SOME RESULTS FROM THE BIOSECURE TALKING FACE EVALUATION CAMPAIGN

Benoît Fauve[1], Hervé Bredin[2], Walid Karam[2,3], Florian Verdet[4,5], Aurélien Mayoue[6], Gérard Chollet[2], Jean Hennebert[4,7], Richard Lewis[1], John Mason[1], Chafic Mokbel[3], Dijana Petrovska[6]

[1]Swansea University, UK, [2]GET-ENST, Dept. TSI, Paris, France, [3]University of Balamand, Lebanon, [4]University of Fribourg, Switzerland, [5]LIA, Université d'Avignon, France, [6]GET-INT, Evry, France, [7]HES-SO Sierre, Switzerland

ABSTRACT

The BioSecure Network of Excellence[1] has collected a large multibiometric publicly available database and organized the BioSecure Multimodal Evaluation Campaigns (BMEC) in 2007[2]. This paper reports on the Talking Faces campaign. Open source reference systems were made available to participants and four laboratories submitted executable code to the organizer who performed tests on sequestered data. Several deliberate impostures were tested. It is demonstrated that forgeries are a real threat for such systems. A technological race is ongoing between deliberate impostors and system developers.

Index Terms— Talking Face Biometrics, Robustness, Impostures, Open Evaluation

1. INTRODUCTION

The BioSecure NoE groups about 30 academic and some industrial laboratories interested in biometrics. It was initiated in June 2004 and held its final review last september 2007. Among its achievements, three large multibiometric databases were collected across Europe over a period on nine months. The BioSecure Multimodal Evaluation Campaigns (BMEC) 2007[2] were held to promote original research work while comparing it with open-source state-of-the-art or baseline reference systems.

This paper focusses on several aspects (from data collection to results) of the Talking Faces campaign. During the BioSecure research program one of the focal points was to develop open source reference systems to limit the repeated efforts of many institutions to catch up with the ever changing state of the art (for example [1]). In the BMEC talking face task, four laboratories submitted executable code to the organizer (University of Fribourg) who performed tests on sequestered data. Several deliberate impostures were developed and tested. In the evaluation results, it is demonstrated that forgeries are a real threat for such systems. A technological race is ongoing between deliberate impostors and system developers.

The outline of this paper is as follows. In Section 2 we present BioSecure databases and BMEC. Section 3 describes the talking face task in BMEC and the different kinds of forgeries. Participants and reference systems are described in Section 4. We present and discuss results in Section 5. Conclusion and perspective are drawn in Section 6.

[1]http://www.biosecure.info/

[2]http://www.int-evry.fr/biometrics/BMEC2007/

2. BIOSECURE DATABASES AND BMEC

A major innovation targeted by the BioSecure project was to acquire a large-scale multimodal database that would include various but realistic recording scenarios, using different kinds of devices and providing reference systems for each of the modalities. This database aims at helping the research community to build reliable biometric-based security systems that can be improved in terms of their accuracy, scalability, robustness to device-dependent data and various environments. Two sessions separated by about one month interval were recorded and three different datasets were acquired by 11 university institutes across Europe:

Internet Dataset: still face images and talking-face recorded through the Internet and under uncontrolled situations. About 1000 volunteers have participated in 2 sessions.

Desktop Dataset: laboratory database with (high/low quality) 2D face, iris, talking-face, signature, (high/low quality) fingerprint and hand modalities. It is PC-based, off-line and supervised data acquisition. About 600 donors were acquired in 2 sessions.

Mobile Dataset: mobile devices under degraded conditions were used to build this dataset. 2D face and talking-face data were acquired in both indoor and outdoor environments. Signature and fingerprint modalities were acquired using the sensors of a PDA. About 700 donors have participated in 2 sessions.

The BioSecure Multimodal Evaluation Campaign (BMEC) has been launched in March 2007 to enable institutions to easily assess the performance of their own monomodal and multimodal algorithms and to compare them to others. At the same time a variety of open-source reference systems were made available online to help every participant site with their development (available on BMEC website).

