IDIAP RESEARCH REPORT

Variance Reduction Techniques in Biometric Authentication

Norman Poh Hoon Thian [a] Samy Bengio [a]

IDIAP–RR 03-17

March 2003

SUBMITTED FOR PUBLICATION

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, CP 592, 1920 Martigny, Switzerland

IDIAP Research Report 03-17

Variance Reduction Techniques in Biometric Authentication

Norman Poh Hoon Thian   Samy Bengio

March 2003

submitted for publication

Abstract. In this paper, several approaches that can be used to improve biometric authentication applications are proposed. The idea is inspired by the ensemble approach, i.e., the use of several classifiers to solve a problem. Compared to using only one classifier, the ensemble of classifiers has the advantage of reducing the overall variance of the system. Instead of using multiple classifiers, we propose here to examine other possible means of variance reduction (VR), namely through the use of multiple real samples, synthetic samples, different extractors (features) and biometric modalities. It is found empirically that VR via modalities is the best technique, followed by VR via real samples, VR via extractors, VR via classifiers and VR via synthetic samples. This order of effectiveness is due to the corresponding degree of independence of the combined objects (in decreasing order). The theoretical and empirical findings show that the combined experts via VR techniques *always* perform better than the average of their participating experts. Furthermore, in practice, *most* combined experts perform better than any of their participating experts.

1 Introduction

Biometric authentication (BA) is the problem of verifying an identity claim using a person's behavioural and physiological characteristics. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement, forgetfulness or reproduction. Examples of biometric modalities are fingerprint, face, voice, hand-geometry and retina scans [1].

However, to date, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases. The focus of this study is to improve the system accuracy by directly minimising the noise via various variance reduction techniques.

Biometric data is often noisy because of deformable templates, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes.

Advancements in biometrics show two emerging solutions: combining several biometric modalities [2–4] (often called multi-modal biometrics) and combining several samples of a single biometric modality [5]. These techniques are related to variance reduction, in that, they exploit the assumed independence of each modality (i.e., independent samples or biometric modality). In this work, we examine several other ways to exploit this (partial) independence, namely via extractors and synthetic samples. In short, all these methods can be termed as follows: Variance Reduction (VR) via classifiers, VR via extractors, VR via samples and VR via different biometric modalities.

To our opinion, VR techniques are potential to improve the accuracy of BA systems because better classifiers or ensemble methods, feature extraction algorithms and biometric-enabled sensors are emerging. VR techniques can be used to combine these new algorithms with existing algorithms to obtain improved results. The added overhead cost will be computation time and possibly hardware cost in the case of adding new sensors (as opposed to other VR techniques which do not require any extra hardware).

For the sake of clarity, we decided to present the concept of variance reduction in Section 2 that follows immediately after this introduction because it is the principal motivation of improving the BA system in this paper. This is followed by Section 3 which presents the general framework of a BA system. Section 4 then explores different VR techniques in the context of BA and also discusses existing works in the literature. Section 5 explains the databases used and it is followed by experimental results in Section 6 and conclusions in Section 7.

2 Variance Reduction: A Theoretical Explanation

This section presents a theoretical justification of variance reduction[1]. A person requesting an access can be measured by his or her biometric data. Let this biometric data be $\mathbf{x}$. This measurement can be done in several methods, to be explored later. Let i denote the i-th extract of $\mathbf{x}$ by a given method. For the sake of comprehension, one method to do so is to use multiple samples. Thus, in this case, i denotes the i-th sample. If the chosen method uses multiple biometric modalities, then i refers to the i-th biometric modality. Let the measured relationship be denoted as $y_i(\mathbf{x})$. It can be thought as the i-th response (of the sample or modality, for instance) given by a biometric system. Typically, this output (e.g. score) is used to make the accept/reject decision. An explanation of the BA system will follow in Section 3. $y_i(\mathbf{x})$ can be decomposed into two components, as follows:

$$y_i(\mathbf{x}) = h(\mathbf{x}) + \eta_i(\mathbf{x}), \tag{1}$$

[1] A similar explanation of this section can be found in [6, Chap. 9], where variance reduction is due to averaging classifier scores. In this section, we have decided to discuss variance reduction due to different samples for easy comprehension. This concept will be generalised to VR via modalities, features and synthetic features as well. Also, note that even though the discussion here is related to regression problems, it can be extended to classification problems, such as BA.

where $h(\mathbf{x})$ is the "target" function that one wishes to estimate and $\eta_i(\mathbf{x})$ is a random additive noise with zero mean, also dependent on $\mathbf{x}$. $h(\mathbf{x})$ can be viewed as the ideal function that consistently gives 1 when $\mathbf{x}$ corresponds to the client and -1 when corresponds to the impostor.

In other words, the score $y_i(\mathbf{x})$ can be viewed as a random variable translated by its mean value $h(\mathbf{x})$ and distorted by $\eta_i(\mathbf{x})$, which in turns, can also be viewed as a random variable with zero mean. It can then be easily observed that the expected value of $y_i(\mathbf{x})$ is $h(\mathbf{x})$.

For the moment, let N be the number of trials, (e.g., the number of samples, assuming that the chosen method uses multiple samples hereinafter). The mean of y over N trials, denoted as $\bar{y}(\mathbf{x})$ is:

$$\bar{y}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} y_i(\mathbf{x}). \tag{2}$$

With enough samples, the expected value of $y_i(\mathbf{x})$, denoted as $E[y_i(\mathbf{x})]$, which approximates the "true" measure, can be written as:

$$\begin{aligned} E[y_i(\mathbf{x})] &= E[h(\mathbf{x})] + E[\eta_i(\mathbf{x})] \\ &= h(\mathbf{x}). \end{aligned} \tag{3}$$

By using one sample for an access, the variance, by definition, is:

$$\begin{aligned} \mathrm{VAR}[y_i(\mathbf{x})] &= E[(y_i(\mathbf{x}) - E[y_i(\mathbf{x})])^2] \\ &= E[(y_i(\mathbf{x}) - h(\mathbf{x})])^2] \\ &= E[\eta_i(\mathbf{x})^2], \end{aligned} \tag{4}$$

where we made use of Equation 3 and Equation 1.

When N samples are available and they are used separately, the *average of variance* made by each sample, independently, is:

$$\begin{aligned} \mathrm{VAR}_{AV}(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{VAR}[y_i(\mathbf{x})] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}[\eta_i(\mathbf{x})^2], \end{aligned} \tag{5}$$

where Equation 4 is used.

However, by combining N samples for an access via averaging, the *variance of average* is:

$$\begin{aligned} \mathrm{VAR}_{COM}(\mathbf{x}) &= E[(\bar{y}(\mathbf{x}) - h(\mathbf{x})])^2] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_i(\mathbf{x}) - h(\mathbf{x})\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} y_i(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{x})\right)^2\right] \\ &= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} \eta_i(\mathbf{x})\right)^2\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N} E\left[\eta_i(\mathbf{x})^2\right] \end{aligned} \tag{6}$$

Equation 6 assumes that all $\eta_i(\mathbf{x})$ has zero mean in general and are uncorrelated, i.e., $E[\eta_i(\mathbf{x})] = 0$ and $E[\eta_i(\mathbf{x})\eta_j(\mathbf{x})] = 0$ for $\forall_{i,j} i \neq j$. In reality, the $\eta_i(\mathbf{x})$ are in general correlated. To compensate these false assumptions, it is necessary to introduce an inequality into Equation 6. Doing so and making use of Equation 5, Equation 6 becomes:

$$\begin{aligned}\mathrm{VAR}_{COM}(\mathbf{x}) &\leq \frac{1}{N^2}\sum_{i=1}^{N} E\left[\eta_i(\mathbf{x})^2\right] \\ &\leq \frac{1}{N}\mathrm{VAR}_{AV}(\mathbf{x}),\end{aligned} \tag{7}$$

where we have written $\mathrm{VAR}_{COM}(\mathbf{x})$ (the variance of average) in terms of $\mathrm{VAR}_{AV}(\mathbf{x})$ (the average of variance).

