IDIAP RESEARCH REPORT

Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments?

Norman Poh [a] Samy Bengio [a]

IDIAP–RR 05-20

May 2005

SUBMITTED FOR PUBLICATION

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

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-20

Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments?

Norman Poh Samy Bengio

May 2005

submitted for publication

Abstract. Combining multiple information sources, typically from several data streams is a very promising approach, both in experiments and to some extents in various real-life applications. A system that uses *more than one* behavioural and physiological characteristics to verify whether a person is who he/she claims to be is called a *multimodal* biometric authentication system. Due to lack of large true multimodal biometric datasets, the biometric trait of a user from a database is often combined with another different biometric trait of yet another user, thus creating a so-called a *chimeric user*. In the literature, this practice is justified based on the fact that the underlying biometric traits to be combined are assumed to be independent of each other given the user. To the best of our knowledge, there is no literature that approves or disapproves such practice. We study this topic from two aspects: 1) by clarifying the mentioned independence assumption and 2) by constructing a pool of chimeric users from a pool of *true* modality matched users (or simply "true users") taken from a bimodal database, such that the performance variability due to chimeric user can be compared with that due to true users. The experimental results suggest that for a large proportion of the experiments, such practice is indeed questionable.

1 Introduction

Biometric authentication (BA) is a 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 essentially verifies "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness. Examples of biometric modalities are fingerprints, faces, voice, hand-geometry and retina scans [9].

Due to inherent properties in each biometric and external manufacturing constraints in the sensing technologies, no single biometric trait can achieve 100% authentication performance. This problem can be alleviated by combining two or more biometric traits, also known as the field of multimodal biometric authentication. In the literature, there are several approaches towards studying fusion of modalities. One practice is to construct a large database containing several biometric traits for each user. This, however, can be very time-consuming and expensive. Another practice is to combine biometric modalities from a database with biometric modalities of another biometric database. Since both databases do not necessarily contain the *same* users, such combination results in *chimeric users*. From the experiment point of view, these biometric modalities belong to the same person. While this practice is commonly used in the multimodal literature, e.g., [18, 7] among others, it was questioned whether this was a right thing to do or not during the 2003 Workshop on Multimodal User Authentication [6]. To the best of our knowledge, there is no work in the literature that approves or disapproves such assumption. In general, the use of chimeric users is justified by the *modality independence assumption*: that two or more biometric traits of a single person are independent of each other. We set out to investigate the validity of this assumption by using two approaches, namely : 1) by pinning down the concept of *independence* and 2) by simulating the effect of chimeric users experimentally and measuring the discrepancy in terms of performance between the use of chimeric users and the use of true users. To verify this hypothesis, we limit our scope to studying such effect to bimodal as generalisation to more than two modalities is direct. It should be emphasised that the use of chimeric users practice is not limited to biometric authentication, but may be in general applicable to problems involving multimodal streams. Hence, this study is of interest to researchers studying multimodal fusion.

Techniques that are designed to combine multimodal systems are referred to as fusion techniques. From an architectural point of view, fusion can be done either at the *feature level* (extracted or internal representation of the data stream) or *score level* (output of a single system). Among the two, the latter is most commonly used in the literature. The use of chimeric users falls into the latter case as well. Some studies further categorize three levels of score level fusion [5], namely, fusion using the scores directly, using a *set of most probable* category labels (called abstract level) or using the *single most probable* category label (called decision level). We will focus on the score for two reasons: the last two cases can be derived from the score and more importantly, by using only labels instead of scores, precious information is lost.

This paper is organised as follows: Section 2 underpins the concept of independence between biometric traits; Section 3 describes the database to be used; Section 4 details the experimental procedure and presents the results; and finally this is followed by conclusions in Section 5.

2 On the Independence Assumption

2.1 Preliminary

Suppose that each authorised person is identified by a unique identity claim $j \in \mathcal{J} \equiv \{1, \ldots, J\}$ and there are J identities. We sometimes call these users as clients to oppose a set of other unauthorised persons known as impostors. Hence, a biometric authentication system is aimed at distinguishing clients from impostors, which is an *aggregated* two-class problem, i.e., a two-class problem with J distinctive users. In this problem, it is common to represent a user by his/her feature template or

model, i.e, a set of parameters derived from the features. Suppose that the output due to comparing a user model C_j to a feature X is y_j. For each client or user model C_j, there is a corresponding impostor model I_j. Lacking a proper definition[1], the impostor model is often *naively* defined as the model of other finite users $\forall_{j'}|j' \in \mathcal{J} - j$. To decide whether to accept or reject the access request represented by feature X claiming identity j, one can evaluate the *posterior probability ratio* in logarithmic domain (called log-posterior ratio, LPR):

$$\begin{aligned} \mathrm{LPR}_j &\equiv \log\left(\frac{P(C_j|X)}{P(I_j|X)}\right) = \log\left(\frac{P(X|C_j)P(C_j)}{P(y|I_j)P(I_j)}\right) \\ &= \underbrace{\log\frac{P(X|C_j)}{P(X|I_j)}} + \underbrace{\log\frac{P(C_j)}{P(I_j)}} \equiv y(j) - \Delta_j, \end{aligned}$$

where we introduced the term $y(j)$ (also called a Log-Likelihood Ratio, LLR) and a threshold Δ to handle the case of different priors. This constant also reflects the different *costs* of false acceptance and false rejection. In both cases, the threshold Δ has to be fixed *a priori*. The decision of accepting or rejecting an access is then:

$$\mathrm{decision}(\mathrm{LPR}_j) = \begin{cases} \text{accept} & \text{if } \mathrm{LPR}_j > 0 \\ \text{reject} & \text{otherwise.} \end{cases} \quad (1)$$

or

$$\mathrm{decision}(y(j)) = \begin{cases} \text{accept} & \text{if } y(j) > \Delta_j \\ \text{reject} & \text{otherwise.} \end{cases} \quad (2)$$