Two different scenarios have been identified for this evaluation: an access control scenario on Desktop Dataset and a mobile scenario on degraded data from Mobile Dataset. The talking-face experiment is part of the mobile scenario.

3. TALKING FACE PROTOCOL AND FORGERIES

3.1. Material for talking face

The video sequences for the talking face evaluation are recorded with a laptop computer and a webcam. For each individual, there is a total of 4 recordings used: 2 indoor from a first session, 2 outdoor from a second session ($\sim$1 month later). A video is around 10s long and individuals say a random and different English phrase in each of the 4 recordings. The actual speech length is 3s on average.

For the protocol the two first recordings are used for training. In the testing phase the client accesses come from the second session,

leading to a total of 1720 client accesses (430 speakers, 2 models per speaker from the indoor set, each model tested against 2 videos from the outdoor set). The number of impostor accesses depends on the forgery scenarios which are described in the following section.

3.2. Forgeries

The systems are challenged against different types of forgeries ranging from the simplest random forgeries to more sophisticated ones.

imp1RND *Random forgeries*. These forgeries are simulated by using video sequences from other users when testing on a specific user model. This category actually does not denote intentional forgeries, but rather accidental accesses by non-malicious users. For the BMEC evaluation, we use the video files from 10 other users taken randomly from the database, leading to 17200 impostor accesses.

imp2CT *Genuine picture animation*. In this scenario the forger has captured a static picture of the genuine user and then uses commercial software to simulate a talking face (see for example "crazytalk" tool[3]). The procedure is illustrated in Figure 1. The speech part was automatically generated using a freely available text-to-speech (TTS) system[4,5]. The gender of the voice is chosen according to the gender of the user to forge. After manual annotation work on the picture to mark the lips and the face positions, the fake video sequence is automatically generated by the software by moving the face according to the sound waveform. One imposture was produced for each user resulting in 430 face animation forgeries for that scenario; these are presented to two models each time, leading to 860 impostor accesses.

imp3PP *Genuine picture presentation*. Here the forger moves a static picture of the targeted user in front of the camera to attempt to break the liveness detection system, if any. For practical reasons, software is used to automatically produce video files (see Figure 2). For the speech part, here again a gender dependent TTS is used. 1720 impostor accesses are available for this condition.

imp4AR *Audio replay attack*. For this scenario the audio from the target user is played back to the system while the forger moves his lips (un-synchronously). In practice, the impostor access uses speech from the outdoor session of the targeted speaker and the video from someone else. 1720 impostor accesses are available for this condition.

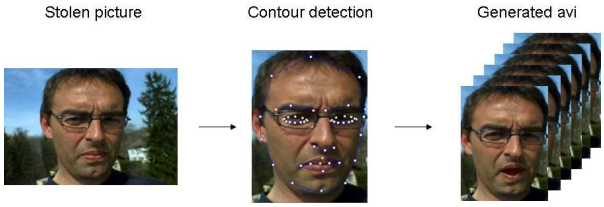

Fig. 1. Genuine picture animation scheme.

3.3. Development data

All participants of the evaluation received a development set of 200 videos from 50 individuals not in the evaluation set. No forgery examples were distributed.

[3]http://crazytalk.reallusion.com/
[4]http ://www.itl.nist.gov/iad/894.03/fing/fing.html
[5]http ://www.cstr.ed.ac.uk/projects/festival/

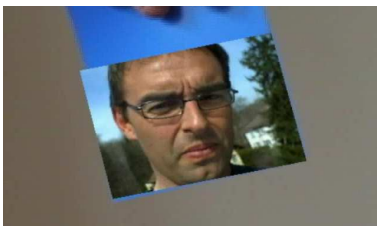

Fig. 2. Frame example from genuine picture presentation forgeries.

4. REFERENCE AND SUBMITTED SYSTEMS

In this section we describe the systems from the 4 participants as well as a BioSecure reference system.

4.1. Reference System

The BioSecure reference system is based on the fusion of face and speaker verification scores.