Basically, Equation 7 shows that by averaging N scores, the average of variance can be reduced by a factor of N (when the assumption is true) or less (when the assumption is violated) with respect to the variance of average.

From Equation 7, one can deduce that $\mathrm{VAR}_{COM}(\mathbf{x}) \leq \mathrm{VAR}_{AV}(\mathbf{x})$ in all situations. To measure *explicitly* the factor of reduction, we introduce α, which can be defined as follows:

$$\alpha = \frac{\mathrm{VAR}_{AV}(\mathbf{x})}{\mathrm{VAR}_{COM}(\mathbf{x})}. \tag{8}$$

Since $\mathrm{VAR}_{COM}(\mathbf{x}) \leq \mathrm{VAR}_{AV}(\mathbf{x})$ and $\mathrm{VAR}[\cdot] \geq 0$, one can conclude that α must be greater than one. Another logical explanation for this observation is as follows: if the error made by each score is correlated, i.e., it makes exactly the same error in the extreme case $\forall_{i,j}, y_i(\mathbf{x}) = y_j(\mathbf{x})$, then $\frac{1}{N}\sum_{i=1}^{N} y_i(\mathbf{x}) = \bar{y}(\mathbf{x})$. As a consequence, $\mathrm{VAR}[y_i(\mathbf{x})] = \mathrm{VAR}_{AV}(\mathbf{x}) = \mathrm{VAR}_{COM}(\mathbf{x})$, which again implies that $\alpha = 1$. Therefore, in averaging scores, if one does not gain, *one does not lose* in the combination neither compared to the *average* performance of using N trials. Furthermore, from Equation 7, it can also be concluded that α is smaller than or equal to N. This leads to the conclusion that $1 \leq \alpha \leq N$.

Ideally, it would have been interesting to show that $\mathrm{VAR}_{COM}(\mathbf{x}) \leq \mathrm{VAR}[y_i(\mathbf{x})]$, for any i. Unfortunately, it is not the case here. In practice, however, as will be shown in Section 6, this *can* happen.

Figure 1 illustrates the effect of averaging scores in a two-class problem, such as in BA where an identity claim could belong either to a client or an impostor. Let us assume that the genuine user scores in a situation where 3 samples are available but are used separately, follow a normal distribution of mean 1.0 and variance ($\mathrm{VAR}_{AV}(\mathbf{x})$ of genuine users) 0.9, denoted as $\mathcal{N}(1, \sqrt{0.9})$, and that the impostor scores (in the mentioned situation) follow a normal distribution of $\mathcal{N}(-1, \sqrt{0.6})$ (both graphs are plotted with "+"). If for each access, the 3 scores are used, according to Equation 8, the variance of the resulting distribution will be reduced by a factor (which is the value α defined in Equation 8) of 3 or less. This reduction factor is derived from Equation 7. Both resulting distributions are plotted with "o". Note the area where both the distributions cross before and after. The later area is shaded in Figure 1. This area corresponds to the zone where minimum amount of mistakes will be committed given that the threshold is optimal[2]. Decreasing this area implies an improvement in the performance of the system.

3 BA: Framework and Performance Evaluation

The discussion in Section II assumes that N ways are available. In particular, N are the number of samples or modalities. In either case, this is often not the situation. In fact, N could be the number of different classifiers, different features, synthetic samples or different biometric modalities. Furthermore, $y_i(\mathbf{x})$ is also vaguely described as a score emitted by a "BA system". This sections discusses the generic framework (i.e., software architecture) of a BA system, and its performance measurement.

[2] Optimal in the Bayes sense, when (1) the cost and (2) probability of both types of errors are equal.

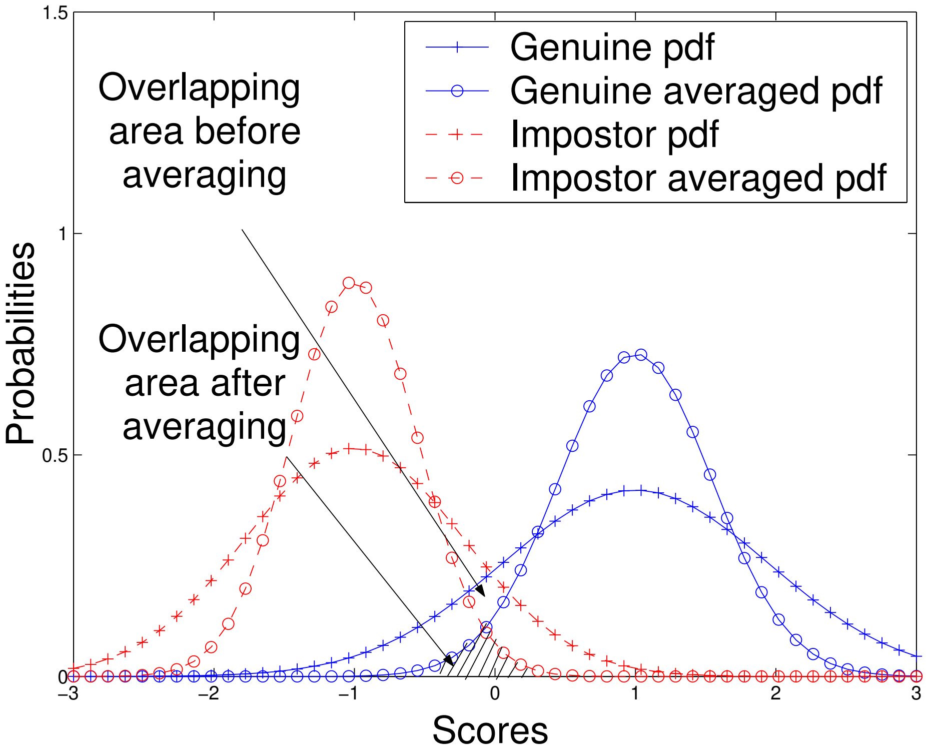


Figure 1: Averaging score distributions in a two-class problem

3.1 A Generic BA Framework

The fundamental problem of BA can be viewed as a binary classification task given a biometric data. One can model such a classification task as a function f in $y = f(\mathbf{x})$, where $\mathbf{x}$ is a given biometric data, f is the BA system and y is a similarity measurement (e.g., output of a neural network, distance measure from a given template)[3]. If y is greater than a pre-defined threshold, the biometric data is accepted as belonging to the claimed identity. Otherwise, it is rejected. In this way, BA is a task of deciding if a biometric data belongs to a claimed identity or not. Improving a BA system in this sense is to minimise the number of errors committed by the system. The lower this number is, the better the system performs.

