Multimodality in Biosecure: Evaluation on Real vs. Virtual Subjects

Bernadette Dorizzi, Sonia Garcia-Salicetti, Lorène Allano

Département Electronique et Physique
Institut National des Télécommunications
9 Rue Charles Fourier, 91011 Evry France
{Bernadette.Dorizzi, Sonia.Salicetti, Lorene.Allano}@int-evry.fr

ABSTRACT

In this paper, we present the BioSecure Network of Excellence and its objectives in terms of biometric evaluation. A particular focus is given in this project to multimodal evaluation, which requires special attention due to the lack of large-size available multimodal databases. We show in this paper that the evaluation of score fusion methods for two a priori independent modalities is possible on standard size (roughly 100 persons) virtual databases, but at the price of a careful statistical protocol.

1. THE BIOSECURE NETWORK OF EXCELLENCE: BIOMETRICS FOR SECURE AUTHENTICATION

The NoE BioSecure has been started in June 2004, in the domain of biometrics, grouping the critical mass of expertise required to promote Europe as a leading force in the field.

The main objective of this network is to strengthen and to integrate multidisciplinary research efforts in order to investigate biometrics-based identity authentication methods, for the purpose of meeting the trust and security requirements in our progressing digital information society. This goal will be attained through various integrating efforts. A common evaluation framework (such as databases, reference systems and assessment protocols) will be developed, participating to standardisation efforts. Identifying and addressing the technical challenges linked to applications will lead to the definition of joint research activities, aiming at the facilitation of the employability and practical use of the technology. The BioSecure Network of Excellence will also promote mobility and international training. A large place will be given to dissemination through large scale events (i.e. conferences, common evaluation campaigns and residential workshops). These efforts will bring together the community and will facilitate the technology transfer to the industry.

A particular effort will be made in the multimodality domain which remains an open issue due in particular to the lack of large scale available databases. Indeed, there are very few available multimodal databases (M2VTS [1], XM2VTS [2], BANCA [3], DAVID [4], SMARTKOM [5]), most of which contain only two biometric modalities, usually face and voice, and only about a hundred subjects. Also, multimodal databases more recently constructed as BIOMET [6], or under construction [7] have the tendency to contain more modalities (4 or 5) but not more subjects. This can be explained by the fact that acquiring multimodal data is more time consuming and expensive than acquiring data from a single modality, and rises some other problems as higher acquisition failure and critical personal data protection. Indeed, acquisition failure is generated because the more modalities there are, the more it is likely that a data sample cannot be acquired in a given modality, thus generating the loss of a complete multimodal sample. This phenomenon is of course amplified whenever several sessions are recorded. Also, regarding personal data protection, the fact that a data collection may contain together fingerprints, signature, iris, and face, among others, of a given person, is obviously critical and not easily acceptable for donators which can be afraid of misuse or forgeries.

2. TOWARDS VIRTUAL MULTIMODAL DATABASES

Many works in the multimodal fusion literature give results on about 100 real subjects, with no insight in the fact that such results may be in fact very biased. We first address this problem in the present work and propose a new protocol for multibiometric systems evaluation on standard size databases of real subjects.

Moreover, it is also natural to wonder about the possibility of using databases of virtual subjects, that is an individual generated by combining different biometric traits (modalities) belonging to different persons. If valid, this procedure would simplify multimodal data construction because it would be sufficient to merge two or more databases of approximately the same number of subjects, containing each specific modality, in order to generate a multimodal data corpus with more modalities. Although this question is crucial for the progress of research in multimodal fusion, few works have studied the validity of using virtual subjects for multimodal systems evaluation [8, 9, 10, 11]. The conclusions are diverse. On one hand, [8,11] motivate the possibility of using databases of virtual subjects under given conditions related to the nature of the combined modalities (fingerprint and face in [8], voice and signature in [11]), namely when the combined modalities are a priori mutually independent (in particular temporally uncorrelated). Moreover, in [11], in order to counter data sparsity due to the presence of very few real persons in the multimodal database, a specific evaluation protocol, based on a bootstrap procedure, is provided in order to exploit virtual subjects. On the other hand, [9] asserts that performance on a database of real subjects is "not equivalent" to performance on a database of virtual subjects, on a very specific and fixed experimental configuration, the XM2VTS audio-video sequences (temporally correlated modalities) database with the Lausanne standard evaluation protocol [2]. Their experimental protocol does not permit to conclude on the validity of the use of virtual subjects databases, because it uses only one split of the real database in a training and a test sets. Indeed, a single split does not permit to take into account the inherent bias due to the limited size of the database. A bootstrap procedure on the real subjects' database is necessary to conclude.