Face verification. It is based on the standard Eigenface approach [2] to represent face images in a lower dimensional subspace. Firstly, 10 frames are extracted from the video at regular intervals. Using the eye positions, each face image is normalized, cropped and projected onto the face space (the face space was built using the 300 images from the BANCA world model and the dimensionality of the reduced space was selected such as 99 per cent of the variance is explained by the PCA analysis). In this way, 10 feature vectors are produced for a given video. Next, the L1-norm distance measure is used to evaluate the similarity between 10 target and test feature vectors. Finally, the face score is the minimum of these 100 distances.

Speaker verification. It is developed using HTK[6] and BECARS[7] open source toolkits. The speech processing is performed on 20ms Hamming windowed frames, with 10ms overlap. For each frame, 15 MFCC coefficients (+energy) and their first-order deltas are extracted. For speech activity detection, a bi-Gaussian model is fitted to the energy component of a speech sample. The threshold t used to determine the set of frames to discard is computed as follows: $t = \mu - 2 * \sigma$, where μ and σ are the mean and the variance of the highest Gaussian component, respectively. Next, a cepstral mean substraction (CMS) is applied to the static coefficients.

A universal background model (UBM) with 256 components has been trained with the EM algorithm using all genuine data of the development database. A speaker model is built by adapting the parameters of the UBM using the maximum a posteriori (MAP) criterion. The speech score is the average log-likelihood ratio being a target model.

Fusion module. The min-max approach [3] is used to fuse the face and speech scores. The fusion parameters have been estimated using all development data.

4.2. GET-ENST

Systems submitted by GET-ENST are based on the weighted sum of normalized speaker, face and client-dependent synchrony verification scores S_{s}, S_{f} and S_{c}. Weights and σ/μ normalization coefficients are estimated on the BMEC development set.

Face verification. Once face detection is applied on each frame of the video sequence (using *Fasel et al.*'s algorithm [4]), *distance*

[6]http://htk.eng.cam.ac.uk/
[7]http://www.tsi.enst.fr/becars/

from face space (DFFS) is computed for every detected face as the distance between the face and its projection on the face space (obtained via principal component analysis) [2]. We define a *reliability* coefficient r as the inverse of the DFFS ($r = 1/\text{DFFS}$): the higher, the more reliable. Finally, a detected face is kept as correct if its r coefficient is higher than a threshold $\theta_r = 2/3 \cdot r_{\max}$, where $r_{\max}$ is the maximum value of r on the current video sequence. Only *eigenface* features corresponding to a correctly detected face are kept to describe the face appearing in the video sequence. Finally, at test time, the *Mahalanobis* distance is computed between the *eigenface* features (of dimension 100, in our case) of each of the N correctly detected faces of the enrollment video sequence and each of the M correctly detected face of the test video sequence, leading to $N \times M$ distances. The negative of the mean of these $N \times M$ distances is taken as the score S_{f} of the face verification module.

Speaker verification. The speaker verification module is similar to the one used in the reference system. The only difference is in the extracted features: 12 MFCC coefficient with first and second order deltas

Client-dependent synchrony measure. The client-dependent measure of the synchrony between acoustic and visual speech features is a new biometric combined feature. All details of implementation for BMEC evaluation can be found in [5].

4.3. Balamand

The Balamand system uses both the speech and the visual modalities for speaker verification. On the visual side, faces are tracked in every frame in the video sequence through a machine learning approach based on a boosted cascade of Haar-like features for visual object detection [6]. Faces are then scaled, cropped, gray-scaled, and histogram equalized. Feature extraction is based on orthogonal 2-D DCT basis functions of overlapping blocks of the face [7]. On the speech side, the feature extraction module calculates relevant vectors from the speech waveform. On a signal "FFT" window shifted at a regular rate, cepstral coefficients are derived from a filter bank analysis with triangular filters. A Hamming weighting window is used to compensate for the truncation of the signal. The toolkit SPro[8] is used.

Classification for both modalities uses GMMs to model the distribution of the feature vectors for each identity. GMM client training and testing is performed on the speaker verification toolkit BECARS[7].

A final decision on the claimed identity of a talking face relies on fusing the scores of both modalities. The speech and face scores are personalized (z-norm) with mean and variance estimated on the development set.