In this section, a biometric-independent framework (see Figure 2) is proposed. This framework consists of a serial concatenation of sensors, extractors, classifiers and supervisors. First, a user's biometric data is captured using sensors. Examples of sensors are Charged Couple Device (CCD) cameras, Infrared-Red (IR) cameras, fingerprint scanners and microphones. Each sensor has its own standard data representation. A set of operations, often based on signal- and image-processing algorithms, constitute the building blocks of extractors. Extractors have two functions: to detect and to extract user-discriminant information. Each extractor produces its own type of vectors or feature vectors, also called templates in a more generic setting. Experts or classifiers are used to categorise these produced vectors. Classifiers are a set of pattern-matching algorithms, which may be learning-based (e.g. Multi-Layer Perceptron, Support Vector Machine, etc) or template-based (dynamic time warping, Euclidean distance, normalised correlation, etc). Classifiers' role is to map a vector to an associated identity. They do so with a certain degree of confidence commonly called a score or a

[3] $y \in \mathbb{R}^+$ is often called a distance; $y \in [0, 1]$ is often viewed as a probability; and $y \in \mathbb{R}$ is called a score.

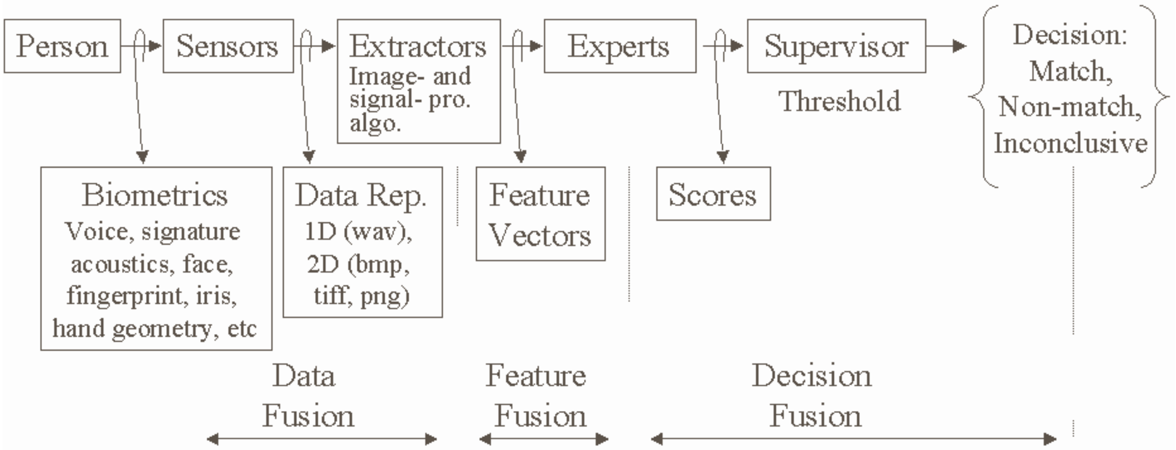


Figure 2: A generic biometric taxonomy and its fusion scheme

confidence measure. It could be a scalar value or a vector when more information is supplied. In some cases, a score can be interpreted as the estimated *a posteriori* probability of the claimed class label given the feature. When there are several classifiers, a supervisor merges different scores to obtain the final decision. To make the final decision, a score is compared with a pre-defined threshold. If the final decision is a match, then the system accepts the identity claim. If the decision is a non-match, then the system rejects the identity claim. Finally and optionally, if the decision is inconclusive, a fall-back procedure should be activated.

The serial concatenation process of sensors, extractors, classifiers and supervisors shows that error may accumulate along the chain because each module depends on its previous module. In a separate study done by Jain and Pankanti [7], they used the terms *information limited behaviour*, *representation limited behaviour* and *invariance limited behaviour* to describe the errors of the first three components (sensors, extractors and classifiers). Note that the supervisor itself can also introduce errors. To our opinion, the term “limitation” is easier to be understood as errors due to sensors, extractors and classifiers, respectively. Unfortunately, in practice, such errors cannot be easily measured. However, one knows that if one can improve any one of the components in this serial process, one can improve the whole biometric system.

3.2 Performance measurement in BA

There are two types of errors in BA: false acceptance (FA) and false rejection (FR). Given a pre-defined threshold, if a score belongs to a impostor but is greater than or equal to the threshold, the system will wrongly accept the impostor's claim. In this case, the system is said to commit a FA. On the other hand, if a score belongs to a client but is smaller than the pre-defined threshold, the system will wrongly reject the client's claim. In this case, the system is said to commit a FR. The error rates of FA and FR are commonly called false acceptance rate (FAR) and false rejection rate (FRR), which are defined as:

$$\mathrm{FAR} = \frac{\text{number of FAs}}{\text{number of impostor accesses}} \times 100\%, \tag{9}$$

$$\mathrm{FRR} = \frac{\text{number of FRs}}{\text{number of client accesses}} \times 100\%, \tag{10}$$

respectively. Note that FAR and FRR change according to the pre-defined threshold. Another commonly used measurement is called the Half Total Error Rate (HTER) that can be defined as:

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2}. \tag{11}$$

Note that HTER assumes the class prior probabilities of clients and impostors to be equal, i.e., 0.5 and the cost of FR and FA to be equal to 1. By varying the decision threshold, one obtains different values of FAR and FRR. This is often plotted on a curve [8] (such as ROC or DET curves, not shown here).

3.3 Threshold Selection

There are several important concepts associated with the way one measures the performance of BA systems. The first concept is about error measurement. HTER is an example of such measures. Other measures could take into account prior information specific to the application.

Since most of these measures are based on the selection of a threshold, the second concept regards how this threshold is selected. If it is selected on the same data set as the one used to compute the performance (we say that it is an *a posteriori* selected threshold), the performance will appear optimistically biased. On the other hand, if the threshold is selected on a separate data set (we say that it is an *a priori* selected threshold), then the performance will be more realistic.

The third concept concerns the criterion used to select the threshold. Various criteria could be used and they should reflect the prior information specific to each application. The most used criterion is the so-called *Equal Error Rate* (EER) which corresponds to FAR = FRR. In practice, one searches the threshold where FAR is closest to FRR by minimising $|\mathrm{FAR} - \mathrm{FRR}|$ [4].

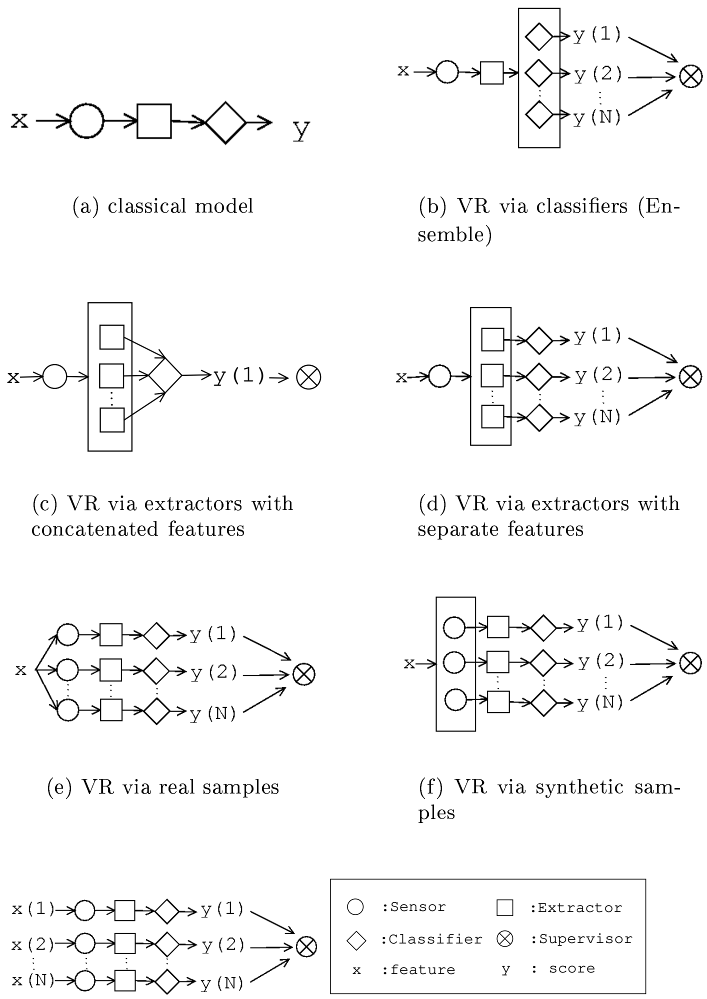
4 Exploring Various Variance Reduction Techniques

