, S. M. anf Y. Rodriguez, S. Shan, Y. Su, and W. Gao, "Performance characterisation of face recognition algorithms and their sensitivity to severe illumination changes," in *LNCS 3832, Proc Int'l Conf. Biometrics*, pp. 1–11, (Hond Kong), 2006.
22. G. Heusch, Y. Rodriguez, and S. Marcel, "Local binary pattern as an image preprocessing face authentication," in *Proc. 7th Int'l Conf. Automatic Face and Gesture Recognition (FGR06)*, pp. 9–14, (Washington, DC), 2006.
23. J. Kittler, Y. Li, and J. Matas, "On matching score for lda-based face verification," in *British Machine Vision Conference (BMVC)*, 2000.
24. D. A. Reynolds, T. Quatieri, and T. Dunn, "Speaker verification using adapted guassian mixture models," in *Digital Signal Processing*, pp. 19–41, 2000.
25. F. Cardinaux, C. Sanderson, and S. Bengio, "User authentication via adapted statistical models of face images," in *IEEE Trans. on Signal Processing*, pp. 361–373, January 2006.

26. R. Gross and V. Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *4th Int'l Conf. Audio and Video-Based Biometric Person Authentication (AVBPA'03)*, pp. 10–18, 2003.

27. D. A. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification.* PhD thesis, Georgia Institute of Technoloy, Atlanta, USA, 1992.

28. P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proc. Eurospeech*, 2001.

29. J. Richiardi, P. Prodanov, and A. Drygajlo, "Speaker verification with confidence and reliability measures," in *Proc. 2006 IEEE International Conference on Speech, Acoustics and Signal Processing*, (Toulouse, France), May 2006.

30. J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. 2005 IEEE International Conference on Speech, Acoustics and Signal Processing*, pp. 73–740, (Philadelphia), 2005.