4.4. Swansea

The Swansea system is a speech only system based on an LFCC front-end and a GMM system for speaker adaptation and testing [8]. It was developed using SPro[8] and ALIZE[9] open source toolkits. The GMM system is as described in [9] and the front-end is an adaptation from the mean-based feature extraction described in [10], found to perform well on short duration tasks. A UBM with only 64 components is trained from all development data. Score normalization is applied with a T-norm cohort made of 100 models coming from development data. All details not mentioned here can be found in [9].

[8] http://gforge.inria.fr/projects/spro/

[9] http://www.lia.univ-avignon.fr/heberges/ALIZE/

The systems in previous publications have been optimized on NIST speaker recognition evaluation (SRE) databases[10], where the recordings come from telephony speech sample at 8 kHz, but in BMEC talking face evaluation the acoustic signal is sampled at 44.2 kHz. In [10] the features are calculated from 24 filterbanks taken between 300Hz and 3.4kHz (telephony band). After experiments on the development data, we retain a front-end where the filterbank width is kept similar to the original configuration by considering 72 linear bands between 300Hz and 12kHZ. Out of the potential 72 LFCC coefficients, we take only the first 29. The new feature size is 59 with 29LFCC+29deltas+delta Energy.

4.5. UNIFRI

The system presented by UNIFRI is modeling independently the speech and face modalities, performing a fusion at the score level. In a similar way as the Balamand system, both modalities use GMMs trained with a MAP adaptation procedure from UBMs.

For the face part, the same face detection procedure as for the reference system is used. Face images are extracted every second from the video sequence. Cropping, resizing to 120×160 pixels, gray level transformations and intensity/contrast normalization are applied consecutively to each face image. A DCTmod2 feature extraction is then applied on 15×15 pixels windows shifted along the x and y axis (50% overlapping) [7]. The feature vectors are then composed from the 25 first DCT coefficients from which the three first ones are replaced by delta values computed from the adjacent windows along the x and y directions. The background GMM is trained using parts of the data of all genuine users from the development set, including indoor and outdoor conditions. To make the impact of illumination uniform, all face images are horizontally flipped. The EM algorithm is used to train this model up to 128 Gaussians using a binary splitting procedure. Genuine models are obtained using a MAP adaptation from this UBM.

For the speech part, the feature extraction is classically based on MFCC features with 13 coefficients. Delta features are not included. A speech activity detection module based on a bi-Gaussian model is used to discard the silence part of the speech signal. The speech detection parameters (essentially the threshold) have been tuned on the development data. Similarly as for the face part, a 32 component background GMM is trained on the development set using the EM algorithm. Genuine models are obtained using a MAP adaptation from this background model.

Log-likelihood ratio scores are computed from the face and speech part and are normalized using a z-norm procedure where normalization coefficients are computed using a cohort composed of users from the development set [11]. A simple sum of normalized scores for each modality is used, without any weighting.

5. RESULTS AND FUSION

5.1. BMEC talking face official results

Table 1 shows the official results for BMEC talking face task in terms of equal error rate (EER). Considering the little amount of development data available, the general level of performance on the usual impostor attack (random imp1RND) is quite good. Noticeably Swansea speech-only system shows promising results when compared to EERs usually observed on (telephony) NIST short duration tasks [10]. Optimization of the speech front-end seems to be fundamental on the BMEC type of data.

[10] http://www.nist.gov/speech/tests/spk/

Table 1. *BMEC Talking Face EERs (%). For each imposture type, the 3 best systems are presented in bold.*

	imp1 RND	imp2 CT	imp3 PP	imp4 RP
RefSys	24.8	**34.9**	35.4	40.3
GET3-speech/face	**21.1**	35.8	35.3	**37.2**
GET6-face only	28.7	47.7	46.7	**29.2**
GET8-synchrony	43.9	43.6	39.6	44.2
Balamand	**19.4**	**27.7**	**22.3**	42.1
Swansea	**16.1**	**16.6**	**14.7**	50.5
UNIFRI	23.3	40.1	**25.5**	**37.2**

Exploitation of the face information proved difficult. Here the adopted PCA approaches proved inadequate to cope with the huge illumination and expression variability.