This section explores various variance reduction techniques that can be applied to the BA problem. Figure 3 shows several possible approaches to improve a biometric authentication system. Figure 3(a) is the usual mono-modal biometric approach. One can improve the system by using multiple classifiers (3b) which can also be called the ensemble method, multiple extractors with concatenated features (3c), multiple extractors with separate features (3d), multiple real samples (3e), multiple synthetic samples (3f), and multiple biometric modalities (3g). Each of the approaches are explained in more details in this section. Exploiting such parallel structure is the key approach towards variance reduction and thus can increase the accuracy of the system.

VR via classifiers is a kind of ensemble method. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of each classifier prediction. The main idea is that ensembles are often much more accurate than the individual classifiers that make them up, provided that the individual classifiers are accurate and diverse (which includes complementary).

Dietterich [9] groups ensemble methods into: (i) Bayesian voting, (ii) manipulation of the training examples (e.g. bagging [10,11], cross-validated committee and boosting [12]), (iii) manipulation of the input features (e.g. sub-tasking), (iv) manipulation of the output targets (e.g. ECOC [13]) and (v) injection of randomness, also known as learning with noise. Biometric features are very susceptible to noise and different deformation. Therefore, these techniques are important considerations in our framework. Kittler *et al* [14] have convincingly shown that a modified version of ECOC called multi-seed ECOC improves face recognition on the XM2VTS database. These are all known methods to improve BA systems.

[4]The experiments carried out in this paper used the "plotdet" software that can be downloaded from http://www.idiap.ch/~marietho.



(a) classical model

(b) VR via classifiers (Ensemble)

(c) VR via extractors with concatenated features

(d) VR via extractors with separate features

(e) VR via real samples

(f) VR via synthetic samples

(g) VR via modalities

(h) legend

Figure 3: Different possible VR techniques in BA

The main idea of **VR via extractors** is that given a raw biometric data, several features are extracted. For example, one can extract the following information from speech features: Linear Predictive Coding Coefficients or Mel Frequency Cepstrum Coefficients [15]. In face verification, common features are principal components, linear discriminant components [16] or more recently independent components [17]. Each feature is often classified by an associated classifier. Since these features are different, one can expect the corresponding trained classifiers to commit different errors. On the often false assumption that features are not correlated, the classifiers are therefore not correlated. In the hope that each classifier operating on different feature space makes different errors, the combined classifier should be able to reduce the errors.

There are basically two variations of VR via extractors: with concatenated features (see Figure 3c) and separate features (3d). In the first case, the extracted features are normalised to the same range, concatenated and fed to a common classifier for training and matching. Often the curse of dimensionality [6] is an obstacle to this approach. In the second case, each feature set is treated separately by its own classifier. A decision fusion scheme is required to merge the scores coming from these classifiers.

These techniques work because different features usually capture different or complementary information. Since they are extracted from the same sample, they are probably dependent (on the sample). Regardless of the degree of dependency, VR via extractors can yield improvement in accuracy due to reasons justified earlier.

In the work of Brunelli and Falavigna [18], two speech experts (using respectively static and temporal derivative features) and three face experts (using respectively eye, nose and mouth areas of the face) are used for person verification. The weighted product approach was used to fuse the opinions, with the weights found automatically via a heuristic approach. The static and dynamic feature experts obtained an identification rate of 77% and 71%, respectively. Combining the two speech experts increased the identification rate to 88%. The eye, nose and mouth experts obtained an identification rate of 80%, 77% and 83% respectively. Combining the three facial experts increased the identification rate to 91%. The work shows that **VR via extractors with separate features** improve the accuracy of a BA system.

For the problem of face verification, Marcel and Bengio [19] have shown that instead of using just face images, one can use normalised face colour histogram as an additional feature to the existing normalised face image to train client specific classifiers. This yields an improved classification result.

Luettin [20] investigated the combination of speech and (visual) lip information using feature vector concatenation. In order to match the frame rates of both feature sets, speech information was extracted at 30 fps (frames per second) instead of the usual 100 fps. In text-dependent configuration, the fusion process resulted in a minor performance improvement. However, in text-independent configuration, the performance slightly decreased. This could probably be due to the curse of dimensionality explained earlier. These works [19,20] showed that **VR via extractors with concatenated features** *may* also improve the accuracy of a BA system.

In this report, we decided to use VR via extractors with separate features for the following reasons: it is better understood; its use has already been justified in Section II; and it is often computationally less expensive compared to its counterpart with concatenated features.

VR via real samples has been demonstrated by Kittler *et al* [5]. In their work, they combined multiple snapshots of a single biometric property using a Bayesian framework. It is observed that as more and more samples are used, the classification error decreases until a point where it is "saturated", i.e., further increase of samples will not decrease the classification error further.

We have shown that **VR via synthetic samples** [21] is also a viable solution when real samples are not available due to some reasons. For instance, the data transfer bandwidth is limited or taking several biometric samples are inconvenient. This approach works only if such transformation can be found. For face images, geometric transformations can be readily applied without the loss of information. Other image-to-image transformations, i.e., quotient image and methods based on the symmetric property of faces can also be used to normalise the face image against lighting variations. The only constraint is that such transformation itself must not require many training data.

Several studies have shown that **VR via different modalities** is superior, on average, to any single-modal biometrics. The following are some strategies proposed in the literature:

- Jain *et al* [22] have proposed a multi-source biometric system design that integrates face and fingerprints to make a personal identification.

- Ross *et al* [2] have used hand-scan, fingerprint and face-scan to improve the overall result.

- Poh [23] used eye features and voice features extracted via wavelets to verify a person's identity. Both the face and voice experts are combined using the AND operation. Experiments showed that combining both experts improved the accuracy of the system.

- Dieckmann *et al* [24] used three experts (frontal face expert, dynamic lip image expert and text-dependent speech expert). A hybrid fusion scheme involving majority voting and opinion fusion was utilised. Two of the experts had to agree on the decision and the combined opinion had to exceed a pre-defined threshold. The hybrid fusion scheme provided better performance than using the underlying experts alone.

- Jourlin *et al* [25] used a form of weighted summation fusion to combine the opinions of two experts: a text-dependent speech expert and a text-dependent lip expert. It was shown that fusion led to better performance than using the underlying experts alone.

- Sanderson [26] used face and noisy speech information and combined both modalities using adaptive weights and various fusion methods. The resultant system provides a good trade-off in both clean and noise conditions.

The above list is certainly not exhaustive. Most multi-modality approaches yield improvement in results. This is a common and promising approach to improve a BA system. According to the VR justification presented in Section II, different biometric modalities provide nearly uncorrelated biometric data comparing to other VR techniques. Indeed, empirical results later in this study does support this argument.