Although both forms are equivalent, the explicit presence of a threshold in the second decision function shows that the log-prior ratio can be adjusted *separately* from the score $y(j)$. Note that $y(j)$ is a direct function of X and the model variable associated to it (say θ_j), i.e., $y(j) = f_{\theta_j}(X)$. We use the function f with parameter θ to explicitly represent the *functional relationship* between the variables $y(j)$ and X. Suppose that $y(j)$ is an instance of the variable $Y(j)$ and is drawn from the distribution $\mathcal{Y}(j)$. The decision function in Eqn. (2) then implies that $E_{\mathcal{Y}(j)|C_j}[Y(j)] > E_{\mathcal{Y}(j)|I_j}[Y(j))]$, where $E_{\mathcal{Z}}[Z]$ is the expectation of Z under the law $\mathcal{Z}$. Typically, on a per user basis, one has an *extremely* few number of samples (depending on the protocols, there are 2 or 3 in the XM2VTS database [14], and 1 or 3 in the BANCA database [1]) to estimate the genuine distribution $\mathcal{Y}(j)|C_j$ whereas one has a relatively large number of samples (typically in the order of hundreds) to estimate the impostor distribution $\mathcal{Y}(j)|I_j$.

Although $y(j)$ is interpreted as an LLR here, many different machine-learning algorithms (e.g., Gaussian Mixture Models, Multi-Layer Perceptrons, Support Vector Machines) can be viewed as an approximation to this relationship, without necessarily giving it a probabilistic interpretation, i.e., $y(j)$ being a probability. This is to contrast with most Bayesian analyses, e.g., [12, 11], that start by making an equivalence between $y(j)$ and $p(C_j|X)$, i.e., $y(j) \equiv p(C_j|X)$, such that the threshold $\Delta = 0.5$. Although this probabilistic interpretation is correct, it does not *explicitly* consider the fact that the threshold *changes* with the priors of client and impostor classes. In fact, the prior has already been integrated in $p(C_j|X) \propto p(X|C_j)p(C_j)$, which is the product between likelihood and the client prior. As a matter of fact, most biometric authentication systems crucially rely on this threshold to make the accept/reject decision. For instance, if the matching score $y(j)$ is based on a distance between a user template X_j and the submitted feature X, i.e., $y(j) \equiv dist(X, X_j)$, where *dist* is a distance measure, then our Bayesian model above is still valid by interchanging between C_j and I_j, such that $E_{\mathcal{Y}(j)|C_j}[Y(j)] < E_{\mathcal{Y}(j)|I_j}[Y(j)]$. This distance measure simply cannot be interpreted in the probabilistic framework with $y(j) \equiv p(C_j|X)$.

Depending on the outcome of the decision (as a function of the threshold Δ_j), a biometric authentication system can commit two types of errors, namely, False Acceptance (FA) and False Rejection

[1]Ideally, this impostor model should be the world population minus the user j. In terms of computation and data collection effort, this is not feasible and in practice not necessary.

(FR). The error rates of FA and FR are defined as:

$$\begin{aligned} \mathrm{FAR}(\Delta_j) &= 1 - p(Y(j)|I_j \leq \Delta_j) \\ \mathrm{FRR}(\Delta_j) &= p(Y(j)|C_j \leq \Delta_j), \end{aligned}$$

where $p(Y(j)|k_j \leq \Delta_j)$ is the cummulative density function of conditional variable $Y(j)$ within the range $[-\infty, \Delta_j]$ for each class k_j. Note that a unique point with Δ_j^* where $\mathrm{FAR}(\Delta_j^*) = \mathrm{FRR}(\Delta_j^*)$ is called Equal Error Rate (EER). EER is often used to characterise a system's performance. Another useful performance evaluation point for *any given threshold* Δ_j (not necessarily Δ_j^*) is called Half Total Error Rate (HTER) and is defined as the average of FAR and FRR, i.e.,:

$$\mathrm{HTER}(\Delta_j) = \frac{1}{2}(\mathrm{FAR}(\Delta_j) + \mathrm{FRR}(\Delta_j)).$$

The discussion until here concerns only a particular client indexed by j. In reality, one has extremely few examples of genuine accesses $y(j)|C_j$ and relatively large impostor accesses $y(j)|I_j$, as mentioned earlier. As a result, the estimation of the threshold Δ_j, i.e., user-specific threshold, is extremely unreliable, and the user-independent versions of FAR, FRR and EER, as well as the threshold are often used. Although there exists abundant literature to estimate user-specific threshold (see for instance a survey in [16, 19]), common threshold is by far a standard practice.

2.2 Different levels of Dependency Assumption

We now introduce two notions of dependencies, i.e, *feature-oriented* dependency and *score-oriented* dependency. The former assumes dependency at the feature-level while *not considering* the dependency at the score level. The latter, on the other hand, assumes *independence at the feature level* but handles dependency uniquely at the score level.

These two dichotomies thus give rise to four types of dependencies in *decreasing order*:

- **Strict Feature Dependence.** It is characterised uniquely by the feature-oriented dependence assumption.
- **Loose Feature Dependence** It is characterised by feature-oriented independence but score-oriented dependence
- **Loose Feature Independence** It is characterised by both feature-oriented and score-oriented independence.
- **Strict Feature Independence.** It is characterised uniquely by the feature-oriented independence assumption.