The GET8 system, which uses new techniques introduced in [5], gave overall poor results. One reason could be the problem of lack of synchrony between video and speech while recording on a mobile device. In general there is very little resistance to forgeries. If a system resists well to a type of imposture (eg speech only Swansea system on imp2CT, GET6-face on imp4RP) it obtains poor results on an other (Swansea on imp4RP, GET6-face on imp2CT).

Given the Swansea system provides good results from the broad band acoustic signal, an interesting option now is post evaluation fusion, described below.

5.2. Post-eval fusion

This fusion is done by logistic regression using the FoCal toolkit[11] with weights learnt on the development data (no optimization on BMEC evaluation results).

Table 2. *Combinations of Swansea speech and a GET3 talking+face and GET6 face only systems, in EER (%).*

	imp1RND	imp2CT	imp3PP	imp4RP
Swan+GET3	12.8	17.9	18.6	43.1
Swan+GET6	13.2	19.9	18.6	40.1

Table 2 presents the combinations of Swansea speech system and a GET talking+face and face only systems. For imp1RND scenario the overall improvement on the best single system from BMEC evaluation is quite significant ($\sim 16\%$ to $\sim 13\%$ EER). On a more negative note the imp4AR (audio replay) imposture still results in really poor performance.

6. CONCLUSIONS AND PERSPECTIVES

In this paper we report the effort of BioSecure NoE members on acquisition of data, definition of protocol and forgeries for talking face biometrics. Despite a relatively small number of participants some original work has been submitted. For practical reasons research is usually judged on random access type of imposture as they are directly available from other genuine accesses (passive impostor access). For such a traditional scenario our results show some relatively good performance. But we also show that deliberate well thought impostures are a real threat for state-of-the-art systems. There is a call here for more research work on forgery scenarios in general.

Now that the BioSecure program is over, a non profit organization called Association BioSecure has been set up to carry on the effort and eventually organize other evaluations on the collected databases. In this framework there are strong possibilities of organizing a BMEC-2009 with a workshop in Alghero during ICB (International Conference on Biometrics) 2009. Building more impostor resistant systems for audio-visual biometrics will definitely be a strong theme for such a new campaign.

[11] www.dsp.sun.ac.za/ nbrummer/focal/

7. ACKNOWLEDGMENTS

BioSecure is a project of the 6th Framework Programme of the European Union. We thank in particular all the sites which participated in the hard task of data collection, annotation and preparation for the evaluation campaign.

8. REFERENCES

[1] H. Bredin, G. Aversano, C. Mokbel, and G. Chollet, "The Biosecure Talking-Face Reference System," in *Proc. Second Workshop on Multimodal User Authentication (MMUA'06)*, 2006.

[2] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3(1), pp. 77–86, 1991.

[3] A. K. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[4] I. Fasel, B. Fortenberry, and J. Movellan, "A Generative Framework for Real Time Object Detection and Classification," *Comput. Vis. Image Underst.*, vol. 98, no. 1, pp. 182–210, 2005.

[5] H. Bredin and G. Chollet, "Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification," in *Proc. ICASSP*, 2007.

[6] R. Lienhart and J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection," *IEEE ICIP*, vol. 1, pp. 900–903, 2002.

[7] C. Sanderson and K. K. Paliwal, "Fast Feature Extraction Method for Robust Face Verification," *IEE Electronics Letters*, vol. 38 (25), pp. 1648–1650, 2002.

[8] D.A. Reynolds, T.F. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[9] J-F. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf, "NIST04 Speaker Recognition Evaluation Campaign: New LIA Speaker Detection Platform Based on ALIZE Toolkit," NIST SRE'04 Workshop. Toledo, Spain, June 2004.

[10] B. Fauve, N. W. D. Evans, and J. S. D. Mason, "Improving the Performance of Text-Independent Short Duration GMM- and SVM-Based Speaker Verification," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2008.

[11] C. Barras and J.-L. Gauvain, "Feature and Score Normalization for Speaker Verification of Cellular Data," in *Proc. ICASSP*, 2003.