5 Databases and Feature Extraction

There are already many works done based on different databases. In this work, the VR techniques are applied on the XM2VTS and LSIIT databases. The XM2VTS database [27] provides a real world data corpus with large amount of data (200 clients). LSIIT database [23] is used to complement what XM2VTS is lacking: providing a small database (30 clients), with multiple audio-video samples for each access while allowing multiple runs of experiments in a reasonable amount of time.

5.1 XM2VTS database

5.1.1 Description of the database

The XM2VTS database [27] contains synchronised video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the pronunciation of a sentence.

The video was captured at a colour sampling resolution of 4:2:0 with 16 bit audio at a frequency of 32 kHz. The video data was compressed at a fixed ratio of 5:1 in the proprietary DV format. When capturing the database the camera settings were kept constant across all four sessions. The head was illuminated from both left and right sides with diffusion gel sheets being used to keep this illumination as uniform as possible. A blue background was used to allow the head to be easily segmented out.

The database is divided into three sets: a training set, an evaluation set and a test set. The training set was used to build client models, while the evaluation set (Eval) was used to compute the decision thresholds (as well as other hyper-parameters) used by classifiers. Finally, the test set (Test) was used to estimate the performance.

The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning of the training and evaluation sets. They are called Lausanne Protocol I and II, denoted as **LP1** and **LP2** in this paper. Thus, besides the data for training the model, the following four data sets are available for evaluating the performance: **LP1 Eval**, **LP1 Test**, **LP2 Eval** and **LP2 Test**. Note that LP1 Eval and LP2 Eval are used to calculate the optimal thresholds that will be used in LP1 Test and LP2 Test, respectively. Results are reported only for the test sets, in order to be as unbiased as possible (using an *a priori* selected threshold). Table 1 is the summary of the data. In both configurations, the test set remains the same. However, there are three training data per client for LP1 and four training data per client for LP2. More details can be found in [28].

Table 1: The Lausanne Protocols of XM2VTS database

Data sets	Lausanne Protocols	
	LP1	LP2
Training client accesses	3	4
Evaluation client accesses	600 (3×200)	400 (2×200)
Evaluation impostor accesses	40,000 ($25 \times 8 \times 200$)	
Test client accesses	400 (2×200)	
Test impostor accesses	112,000 ($70 \times 8 \times 200$)	

5.1.2 Feature Used for the XM2VTS Database

For the face data, a bounding box is placed on a face according to eyes coordinates located manually. This assumes a perfect face detection[5]. The face is cropped and the extracted sub-image is down-sized to a 30×40 image. After enhancement and smoothing, the face image has a feature vector of dimension 1200.

In addition to these normalised features, RGB (Red-Green-Blue) histogram features are used. To construct this additional feature set, a skin colour look-up table must first be constructed using a large number of colour images which contain only skin. In the second step, face images are filtered according to this look-up table. Unavoidably, non-skin pixels are captured as well. The resultant features will be submitted to a classifier to discriminate its degree of relevance. For each colour channel, a histogram is built using 32 discrete bins. Hence, the histograms of three channels, when concatenated, form a feature vector of 96 elements. More details about this method, including experiments, can be obtained from [19].

Another feature set derived from Discrete Cosine Transform (DCT) coefficients [29, 30] has also given good performance. The idea is to divide images into overlapping blocks. For each block, a subset of DCT coefficients are computed. The horizontal, vertical and diagonal (with respect to a reference block of) DCT coefficients can also be derived. It has been shown that these features are comparable (in terms of performance in the context of BA) to features derived from Principal Component Analysis [29].

For the speech data, the feature set used in the experiments are Linear Filter-bank Cepstral Coefficients (LFCC) [15]. These features are obtained from DCT transformation of Short Term Fourier Transform coefficients for each frame of a fixed size window. It is often necessary to remove non-speech frames [15], because they do not contain any discriminative or useful information for the purpose of BA. This problem is called segmentation of speech/non-speech. This segmentation is done in this case using two competing Gaussians trained in an unsupervised way by maximising the likelihood of the data given a mixture of the 2 Gaussians. One Gaussian will end up modelling the speech and the other will end up modelling the non-speech feature frames [31]. In general, the segmentation given by this technique is satisfactory.

5.2 LSIIT database

The LSIIT database[6] contains face images and voice passwords of 30 clients. It is used to simulate a BA scenario of a multi-modal biometric system in a secured workplace with a small population of users.

A generic PC web cam is used for sampling a 320×240 RGB image. Within this area of viewing, a face image is cropped out to the dimension of 150×225. The cropped out image is saved in the

[5]Hence, even if this is often done in the literature, the final results using face scores could be optimistically biased due to this manual detection step. Note on the other hand that due to the clean and controlled quality of XM2VTS, automatic detectors often yield detection rates around 99%.

[6]The LSIIT database can be obtained from http://hydria.u-strasbg.fr/~norman/BAS.

24-bit bitmap (BMP) format. When taking the photo, the person is requested to move his face into the area of interest where the cropped face image is expected. The recorded image contains an upright frontal image.

Each voice password is sampled for a duration of three seconds at 8kHz on a mono-channel microphone. The data is saved in a wave (WAV) file format of approximately 24K bytes. The password of each client could be any short word such as his or her name. The voice-scan is taken in the laboratory environment to model a typical indoor environment. No effort is made to make the problem more challenging or particularly easier.

Contrary to the XM2VTS database, there are only 30 clients. Furthermore, there is only one session. However, with this session, 10 face images and 10 recordings of speech were recorded. There are only two data sets: training and test sets. 5 samples are randomly selected for training and the remaining 5 samples are used for testing. Note that there is no data set reserved for tuning the threshold parameters as in the XM2VTS database. Due to the lack of the third data set, the HTERs obtained on this test data set are *a posteriori*. Therefore, the HTERs on both of the databases are not comparable.

For the face verification task, Principal Component Analysis (PCA) [32] and Linear Discriminant Analysis (LDA) [33] features have been used for this database. LDA and PCA are linear transformations which, in our case, reduced the face image from 150×225 pixels to 50 and 144 real values, respectively. The number of components in the PCA is determined such that the eigen-value ratio retains approximately 95% of the variance of the face data. Due to lack of intra-class data, the LDA features are derived directly from the PCA features. The speech information is extracted using Morlet wavelets. The speech features consist of wavelet coefficients in different scales that are truncated and sub-sampled into a fixed length feature. More details can be obtained from [4, 23].

6 Empirical Results

In order to analyse the effects due to VR techniques, we first present the baseline experimental results. This is followed by results obtained by various VR techniques.

6.1 Baseline Performance on The XM2VTS Database

Three types of face features, one type of speech features and two classifiers are used here. The features are:

1. **FH**: It is a normalised **f**ace image concatenated with its RGB **H**istogram (thus the abbreviation **FH**)

2. **DCTs**: It is a set of face features derived from a subset of DCT-derived coefficients. The DCT algorithm used overlapping windows (block of sub-image) having the size of 40×32 pixels. (**s** indicates the use of this small image comparing to the bigger size image with the abbreviation **b**).

3. **DCTb**: Similar to DCTs except that it uses overlapping windows having the size of 80×64.