Suppose that X_1 and X_2 are features of two different biometric modalities. Using the same Bayesian formulation (with focus on LLR) as in the previous Section, the four categories can be formally stated as follows:

- **Strict Feature Dependence:**

$$\begin{aligned} y_{SD}(j) &= \log \frac{p(X_1, X_2|C_j)}{p(X_1, X_2|I_j)} \quad (3) \\ &\equiv f_{\theta_j}(X_1, X_2), \quad (4) \end{aligned}$$

  where the function f explicitly represents any classifier with the associated parameter θ_j. By so doing, we actually provide a Bayesian interpretation of the classifier f. One possible weakness of this approach is known as the "curse of dimensionality", whereby modeling the joint features in higher dimension can cause a degraded performance compared to methods resulting from the other assumptions (to be discussed below).

- **Strict Feature Independence:**

$$
\begin{aligned}
y_{SI}(j) &= \log \frac{p(X_1|C_j)p(X_2|C_j)}{p(X_1|I_j)p(X_2|I_j)} && (5)\\
&= \log \frac{p(X_1|C_j)}{p(X_1|I_j)} + \log \frac{p(X_2|C_j)}{p(X_2|I_j)} && (6)\\
&= y_1(j) + y_2(j) && (7)\\
&\equiv f_{\theta_j^1}(X_1) + f_{\theta_j^2}(X_2) && (8)
\end{aligned}
$$

where $y_i(j) \equiv \log \frac{p(X_i|C_j)}{p(X_i|I_j)}$ and θ_j^i is the model parameter associated to modality i and user j. Note that in theory the two classifiers involved, $f_{\theta_j^i}|i = \{1, 2\}$, do not have to be homogeneous (the same type). In practice, however, some form of normalisation may be needed if they are not homogeneous, e.g., from different vendors or based on different algorithms. It can be seen that using this Bayesian framework, the independence assumption leads to the well-known sum rule. On the other hand, using the probabilistic framework $y(j) \equiv p(C_j|X)$, this dependency would have led to the well-known product rule (proof not shown here).

- **Loose Feature Dependence:**

$$
\begin{aligned}
y_{LD}(j) &= \log \frac{p(y_1(j), y_2(j)|C_j)}{p(y_1(j), y_2(j)|I_j)} && (9)\\
&\equiv f_{\theta_j^{COM}}(y_1(j), y_2(j)) && (10)\\
&= f_{\theta_j^{COM}}\left(f_{\theta_j^1}(X_1), f_{\theta_j^2}(X_2)\right), && (11)
\end{aligned}
$$

where $f_{\theta_j^{COM}}$ can be considered as a second-level classifier, also called a fusion classifier. The loose feature dependence is a result of committing to the feature independence assumption – which means that the scores $y_1(j)$ and $y_2(j)$ can be derived separately – and score-oriented dependence assumption – implying that the dependency at the score level should be modeled. This formulation actually motivates the use of trainable classifiers in fusion. Suppose that $\boldsymbol{y}(j) = [y_1(j), y_2(j)]^T$ is a vector and an instance of the variable $\boldsymbol{Y}(j)$. If $\boldsymbol{Y}(j)$ is drawn from a class-conditional Gaussian distributions and that both the client and impostor distributions share a common covariance matrix $\boldsymbol{\Sigma}$, it is possible to show that:

$$f_{\theta_j^{COM}} = w_1(j)y_1(j) + w_2(j)y_2(j), \tag{12}$$

where $\boldsymbol{w}(j) = [w_1(j), w_2(j)]^T$ has the following solution:

$$\boldsymbol{w}(j) \propto \boldsymbol{\Sigma}^{-1}\left(E[\boldsymbol{Y}(j)|C_j] - E[\boldsymbol{Y}(j)|I_j]\right) . \tag{13}$$

The linear opinion pool (or weighted sum) shown here is a typical solution given by Fisher's linear discriminant [3, Sec. 3.6]. Other solutions using the same linear discriminant function (but possibly *more powerful* since they do not make the class-conditional Gaussian assumption) includes Support Vector Machines with a linear kernel [20] and the perceptron algorithm [3, Chap. 6], the latter of which generalises to the least square and the logistic discrimination/regression solutions (depending on the error criterion). It can thus be seen that the loose feature dependence assumption motivates the use of a fusion classifier. It should be noted that the Bayesian framework using Eqn. (9) as a departure point does not dictate that a linear classifier has to be used. In practice, however, to the best of our knowledge, non-linear classifiers have not been reported to provide *significantly* better results over their linear counterparts in this application. Often, due to small training sample size on a *per user basis*, the classifier at this level is trained across all users. Although user-specific fusion classifiers have been proposed, e.g., [7], global fusion classifier is by far the most commonly used approach. We will study this case here. Hence,

as long as fusion is concerned, the index j in the term $f_{\theta_j^{COM}}$ of *Eqn.* (10) can be dropped, so as for the weights in Eqn. (12).