As already been said, our aim in this work is to estimate performance variability on a real subjects database of limited size by exploiting virtual subjects created from the original real subjects database. The real subjects database used is the BIOMET database [6]. This methodology permits us to do a comparative study of the behaviour of a bimodal fusion system (on-line signature and voice) on the real subjects and on several databases of virtual subjects generated from BIOMET. Indeed, the originality of this work is that we set the problem of using virtual subjects for systems evaluation relatively to the use of real subjects in multimodal databases. In fact, this methodology permits first to have a better insight into the nature of a real subjects database relatively to a virtual subjects one and, second, to enlarge the scope of our study: how evaluation should be performed in both cases.

The fusion method that we use in [11] was a training-based method, a Support Vector Classifier. The associated protocol was restricted because of the intrinsic computational load of such method. In this paper, we use a lighter but still efficient fusion method, the Arithmetic Mean Rule (AMR) after a previous normalisation of scores [12], to improve the evaluation protocols.

As mentioned above, our work is limited to two temporally uncorrelated modalities, voice and on-line signature, already combined in [12]. Of course, the choice of the modalities is a delicate question since it rises the problem of their mutual dependence/independence. We focus here in the combination of modalities that are a priori mutually independent, since it is the only framework that we may consider in order to build virtual subjects.

3. SCORE FUSION: METHOD AND PROTOCOLS

3.1 Fusion of On-line Signature and Voice

Our study considers two mono-modal biometric systems: a signature verification system and a text-independent Speaker Verification system, both described in [13]. The scores provided by each system are first normalised before being averaged. The normalisation used is a rescaling of scores in the [0, 1] interval using the Min-Max rule [8]. The aim of such normalisation is to obtain two comparable scores in order to combine them in an efficient way by the AMR. This fusion method avoids a time-consuming learning phase, but still requires a dedicated development set of scores to compute the normalisation factors.

We build what we call a bimodal database through the association of the scores of the two experts (Signature and Voice). For each person, we have at disposal 4 or 5 bimodal client accesses and in average 10 bimodal impostor accesses (this number varies across persons from 6 to 12 impostor accesses).

3.2 The evaluation protocol on BIOMET

This bimodal database of 77 persons is then split in 2 subsets (of 38 and 39 persons): one is devoted to the computation of the normalisation parameters, named *FLB* (Fusion Learning Base), and the other is devoted to testing purposes, named *FTB* (Fusion Test Base). Data sparsity due to the presence of only 77 real subjects in BIOMET generates a bias effect if we consider only one split. To counter this phenomenon, we consider 100 different splits.

We have chosen a Cross-Validation (CV) procedure because it permits to obtain results on the entire database instead of only on a predefined test subset. We consider a 2-fold Cross-Validation (CV) protocol that consists in using first the 2 subsets (*S1* and *S2*) as (*FLB=S1, FTB=S2)* and then to interchange their roles, that is (*FLB=S2, FTB=S1)*. After these 2 steps, we compute an error rate per type (FAR, FRR) and per split. This process is repeated for each value of the decision threshold, leading to a DET curve for each split.

As shown in **Fig. 1**, error rates are averaged on the 100 splits generated, in order to obtain a Real Mean DET Curve (RMDC). Standard deviations of error rates are also estimated to represent the variance among 100 different couples of *(FLB, FTB)*.

3.3 Creation of virtual subjects with BIOMET

We create a virtual subject by pairing randomly signature data of a given subject to the speech data of another subject. We chose in this work to create up to 1000 data sets of virtual subjects, as in [8, 11].

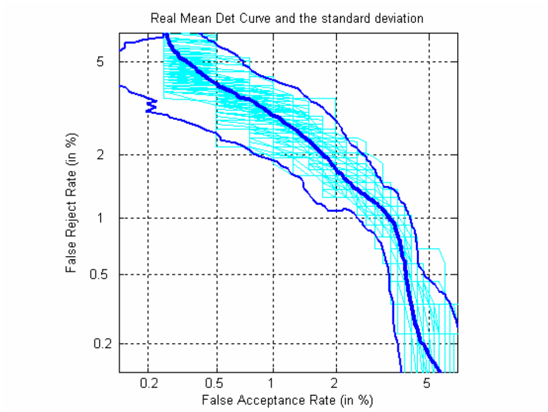

Fig. 1. 100 splits of 2-fold CV for the real subjects Database with the mean and the mean ±2*standard-deviation of errors.