4. **LFCC**: The Linear Filter-bank Cepstral Coefficient (LFCC) speech features were computed with 24 linearly-spaced filters on each frame of Fourier coefficients sampled with a window length of 20 milliseconds and each window moved at a rate of 10 milliseconds. 16 DCT coefficients are computed to decorrelate the 24 cepstrum coefficients obtained from the linear filter-bank.

Two different types of classifiers were used for these experiments: a Multi-Layer Perceptron (MLP) [6] and a Bayes Classifier using Gaussian Mixture Models (GMMs) to estimate the class distributions [6]. While in theory both classifiers could be trained using any of the previously defined feature sets, in practice only some specific combinations appear to yield reasonable performance.

Whatever the classifier is, the hyper-parameters (e.g. the number of hidden units for MLPs or the number of Gaussian components for GMMs) are tuned on the evaluation set LP1 Eval. The same set of hyper-parameters are used in both LP1 and LP2 configurations of the XM2VTS database.

For each client-specific MLP, the samples associated to the client are treated as positive patterns while all other samples *not* belong to the client are treated as negative patterns. This is commonly called the one-against-all strategy. All MLPs reported here were trained using the stochastic version of the error-backpropagation training algorithm [6].

For the GMMs, two competing models are often needed: a world and a client-dependent model. Initially, a world model is first trained from an external database (or a sufficiently large data set) using the standard Expectation-Maximisation algorithm [6]. The world model is then adapted for each client to the corresponding client data of the training set of the XM2VTS database using the Maximum-A-Posteriori adaptation [34] algorithm.

The baseline experiments based on DCT coefficients were reported in [30] while those based on normalised face images and RGB histograms (FH features) were reported in [19]. Details of the experiments, coded in terms of pairs of **(features, classifiers)**, are given below:

1. **(FH,MLP)** Features are normalised **F**ace concatenated with **H**istogram features. The client-dependent classifier used is an MLP with 20 hidden units. The MLP is trained with geometrically transformed images [19].
2. **(DCTs,GMM)** The face features are DCT-derived coefficients with each overlapping window (block of sub-image) having the size of 40×32 pixels There are 64 Gaussian components in the GMM. The world model is trained using *all the clients* in the training set [30].
3. **(DCTb,GMM)** Similar to (DCTs,GMM), except that the features used are DCT-derived coefficients with the overlapping window-size of 80×64. The corresponding GMM has 512 Gaussian components [30].
4. **(DCTs,MLP)** Features are the same as those in (DCTs,GMM) except that an MLP is used in place of a GMM. The MLP has 32 hidden units [30].
5. **(DCTb,MLP)** The features are the same as those in (DCTb,GMM) except that an MLP with 128 hidden units is used [30].
6. **(LFCC,GMM)** This is the Linear Filter-bank Cepstral Coefficients (LFCC) obtained from the speech data of the XM2VTS database. The GMM has 500 Gaussian components and this is the best known model currently available.

The baseline experiments are shown in Table 2.

As can be seen, among the face experiments, (DCTb,GMM) performs the best across all configurations while (DCTb,MLP) performs the worst. Regardless of strong or weak classifiers, as long as their correlation is weak, they can be used in the VR techniques. Note that in the LP2 configurations, (FH,MLP), (DCTb,GMM) and (LFCC,GMM) are the only results currently available.

6.2 VR via Different Modalities

From the 5 baseline experiments on the face and 1 baseline experiment on the speech in LP1 configuration, we combined each baseline face experiment with the single speech baseline. In LP2, only 2 face baseline experts and 1 speech baseline expert are available. Therefore, in total, there are 5 and 2 experiments that can be performed on LP1 and LP2, respectively. Note that all scores are normalised to unit variance (i.e., a score is subtracted by its mean and divided by its standard deviation based on the evaluation set). The scores are combined using the average operation defined in Equation 2.

The "Average HTER" in Table 3 is calculated by taking the average HTER of its corresponding participating baseline expert taken from Table 2. Two important observations can be made here: (i) the combined experts are better than the average of their participating experts, and (ii) the combined

Table 2: Baseline performance of different modalities evaluated on the XM2VTS database based on *a priori* selected thresholds

Data sets	(Features, classifiers)	FAR	FRR	HTER
Face LP1 Test	(FH,MLP)	1.751	2.000	1.875
Face LP1 Test	(DCTs,GMM)	4.454	4.000	4.227
Face LP1 Test	(DCTb,GMM)	1.840	1.500	1.670
Face LP1 Test	(DCTs,MLP)	3.219	3.500	3.359
Face LP1 Test	(DCTb,MLP)	4.443	8.000	6.221
Voice LP1 Test	(LFCC,GMM)	1.029	1.250	1.139
Face LP2 Test	(FH,MLP)	1.469	2.250	1.860
Face LP2 Test	(DCTb,GMM)	1.039	0.250	0.644
Voice LP2 Test	(LFCC,GMM)	1.349	1.250	1.300

experts *happen to be* better than any of their participating experts. Therefore, VR via different modalities is a promising approach.

6.3 VR via Samples

For experiments on VR via samples, (see Table 4), we decided to use the LSIIT database because it has 5 test samples and each sample is not necessarily taken in the order of time sequence. The face feature is extracted using the PCA algorithm. The speech feature is extracted using wavelets transformation. As more and more samples are used for testing (from 1 to 5), the HTER for each modality decreases, but the rate of decrease is also smaller each time. This indicates that continuing to add more samples may not help, because the HTER may be 'saturated", i.e, variance cannot be decreased further. Given the small size of the database, 0 HTER error rate is reached for the face features.

6.4 VR via Synthetic Samples

To illustrate the concept of VR via synthetic samples, the trained (FH,MLP) baseline model is reused. However, for each original test and evaluation face image, 329 synthetically transformed face images (using translation, scaling and mirroring operations) are added into the evaluation and test set. The resultant 330 scores are simply averaged, according to the variance reduction theory discussed earlier. The results are shown in Table 5. Each odd line (labelled "Original", i.e., the baseline found in Table 2) in Table 5 should be compared to its corresponding even line (labelled "Averaged", i.e., with synthetic samples). For both data sets LP1 Test and LP2 Test, VR via synthetic samples always outperforms the original baseline. Although generating synthetic samples does not give additional information, surprisingly it does help in reducing variance of scores obtained from the MLP. If training with noise may help in generalisation, testing with noise can also probably be helpful. This phenomenon can be explained by VR via synthetic samples described earlier. Further details can be obtained from [21].

6.5 VR via Extractors

In this section, both the LSIIT and XM2VTS databases are used. Experiments on VR via extractors are repeated 10 times on the LSIIT face database. The LDA and PCA features are used to provide two different types of features. 5 samples are used for training and the remaining 5 samples are used for testing. These results are shown in Table 6. μ and σ in Table 6 are, respectively, mean and standard deviation of the HTER values above. The column labelled "Combined" shows the HTER based on the

Table 3: Performance of VR via modalities of the XM2VTS database based on *a priori* selected thresholds

Data sets	(Features, classifiers)	FAR	FRR	HTER	Average HTER
LP1 Test	(FH,MLP) (LFCC,GMM)	0.299	0.500	0.399†	1.507
LP1 Test	(DCTs,GMM) (LFCC,GMM)	0.574	0.500	0.537†	2.683
LP1 Test	(DCTb,GMM) (LFCC,GMM)	0.291	0.750	0.520†	1.405
LP1 Test	(DCTs,MLP) (LFCC,GMM)	0.181	1.000	0.591†	2.249
LP1 Test	(DCTb,MLP) (LFCC,GMM)	0.494	0.500	0.497†	3.680
LP2 Test	(FH,MLP) (LFCC,GMM)	0.052	0.250	0.151†	1.580
LP2 Test	(DCTb,MLP) (LFCC,GMM)	0.045	0.250	0.147†	0.972

† indicates better performance that the participating experts

Table 4: Performance of VR via real samples evaluated on the LSIIT database using *a posteriori* selected thresholds

No. of samples used	Face HTER	Voice HTER
1	7.184	6.897
2	2.701	4.828
3	1.207	4.540
4	0.000	2.126
5	0.000	2.414

average of normalised scores obtained from two MLP classifiers (for LDA and PCA features), while the column labelled "Average" shows the HTER based on the average HTER of the two face experts. Both the classifiers happen to have 8 hidden units. These numbers of hidden units are selected by validation on the test set, hence all results become optimistically biased but the overall behaviour can still be analysed.