- **Loose Feature Independence:**

$$
\begin{aligned}
y_{LI}(j) &= \log \frac{p(y_1(j)|C_j)p(y_2(j)|C_j)}{p(y_1(j)|I_j)p(y_2(j)|I_j)} && (14)\\
&= \log \frac{p(y_1(j)|C_j)}{p(y_1(j)|I_j)} + \log \frac{p(y_2(j)|C_j)}{p(y_2(j)|I_j)} && (15)\\
&\equiv f_{\theta_j^1}(y_1(j)) + f_{\theta_j^2}(y_2(j)) && (16)\\
&= f_{\theta_j^1}\left(f'_{\theta_j^1}(X_1)\right) + f_{\theta_j^2}\left(f'_{\theta_j^2}(X_2)\right), && (17)
\end{aligned}
$$

where $f'_{\theta_j^i}$ is a classifier taking features X_i and $f_{\theta_j^i}$ is another classifier taking the score $y_i(j)$, for all $i \in \{1,2\}$. Since $f_{\theta_j^i}$ is a one-input one-output function, this procedure is also called *score normalisation* [8]. Among the score normalisation techniques, user-specific Z-score normalisation is perhaps the most representative one. Z-norm and other techniques are surveyed in [16]. It turns out that the fusion classifier is a sum rule. Again, due to lack of user-specific data, the score normalisation is treated the same across users. Hence, we can replace $f_{\theta_j^i}$ by f_{θ^i} (without the subscript j) in Eqns. (16) and (17), for all $i = \{1,2\}$.

The above four types of architecture as a result of different levels of dependence assumption are certainly not exhaustive. It is possible to combine say strict feature dependence and strict feature independence assumption such that the resultant architecture compensates for both assumption (see for instance [17]).

As can be seen, depending on the level of dependence between X_1 and X_2 that one is willing to commit to, one arrives at any of the four choices of architectures. In multimodal biometrics, where two (or more) biometric modalities are captured using different sensors, it is well accepted that the strict feature dependence assumption (the first one) is *in general* not true [18]. Hence, as long as the use of chimeric users is concerned, only the last three levels of dependence are relevant. In the experimental setting with chimeric users, one simply uses the concatenated score with modalities of *other users* , i.e.,

$$\boldsymbol{y}_{chimeric} = [y_1(j), y_2(j')]^T \text{ where } j \neq j'.$$

and combines the concatenated score by using classifiers such as Eqns. (7), (10) and (16), respectively for the last three levels of dependency.

Thus we arrive at the crucial question: "**Do the different levels of dependency allow to switch the identities?**". If one follows strictly (and agrees with) the Bayesian framework presented so far, none of these assumptions provide any hint about the use of chimeric users in practice. They merely guide how one should model the final score y just before making the accept/reject decision. Lacking any plausible justification and theoretical explanation, we resolve to an experiment-driven approach to study the effects of switching identities. Before presenting the experimental approach, we first present the database used in the next section.

3 The XM2VTS Database

There exists several bimodal biometric authentication databases for this purpose, e.g., M2VTS, XM2VTS and BANCA databases. We will use the XM2VTS for two reasons: it has among the largest number of users, i.e., 200 clients and 95 casual impostors; and the results of many single modal experiments (in scores) are available for fusion. These scores are also publicly available[2] and are reported in [15].

[2] http://www.idiap.ch/∼norman/fusion

Table 1: The Lausanne Protocols as well as the fusion protocol of XM2VTS database.

Data sets	Lausanne Protocols		Fusion Protocols
	LP1	LP2	
LP Train client accesses	3	4	NIL
LP Eval client accesses	600 (3×200)	400 (2×200)	Fusion dev
LP Eval impostor accesses	40,000 ($25 \times 8 \times 200$)		Fusion dev
LP Test client accesses	400 (2×200)		Fusion eva
LP Test impostor accesses	112,000† ($70 \times 8 \times 200$)		Fusion eva

†: Due to one corrupted speech file of one of the 70 impostors in this set, this file was deleted, resulting in 200 less of impostor scores, or a total of 111,800 impostor scores.

The XM2VTS database [14] 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 recital of a sentence. The database is divided into three sets: a training set, an evaluation set and a test set. The training set (LP Train) was used to build client models, while the evaluation set (LP Eval) was used to compute the decision thresholds (as well as other hyper-parameters) used by classifiers. Finally, the test set (LP 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 approaches of the training and evaluation sets. They are called Lausanne Protocol I and II, denoted as **LP1** and **LP2** in this paper. In both configurations, the test set remains the same. Their difference is that there are three training shots per client for LP1 and four training shots per client for LP2. Table 1 is the summary of the data. More details can be found in [13]. The first column shows the data set, divided into training, evaluation and test sets. Columns two and three show the the partition of the data according to LP1 and LP2 whereas column four shows the partition of data for the fusion protocols that are *consistent* with the Lausanne Protocols. As far as fusion is concerned, there are only two data sets, labeled as "Fusion dev" (for development) and "Fusion eva" (for evaluation), since the data used in LP training sets are reserved to construct the base systems.

Note that the fusion development set is used to calculate the parameters of fusion classifier as well as the optimal global threshold. They are then applied to the fusion evaluation set. Since the threshold is *calculated from the development set*, the reported HTER obtained from the evaluation set is thus called an *a priori* HTER.

4 An Experimentally Driven Approach

This Section aims at answering the following question: "Is an experiment carried out using chimeric users *equivalent* to the one carried out using true users in terms of a given performance measure?". Suppose that the performance measure of interest is *a priori* HTER. The above question can then be rephrased as: "Is the *a priori* HTER obtained using chimeric users *similar to* (or *not significantly* different from) the one obtained using the true users?". We can formally specify the null hypothesis and its corresponding alternative hypothesis as follows:

- H_0: The *a priori* HTER obtained from chimeric users is *equivalent* to the one obtained from true users.
- H_1: The *a priori* HTER obtained from chimeric users is *different from* the one obtained from true users.