For every database of virtual subjects, we exploit exactly the same protocol as the one used on the database of real subjects: in other words, we perform 100 different splits in two sets (38 and 39 persons) and a 2-fold CV on each split; we then average the resulting 100 error rates for each value of the decision threshold and obtain a Virtual Mean DET Curve (VMDC) for each virtual database.

4. COMPARATIVE FUSION EXPERIENCES ON REAL AND VIRTUAL SUBJECTS

We perform experiments to compare the behaviour of the fusion system on the real database (BIOMET) and the virtual databases built from the same persons. In **Fig. 2**, we compare the Real Mean DET Curve (RMDC) obtained on the BIOMET database to 1000 Virtual Mean DET Curves (VMDC) associated to 1000 databases of virtual subjects. As explained in sections 3.2 and 3.3, each curve represents average error rates over 100 different splits and a 2-fold CV on each split. We also represent in **Fig. 2**, the Average of Virtual Mean DET Curves (AVMDC) over the 1000 databases of virtual subjects.

Fig. 2 shows that the Real Mean DET Curve (RMDC) is inside the band generated by the 1000 Virtual Mean DET Curves (VMDC). In particular, for values of the threshold close to the Equal Error Rate (EER) point, this curve is close to the Average Virtual Mean DET Curve (AVMDC), while for other values of the decision threshold it behaves as any other of the 1000 VMDC. This first result permits to conclude that the system behaves on the database of real subjects (when averaging error rates on 100 splits) as on any of the databases of virtual subjects (with the same CV protocol). This also supports the mutual independence assumption between the two modalities that we consider on-line signature and voice.

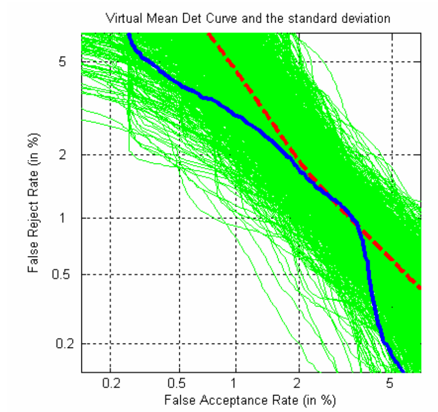

Fig. 2. Real Mean DET Curve (boldface), 1000 Virtual Mean DET Curves and their average (dashed-line).

Moreover, this suggests that the use of virtual subjects sets permits to have an estimation of performance variability, providing in fact a "confidence interval" for performance obtained on a real subjects data set of standard size (100 persons). In other words, the database of real subjects is a data set with an inherent bias. This bias is greatly increased if a single split in a Fusion Learning and Fusion Test Bases *(FLB,FTB)* is considered, like widely done in the literature, in particular in [9,10]. Indeed, the statistics of bimodal data found in the test set (represented by the real subjects present in such set) may be very different from that present in the training set, leading this way to an unreliable and misleading evaluation of the fusion system. It is thus necessary to perform different splits that correspond to different individuals in *FLB* and *FTB* respectively, to apply on each a Cross-Validation protocol and to average error rates over those splits.

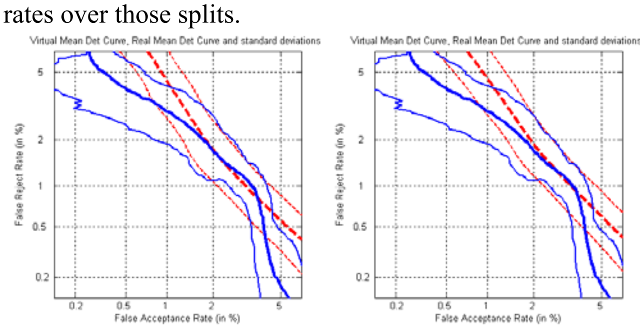

Fig. 3. RMDC vs. AVDMC and their standard deviation for 100 (left) and 1000 (right) virtual databases

We now report in **Fig. 3**, the Average of Virtual Mean DET Curves (AVMDC) with the Real Mean DET Curve and the average of standard deviations for 1000 virtual subjects databases and for the real subjects database due to 100 splits. We consider in **Fig. 3** respectively 100 (left) and 1000 (right) virtual subjects databases. We notice that the mean curves have the same behaviour around the EER point, in terms of mean but also of standard deviation. This shows that when isolating 2 different VMDC, the performance will differ from each other and of course from the RMDC (because of the small size of BIOMET), but the variance is of the same order of magnitude. **Fig. 3** shows that the same result is observed for 100 and 1000 virtual subjects databases.