It can be observed that in this particular case, the average HTER of the combined expert (forth column of Table 6) is lower than any of the two participating experts, even though the VR justification in Section II modestly asserts that the combined HTER will always be lower than the average HTER of the two experts. Furthermore, the standard deviation of the combined expert is also lower than any of its two underlying expert. This means that the scores of the combined expert are more stable. This is another advantage of using VR techniques.

In the second experiment using this approach, we used the more realistic XM2VTS database. However, several runs of the experiment is not possible as in the case of the LSIIT database. The 5 baseline face experts in LP1 configuration shown in Table 2 were used. As for the LP2 configuration, the 2 face experts are also used. These two sets of experiments are carried out on the LP1 Eval, LP1 Test, LP2 Eval and LP2 Test data sets. However, only the LP1 Test and LP2 Test results, which give *a priori* HTERs, are shown in Table 7.

Table 5: Performance of VR via synthetic samples evaluated on the XM2VTS database based on *a priori* selected thresholds

Data sets	Models	FAR	FRR	HTER
LP1 Test	Original	1.751	2.000	1.875
LP1 Test	Averaged	1.474	1.750	1.612
LP2 Test	Original	1.469	2.250	1.860
LP2 Test	Averaged	1.285	1.750	1.518

Table 6: Performance of VR via extractors evaluated on 10 runs of experiments of the LSIIT database using *a posteriori* selected thresholds

Experiments	HTER			
	LDA	PCA	Combined	Average
1	3.460	15.069	1.333†	9.264
2	6.621	17.471	3.368†	12.046
3	3.655	10.310	1.437†	6.983
4	6.598	20.862	3.253†	13.730
5	4.034	16.057	2.529†	10.046
6	0.161	16.724	0.805	8.443
7	1.345	14.172	3.333	7.759
8	3.989	15.483	2.023†	9.736
9	0.046	15.287	1.437	7.667
10	0.667	14.460	1.333	7.563
μ	3.058	15.590	2.085	9.324
σ	2.442	2.675	0.965	2.162

† indicates better performance that the participating experts

Two observations can be made: (i) the HTER of all 5 combined experts are better than the average of their participating experts, and (ii) out of 5 combined experts, 3 have lower HTER than their participating experts (see Table 2 for comparison).

6.6 VR via Classifiers

VR via classifiers can be viewed as an ensemble algorithm. The simplest ensemble, with well-founded theory according to Bishop [6, Chap. 9], is to average the output of classifiers trained on the *same* feature type. This is also in accordance to the demonstration shown in Section II. The baseline experts of Table 2 that meet this criterion are the pairs {(DCTs,GMM), (DCTs,MLP)} and {(DCTb,GMM), (DCTb,MLP)}, which are available in XM2VTS LP1 protocol only. For this reason, experiments are carried out on LP1 only. The combined experts are shown in Table 8.

Similar observations can be made: (i) all combined experts are better than the average experts; and (ii) 1 out of 2 combined expert is better than its participating experts.

6.7 Comparison of various VR Techniques

Among various VR techniques, which one is better? Based on the experiments already done, it is possible to calculate the factor of reduction of a given combined expert with respect to the average

Table 7: Performance of VR via extractors evaluated on the XM2VTS database based on *a priori* selected thresholds

Data sets	(Features, classifiers)	FA	FR	HTER	Average HTER
LP1 Test	(FH,MLP) (DCTs,GMM)	2.033	1.250	1.641†	3.051
LP1 Test	(FH,MLP), (DCTb,GMM)	1.495	0.750	1.123†	1.773
LP1 Test	(FH,MLP), (DCTs,MLP)	1.450	1.500	1.475†	2.617
LP1 Test	(FH,MLP), (DCTb,MLP)	1.146	2.750	1.948	4.048
LP2 Test	(FH,MLP) (DCTb,GMM)	0.792	1.000	0.896	1.252

† indicates better performance than the participating experts

Table 8: Performance of VR via different classifiers evaluated on the XM2VTS database based on *a priori* selected thresholds

Data sets	(Features, classifiers)	FA	FR	HTER	Average HTER
LP1 Test	(DCTs,GMM), (DCTs,MLP)	3.245	2.500	2.873†	2.949
LP1 Test	(DCTb,GMM), (DCTb,MLP)	2.046	3.750	2.898	4.790

† indicates better performance than the participating experts

HTER of its participating experts. Let us call this factor β, which can be defined formally as follows:

$$\beta = \frac{\mathrm{mean}_i(\mathrm{HTER}_i)}{\mathrm{HTER}_c}, \tag{12}$$

where i is the index to each participating expert and HTER_c is the HTER of the combined expert.

Note that β is analogous to α defined in Equation 8. They are however not the same because α measures the factor of variance reduction while β measures the factor of reduction in terms of HTER. This will give a relative measure to quantify empirically the importance of each class of VR techniques. β can easily be calculated from Tables 3, 7, 8 and 5 by using the average HTER and the HTER of the combined experts, all based on the XM2VTS database. Note that for the experiments on VR via real samples (Table 6), experiments are based on the LSIIT database, different from the rest of the experiments. Furthermore, it uses *a posteriori* selected thresholds, different from the rest of the results that use *a priori* selected thresholds. As a result, VR via real samples may be overly optimistic, i.e., its β value may appear slightly higher than it would have been on unseen data. The results are shown in Table 9. The third column in Table 9 shows the number of data that is used to calculate the corresponding mean, standard deviation (fourth column) and median (fifth column). It can be observed that VR techniques via modalities (first line) has the highest β value. However, from its variance, one can observe that its HTERs fluctuate very much. This is due to the possible inaccuracy when the HTER of combined expert becomes very small.

Both the median and mean of β show that VR via modalities is the best VR technique. VR via real samples is the second most effective VR technique. Again, this should be interpreted with care

Table 9: Comparison of various VR techniques based on all experiments carried out

VR techniques	Table	No. of data	mean(β) $\pm$ std.dev.	median(β)
Modalities	3	7	5.680$\pm$ 2.683	4.996
Real samples	6†	10	5.222 $\pm$ 2.234	4.836
Extractors	7	5	1.730 $\pm$ 0.249	1.774
Classifiers	8	2	1.340 $\pm$ 0.443	1.340
Synthetic samples	5	2	1.230 $\pm$ 0.00004	1.230

† indicates performance taken from the LSIIT database using *a posteriori* selected thresholds

because its HTERs are obtained *a posteriori*. Fortunately, between VR via real samples and VR via extractors, there is a considerable gap of β difference, i.e., between the mean of 5.680 and the mean of 1.730 of VR via modalities and VR via extractors, respectively. The next most effective VR technique is VR via extractors, followed by VR via classifiers and VR via synthetic samples.