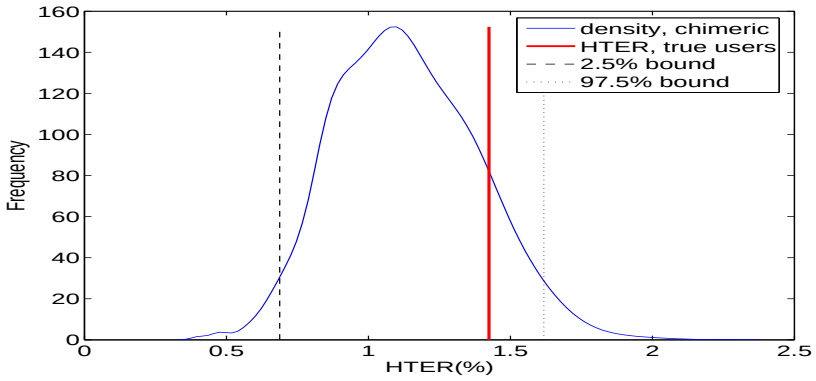

Figure 1: The distribution of *a priori* HTER (thin line) estimated from 1000 random samplings of chimeric users versus the HTER of true users (bold line). All thresholds were calculated to minimise HTER on the development set. The HTER of the true users is in the 87.7 percentil (or 1.42% HTER) and is within the 2.5 (dashed line) percentil (or 0.69% HTER) and 97.5 percentil (dotted line) (or 1.62% HTER). Hence, this experiment supports the null hypothesis.

Suppose that the HTER value due to chimeric users, v, is an instance of a random variable V which follows an unknown distribution. We are interested in:

$$p(v \in {}^{c}[a,b]|H_0) = \alpha, \tag{18}$$

where ${}^{c}[a,b]$ is the complementary of $[a,b]$ – or the *critical region*, i.e., the set of values for which we will reject H_0 – and α is the level of the test – or the Type I error, i.e., the probability of selecting H_1 when H_0 is true. By convention, α is usually set to 1% or 5%. Note that the critical region is computed such that the Type I error is only meaningful for a given α level.

Since the distribution of HTER due to chimeric users is unknown, we need to estimate it using a random permutation procedure such that in each permutation, a biometric modality of one user is paired with another biometric modality of yet another user. This procedure is somewhat similar to the bootstrap-based non-parametric statistical test [4, 10] but different in two aspects: a bootstrap manipulates samples whereas the permutation process here manipulates user identities; and a bootstrap draws samples with replacement whereas the permutation process, as its name implies, permutes identities, which means it draws identity *without replacement*. Since each permutation creates a "new" set of fusion scores, a fusion classifier has to be constructed before the HTER value can be computed. By repeatedly applying the random permutation procedure, we can then obtain a set of HTER values, which represents our non-parametric estimate of the distribution $\mathcal{V}$. Evaluating Eqn. (18) is simply a matter of determining if the HTER due to true users is in $[a,b]$ (hence in favour of H_0) or in its complement ${}^{c}[a,b]$ (hence in favour of H_1). The values a and b are chosen such that $p(v \in [a,b]) = 1-\alpha$ for a given α and $p(v < a) = p(v > b)$. Under such constraints, it is obvious to see that $p(v < a) = p(v > b) = \alpha/2$. To illustrate this idea, we took an experiment from the XM2VTS score-level fusion benchmark database, and applied the hypothesis test procedure mentioned. The results are plotted in Figure 1.

Two fusion classifiers are used in the experiments, namely the mean operator and the Gaussian Mixture Model (GMM). Both of these fusion classifiers are representative approaches of the *loose feature independence* assumption and the *loose feature dependence* assumption, respectively. For the mean operator, prior to fusion, scores are normalised to zero mean and unit variance such that none of the two expert scores dominate just because of a larger variance. The normalisation parameters are calculated from the development set. For the GMM, the number of Gaussian components is tuned by simple validation.

Table 2: The HTER range (whose confidence falls between 2.5% and 97.5% quantiles, corresponding to the usual middle 95% confidence bound) of 1000 samples of random identity match (chimeric-user effect) versus the HTER of true identity match for both the mean operator and the GMM fusion classifiers, for each of the 21 fusion datasets. For the values of HTER of true identity match falling outside the confidence range, a "$*$" sign is marked.