5. CONCLUSIONS

We have studied in this work the problem of evaluation of score fusion algorithms on relatively small size real-person databases of bimodal score values, as well as the question of using virtual persons (built through different pairings of the mono-modal scores) instead of real ones. The data at our disposal comes from 77 subjects of the BIOMET database and we considered two a priori independent modalities: on-line signature and speech. 1000 databases of virtual subjects were constructed from BIOMET bimodal data. Our first conclusion is that, if the multibiometric system is evaluated with a precise statistical protocol based on Cross-Validation, a standard size database (about 100 subjects) of real subjects behaves as any virtual subjects set of the same size in terms of mean and standard deviation of error rates. This of course supports the mutual independence assumption of the two biometric traits that we consider. In other words, this confirms a natural intuition that a database of real subjects has an inherent bias. Therefore, considering only one split in a learning set and a test set does not permit a reliable evaluation of the multibiometric system, even if the database is real! To cope with this fact, we propose a protocol for multibiometric systems evaluation on standard size databases (about 100 subjects) of real subjects, consisting in creating several splits of the data set in a Fusion Learning Base and a Fusion Test Base and performing a 2-fold Cross-Validation on each split; then error rates are averaged over such 100 splits for each value of the decision threshold. We can conclude that, in the case of mutual independence of the modalities that are considered, the use of virtual subjects with the protocol above given is a powerful tool to estimate the performance variability, providing a “confidence interval” for performance obtained on a real subjects data set of standard size (100 persons). It is thus recommended to create virtual sets from real ones for a complete and reliable evaluation of multibiometric systems on real multimodal databases of standard size.

Further work will be focused on the study of correlation of the combined modalities through the use of correlation measures. Moreover, we will consider other configurations of data and classifiers.

6. ACKNOWLEDGEMENTS

The authors would like to thank G. Chollet (ENST) and his team for having provided the scores of the speech verification system.

This work was partially funded by the IST-FP6 project BioSecure.

7. REFERENCES

[1] S. Pigeon & L. Vandendorpe, "The M2VTS Multimodal Face Database", In Proceedings of AVBPA 97, Springer LNCS, Bigün et al. Eds, 1997.

[2] K Messer J Matas J Kittler, J Luettin and G Maître, “XM2VTSDB: The Extended M2VTS Database”, Second International Conference on Audio and Video-based Biometric Person Authentication, 1999.

[3] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, Jo. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruíz, and J.-P. Thiran. “The BANCA Database and Evaluation Protocol”, Audio- and Video-Based Biometric Person Authentication (AVBPA), Guilford, 2003.

[4] J.S.D Mason, F. Deravi, C. Chibelushi & S. Gandon, “BT DAVID - Final Report”, Speech and Image PDETessing Research Group, Dept. of Electrical and Electronic Engineering, University of Wales Swansea, UK.

[5] http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html

[6] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux-Les Jardins, J. Lunter, Y. Ni, D. Petrovska-Delacretaz, "BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", Proc. of 4th International Conference on Audio and Video-Based Biometric Person Authentication, pp. 845-853, Guildford, UK, July 2003.

[7] A. Ross, A. Jain, "Information Fusion in Biometrics", Pattern Recognition Letters 24, pp. 2115-2125, 2003.

[8] M. Indovina, U. Uludag, R. Snelick, A. Mink, A. Jain, “Multimodal Biometric Authentication Methods : A COTS Approach”, Workshop on Multimodal User Authentification (MMUA), pp. 99-106, Santa Barbara, California, USA, Dec. 2003.

[9] N. Poh, S. Bengio, “Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments?”, in Proc. of MLMI 2005, 11-13 July, Edinburgh, UK.

[10] N. Poh, S. Bengio, “Using Chimeric Users to Construct Fusion Classifiers in Biometric Authentication Tasks: An Investigation”, IDIAP-RR 05-59, 2005.

[11] S.Garcia-Salicetti, M.A.Mellakh, B. Dorizzi, “A generic protocol for multibiometric systems evaluation on virtual and real subjects”, in Proc. of AVBPA 2005, pp. 494-502, Hilton Rye Town, NY, USA, July 20-22, 2005.

[12] S. Garcia-Salicetti, M.A. Mellakh, L. Allano, B. Dorizzi, “Multimodal Biometric Score Fusion: the Mean rule vs. Vector Support Classifiers”, in Proc. of EUSIPCO 2005, Antalya, Turkey, September 4-8, 2005.

[13] B. Ly Van, R. Blouet, S. Renouard, S. Garcia-Salicetti, B. Dorizzi, G. Chollet : “Signature with text-dependent and text-independent speech for robust identity verification”, in Proc. MMUA 2003, Santa Barbara, California, USA, Dec. 2003, pp. 13-18.