This order of reduction factor is in fact not a coincidence. It reveals their degree of independence. Higher β can be interpreted as higher independence. The degree of independence due to modalities is higher than that due to real samples, the degree of independence due to real samples is higher than that due to features, and etc.

Interestingly, from the application point of view, VR via modalities also has the highest overhead cost, i.e., cost of adding a new modality, software, computation time, etc. VR via real samples has it's own overhead cost as well: more access time needed to scan multiple biometric samples. For the rest of VR techniques, the computation cost also decreases from VR via extractors to VR via classifiers. In a nutshell, from the point of view of VR techniques, the accuracy of a BA system can be boosted by means of adding more computation cost. Fortunately, this cost increases only linearly in the number of units added in the VR techniques.

7 Conclusion

Variance reduction (VR) is an important technique to increase accuracy in regression and classification problems. In this study, several approaches are explored to improve Biometric Authentication systems, namely VR via modalities, VR via extractors, VR via classifiers, VR via real samples and VR via synthetic samples. A brief survey is made on the literature according to these techniques. The experiments carried out on the XM2VTS database have shown that the combined experts due to VR techniques *always* perform better than the average of their participating experts. Furthermore, *most* combined experts outperform the best participating expert based on the HTER. It is shown empirically that VR via modalities is the best VR techniques, followed by VR via real samples, VR via extractors, VR via classifiers and VR via synthetic samples. This can be explained by the independence of the scores due to these VR techniques. The higher the degree of independence of a given VR technique, the higher the reduction factor, i.e., the ratio between the average HTER of the participating experts and the HTER of the combined expert. As new and more powerful extraction and classification algorithms become available, they can all be integrated into the VR framework. Therefore, VR techniques are potentially very useful for biometric authentication.

Acknowledgement

The authors would like to thank Fabien Cardinaux for supplying some of the data to carry out the experiments.

References

[1] A. Jain, R. Bolle, and S. Pankanti, *Biometrics: Person Identification in Networked Society.* Kluwer Publications, 1999.

[2] A. Ross, A. Jain, and J.-Z. Qian, "Information fusion in biometrics," in *The 3rd International Conference on Audio-Visual Biometric Person Authentication (AVBPA)*, 2001, pp. 354–359.

[3] L. Hong, A. Jain, and S. Pankanti, "Can multibiometrics improve performance?" MSU-CSE" Technical Report MSU-CSE-99-39, 12 1999.

[4] N. Poh and J. Korczak, "Hybrid biometric authentication system using face and voice features," in *The 3rd International Conference on AVBPA*, 2001, pp. 348–353.

[5] J. Kittler, G. Matas, K. Jonsson, and M. Sanchez, "Combining evidence in personal identity verification systems," 1997. [Online]. Available: citeseer.nj.nec.com/kittler97combining.html

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

[7] A. Jain and S. Pankanti, "Biometrics systems: Anatomy of performance," Department of Computer Science, Michigan State University, East Lansing, Michigan, Tech. Rep. MSU-CSE-00-20, September 2000.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech'97, Rhodes, Greece*, 1997, pp. 1895–1898.

[9] T. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, 2000, pp. 1–15. [Online]. Available: citeseer.nj.nec.com/dietterich00ensemble.html

[10] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. [Online]. Available: citeseer.nj.nec.com/breiman94bagging.html

[11] R. Maclin and D. Opitz, "An empirical evaluation of bagging and boosting," in *AAAI/IAAI*, 1997, pp. 546–551. [Online]. Available: citeseer.nj.nec.com/maclin97empirical.html

[12] Y. Freund and R. Schapire, "A short introduction to boosting," in *J. Japan. Soc. for Artificial Intelligence*, 1999, pp. 771–780. [Online]. Available: citeseer.nj.nec.com/freund99short.html

[13] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995. [Online]. Available: citeseer.nj.nec.com/dietterich95solving.html

[14] J. Kittler, R. Ghaderi, T. Windeatt, and J. Matas, "Face identification and verification via ecoc," in *The 3rd International Conference on Audio-Visual Biometric Person Authentication (AVBPA)*, 2001, pp. 1–13.

[15] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Oxford University Press, 1993.

[16] R. Chellappa and S. Sirohey, "Human and machine recognition of faces: A survey," in *Proc. of the IEEE*, vol. 83, no. 9, May 1995, pp. 705–740.

[17] C. Havran, L. Hupet, J. Czyz, J. Lee, L. Vandendorpe, and M. Verleysen, "Independent component analysis for face authentication," in *International Conference on Knowledge-Based Intelligent Information and Engineering*, September 2002, pp. 1207–1211.

[18] R. Brunelli and D. Falavigna, "Personal identification using multiple cues," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17(10), 1995, pp. 955–966.

[19] S. Marcel and S. Bengio, "Improving face verification using skin color information," in *Proceedings of the 16th International Conference on Pattern Recognition.* IEEE Computer Society Press, 2002.

[20] J. Luettin, "Visual speech and speaker recognition," Ph.D. dissertation, Department of Computer Science, University of Sheffield, 1997.

[21] N. Poh, S. Marcel, and S. Bengio, "Improving face authetication using virtual samples," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, no. 40, 2003, IDIAP-RR, iDIAP-RR.

[22] L. Hong and A. Jain, "Multi-model biometrics," in *Biometrics: Person Identification in Networked Society*, 1999.

[23] N. Poh, "Biometric authentication system," Master's thesis, USM, Penang, August 2001.

[24] U. Dieckmann, P. Plankensteiner, and T. Wagner, "Sesam: A biometric person identification system using sensor fusion," in *In Pattern Recognition Letters*, vol. 18(9), 1997, pp. 827–833.

[25] P. Jourlin, J. Luettin, D. Genoud, and H. Wassne, "Acoustic-labial speaker verification," in *Pattern Recognition Letters*, vol. 18(9), 1997, pp. 853–858.

[26] Conrad Sanderson and Kuldip K. Paliwal, "Information Fusion and Person Verification Using Speech & Face Information," IDIAP, IDIAP-RR 33, September 2002.

[27] 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 the xm2vts database," in *Proceedings of the 15th ICPR*, vol. 4, 2000, pp. 858–863.

[28] J. Lüttin, "Evaluation protocol for the XM2FDB database (lausanne protocol)," IDIAP, Tech. Rep. COM-05, 1998.

[29] C. Sanderson and K. Paliwal, "Polynomial features for robust face authentication," in *Proceedings of International Conference on Image Processing*, vol. 3, 2002, pp. 997–1000.

[30] F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of MLP and GMM classifiers for face verification on XM2VTS," IDIAP, IDIAP-RR 10, 2003.

[31] Mariéthoz, J. and Bengio, S., "A comparative study of adaptation methods for speaker verification," in *International Conference on Spoken Language Processing ICSLP*, Denver, CO, USA, September 2002, pp. 581–584, iDIAP-RR 01-34.

[32] M. Turk and A. Pentland, "Eigenfaces for recognition," in *J. of Cognitive Neuroscienc*, vol. 3(1), 1991, pp. 71–86.

[33] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," in *ECCV (1)*, 1996, pp. 45–58. [Online]. Available: citeseer.nj.nec.com/belhumeur96eigenfaces.html

[34] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture obervation of markov chains," in *IEEE Transactions on Speech Audio Processing*, April 1994, pp. 290–298.