No.	LP	Data set (Face) (Speech) experts	HTER (%)			
			Mean		GMM	
			chimeric	true	chimeric	true
1	1	(FH,MLP)(LFCC,GMM)	[0.36, 1.02]	0.79	[0.10, 0.60]	0.35
2	1	(FH,MLP)(PAC,GMM)	[0.70, 1.36]	1.13	[0.38, 1.13]	1.08
3	1	(FH,MLP)(SSC,GMM)	[0.54, 1.24]	0.87	[0.32, 1.03]	0.72
4	1	(DCTs,GMM)(LFCC,GMM)	[0.16, 0.68]	0.53	[0.11, 0.58]	0.44
5	1	(DCTs,GMM)(PAC,GMM)	[0.71, 1.59]	1.44	[0.69, 1.62]	1.42
6	1	(DCTs,GMM)(SSC,GMM)	[0.60, 1.38]	1.14	[0.55, 1.39]	1.21
7	1	(DCTb,GMM)(LFCC,GMM)	[0.13, 0.47]	$*$ 0.55	[0.04, 0.51]	0.47
8	1	(DCTb,GMM)(PAC,GMM)	[0.30, 0.93]	$*$ 1.13	[0.29, 0.97]	$*$ 1.06
9	1	(DCTb,GMM)(SSC,GMM)	[0.27, 0.82]	0.75	[0.22, 0.82]	$*$ 0.86
10	1	(DCTs,MLP)(LFCC,GMM)	[0.52, 1.16]	0.84	[0.09, 0.58]	0.50
11	1	(DCTs,MLP)(PAC,GMM)	[0.95, 1.77]	1.12	[0.54, 1.40]	0.86
12	1	(DCTs,MLP)(SSC,GMM)	[0.84, 1.64]	1.37	[0.45, 1.19]	1.02
13	1	(DCTb,MLP)(LFCC,GMM)	[1.31, 2.62]	1.62	[0.23, 1.08]	0.58
14	1	(DCTb,MLP)(PAC,GMM)	[2.42, 3.84]	3.65	[1.41, 2.91]	2.60
15	1	(DCTb,MLP)(SSC,GMM)	[2.07, 3.43]	2.88	[1.00, 2.22]	1.55
16	2	(FH,MLP)(LFCC,GMM)	[0.34, 0.91]	0.69	[0.01, 0.64]	0.13
17	2	(FH,MLP)(PAC,GMM)	[0.53, 1.21]	1.14	[0.27, 0.98]	0.73
18	2	(FH,MLP)(SSC,GMM)	[0.50, 1.10]	0.98	[0.17, 0.83]	$*$ 0.89
19	2	(DCTb,GMM)(LFCC,GMM)	[0.00, 0.33]	0.13	[0.00, 0.38]	0.38
20	2	(DCTb,GMM)(PAC,GMM)	[0.04, 0.46]	0.18	[0.03, 0.51]	0.16
21	2	(DCTb,GMM)(SSC,GMM)	[0.01, 0.38]	0.18	[0.01, 0.51]	0.17

According to the fusion protocol, there are 21 available multimodal data sets. The HTER distribution due to random identity match is sampled 1000 times and there are 200 users. This means that the 1000 samples are a sheer portion of $1000/200! \approx 0$, i.e., one cannot possibly evaluate all the possible permutations. Table 2 lists the HTER range at 95% of confidence due to 1000 samples of random identity match (chimeric-user effect) and the corresponding HTER of true identity match. The first 15 are fusion datasets taken from LP1 while the rest are from LP2. For the values of HTER of true identity match falling outside the confidence range, a $*$ sign is marked. There are two $*$'s for the mean operator and three for the GMM.

Since Table 2 is limited to the criterion of HTER only, we also plot the whole spectrum of the so-called Expected Performance Curve (EPC) [2], which selects a different thresholds for various criteria, on a separate validation set, as follows:

$$\Delta_* = \arg\min_{\Delta} \alpha \mathrm{FAR}(\Delta) + (1-\alpha)\mathrm{FRR}(\Delta) \tag{19}$$

with α a parameter that ranges from 0 to 1. Using this threshold, the EPC then plots the corresponding HTER on the test set, with respect to α. This enables to obtain unbiased estimates of the HTER since all hyper-parameters, including the threshold, are selected on some separate validation set.

Figures 2 and 3 thus show EPC curves of the distribution due to random identity match (with a 95% confidence interval) and the EPC curve of true identity match, for the mean operator and

the GMM, respectively. As can be observed, there are much more points where the HTER of true identity match falls out of the 95% confidence range. Precisely, exactly 8/21 of experiments for the mean operator and 7/21 of experiments for the GMM. Hence, based on the available fusion datasets, about one third of them shows that the experiments with chimeric users are *inconsistent* with those carried out with the true identity match setting. Considering the fact that the mean operator has no parameters to be estimated and that the GMM has some, the free parameters in the fusion classifier *does*, to some extents, contribute to the variability observed by HTER due to the chimeric-user effect. Note that in both experiments, the 1000 random identity permutations were constrained to be the *same*. This is essential to keep the possible experiment-induced variation to be minimal.

5 Conclusions

In this paper, the following issue was addressed: "Can chimeric persons be used in multimodal biometric authentication experiments?". This topic was tackled by 1) identifying the different levels of dependence assumptions as a result of two dichotomies: feature-oriented dependence and score-oriented dependence; and 2) by experimentally comparing the effects due to using chimeric users with those using the original true modalities of same users (or simply "true users"). One major conclusion from the first approach is that the independence assumption does not imply that one can use the chimeric users in experiments. Instead, such assumption only guides how one should construct a classifier to combine information from different modalities. Neither does the second more empirical approach support the use of chimeric users. Indeed based on 21 fusion datasets and two fusion classifiers, only about two thirds of the data indicate that chimeric users can be used, or in other words, the use of true users does not vary significantly, at 95% of confidence, compared to the case when chimeric users are used in experiments. The rest of the rather large one-third of datasets suggest that the use of chimeric users cannot appropriately replace the dataset of the true modality matched dataset. Considering the high variability of HTER due to the effect of chimeric users, several runs of fusion experiments with different identity match are *strongly recommended*. Although such remedial procedure does not necessarily reflect the case when true modality matched identity is used, it *at least* gives a more accurate figure about the possible range of HTER values when the true identities are used. If the 21 fusion datasets are representative of this scenario, then, one might have a 2/3 chance of better reflecting the real HTER, after performing a large number of fusion experiments (1000 in our case!). However, one should *probably not* use the obtained HTER as a claim that the performance reflects the actual case where the real multimodal datasets are used. The current experimental approach adopted here is somewhat preliminary and in some ways limited in scope. It does not answer for instance, "how far the score distribution estimated with the independence assumption is from the one estimated with the dependence assumption?". Secondly, it does not yet answer the question: "Are the relative HTER values, in contrast to absolute values as done here (e.g., in comparing two fusion methods) *consistent* between experiments with chimeric users and those with true users?" These issues will be dealt with in the near future.

Acknowledgement

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

[1] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. In *Springer LNCS-2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03*. Springer-Verlag, 2003.

[2] S. Bengio and J. Mariéthoz. The Expected Performance Curve: a New Assessment Measure for Person Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 279–284, Toledo, 2004.

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

[4] Ruud M. Bolle, Nalini K. Ratha, and Sharath Pankanti. Error analysis of pattern recognition systems: the subsets bootstrap. *Computer Visioin and Image Understanding*, 93(1):1–33, 2004.

[5] R. Brunelli and D. Falavigna. Personal Identification Using Multiple Cues. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[6] J-L. Dugelay, J-C. Junqua, K. Rose, and M. Turk (Organizers). *Workshop on Multimodal User Authentication (MMUA 2003)*. no publisher, Santa Barbara, CA, 11–12 December, 2003.

[7] 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*, volume 5404, pages 544–554, 2004.

[8] A. Jain, K. Nandakumar, and A. Ross. Score Normalisation in Multimodal Biometric Systems. *Pattern Recognition (to appear)*, 2005.

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

[10] M. Keller, J. Mariéthoz, and S. Bengio. Significance Tests for *bizarre* Measures in 2-Class Classification Tasks. IDIAP-RR 34, IDIAP, 2004.

[11] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas. On Combining Classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[12] L.I. Kuncheva. A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(2):281–286, February 2002.

[13] J. Lüttin. Evaluation Protocol for the XM2FDB Database (Lausanne Protocol). Communication 98-05, IDIAP, Martigny, Switzerland, 1998.

[14] 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 *Proc. 15th Int'l Conf. Pattern Recognition*, volume 4, pages 858–863, Barcelona, 2000.

[15] N. Poh and S. Bengio. Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication. Research Report 04-44, IDIAP, Martigny, Switzerland, 2004. Accepted for publication in *AVBPA 2005*.

[16] N. Poh and S. Bengio. Improving Single Modal and Multimodal Biometric Authentication Using F-ratio Client Dependent Normalisation. Research Report 04-52, IDIAP, Martigny, Switzerland, 2004.

[17] A. Ross and R. Govindarajan. Feature Level Fusion Using Hand and Face Biometrics. In *Proc. SPIE Conf. on Biometric Technology for Human Identification II*, volume 5779, pages 196–204, Orlando, 2005.

[18] A. Ross, A. Jain, and J-Z. Qian. Information Fusion in Biometrics. *Pattern Recognition Letter*, 24(13):2115–2125, September 2003.

[19] K.-A. Toh, X. Jiang, and W.-Y. Yau. Exploiting Global and Local Decision for Multimodal Biometrics Verification. *IEEE Trans. on Signal Processing*, 52(10):3059–3072, October 2004.

[20] V. N. Vapnik. *Statistical Learning Theory.* Springer, 1998.

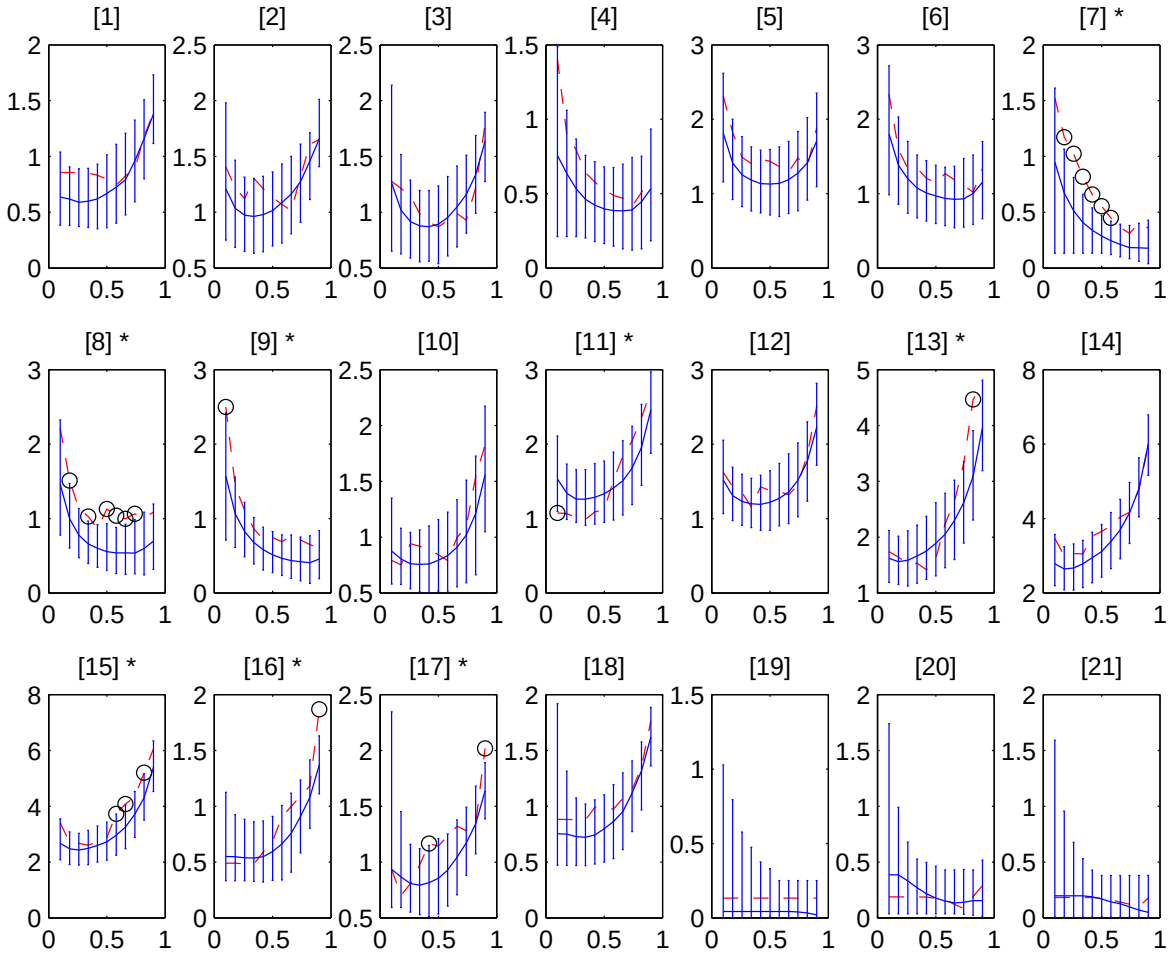

Figure 2: The EPC curve range (due to 1000 samples of random identity match) at 95% of confidence versus the EPC curve of true identity match, for each of the 21 experiments, using the *mean operator* as the fusion classifier. They are labeled accordingly from 1 to 21 corresponding to the experiment numbers in Table 2. A $*$ sign is marked for the experiments whose one or more HTERs of true identity match fall outside the confidence range. For these points, circles are plotted on the corresponding EPC curve.

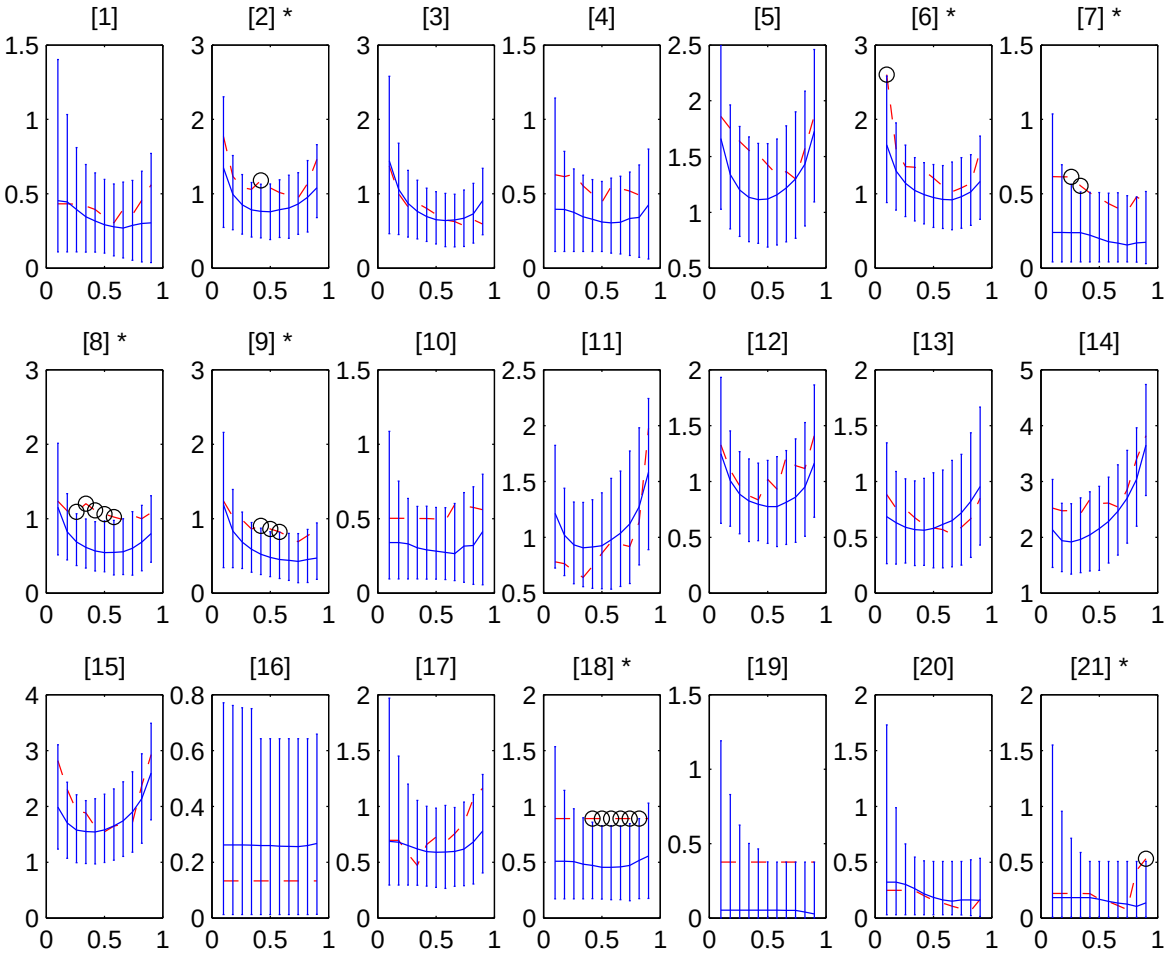

Figure 3: As per Figure 2, except that a Gaussian Mixture Model fusion classifier is used in place of the mean operator. There are 7 data sets reporting that the EER due to true identity match is *significantly* different from the EER distribution due to random identity match at 95% of confidence, contrary to 8 in Figure 2.