INVESTIGATING FEATURE-LEVEL FUSION FOR CHECKING LIVENESS IN FACE-VOICE AUTHENTICATION

Girija Chetty and Michael Wagner
HCC laboratory, School of ISE, University of Canberra, Australia
Girija.chetty@canberra.edu.au

ABSTRACT

In this paper we propose a feature level fusion approach for checking liveness in face-voice person authentication. Liveness verification experiments conducted on two audiovisual databases, VidTIMIT and UCBN, show that feature-level fusion is indeed a powerful technique for checking liveness in systems that are vulnerable to replay attacks, as it preserves synchronisation between closely coupled modalities, such as voice and face, through various stages of authentication. An improvement in error rate of the order of 25-40% is achieved for replay attack experiments by using feature level fusion of acoustic and visual feature vectors from lip region as compared to classical late fusion approach.

1. INTRODUCTION

By using multiple cues concurrently for authentication, systems gain more immunity to intruder attacks [1], as it will be more difficult for an impostor to impersonate another person with multiple cues, such as audio and visual cues simultaneously. In addition, multiple cues such as those from face and voice, also help improve system reliability. For instance, while background noise has a detrimental effect on the performance of voice biometrics, it does not have any influence on face biometrics. On the other hand, while the performance of face recognition systems depends heavily on lighting conditions, lighting does not have any effect on the voice quality [2].

However, current audiovisual authentication systems mostly verify a person's face statically, and hence these systems remain vulnerable to replay attacks that present pre-recorded audio together with a still photograph. To resist such attacks, audiovideo authentication should include verification of the "liveness" of the audiovideo data presented to the system [3]. Until now, although there has been much published research on the liveness, for example, of fingerprints, research on live-ness verification in audiovisual authentication systems has been very limited.

Moreover, the classical approaches to multimodal fusion, late fusion and its variants in particular, have been investigated in great depth. Late fusion, or fusion at the score level, involves combining the scores of different classifiers, each of which has made an independent decision. This means, however, that many of correlation properties of the joint audiovideo data are lost. Fusion at feature-level on the other hand, can substantially improve the performance of the multimodal systems as the feature sets provide a richer source of information than the matching scores, and because in this mode, features are extracted from the raw data and subsequently combined. Feature-level fusion allows synchronisation between closely coupled modalities such as voice and lip-movements to be preserved throughout various stages of authentication, facilitating liveness verification in systems that would otherwise be more vulnerable to replay attacks.

This paper proposes feature-level audiovisual fusion as a powerful technique for checking liveness for face-voice authentication. The remainder of this paper is organised as follows. The next section details the speaking face data used for different experiments. The description of feature extraction method used is given in section 3. The examination of system performance for feature fusion and late fusion, in the presence of adverse environmental conditions such as acoustic noise and visual artefacts, as well as sensitivity to training data size and utterance length are discussed in section 4, followed by concluding remarks in section 5.

2. SPEAKING FACE DATA

We used two different types of speaking face data to evaluate proposed feature-level fusion. The first database used for evaluation is the multimodal person authentication database VidTIMIT [4]. The VidTIMIT database consists of video and corresponding audio recordings of 43 people (19 female and 24 male). The mean duration of each sentence is around 4 seconds, or approximately 100 video frames. A broadcast-quality digital video camera in a noisy office environment was used to record the data. The video of each person is stored as a sequence of JPEG images with a resolution of 512 × 384 pixels (columns × rows), with corresponding audio provided as a monophonic, 16 bit, 32 kHz PCM file.

The second type of data used is the UCBN database, a free to air broadcast news database. The broadcast news is a continuous source of video sequences, that can be easily obtained or recorded, and has optimal illumination, colour, and sound recording conditions. However, some of the attributes of broadcast news database such as near-frontal images, smaller facial regions, multiple faces and complex backgrounds require an efficient face detection and tracking scheme to be used.

0-7803-9243-4/05/$20.00 ©2005 IEEE

The database consists of 20-40 second video clips for anchor persons and newsreaders with frontal/near-frontal shots of 10 different faces (5 female and 5 male). Each video sample is a 25 frames per second MPEG2 encoded stream with a resolution of 720 × 576 pixels, with corresponding 16 bit, 48 kHz PCM audio.

These two types of databases represent very different types of speaking face data, VidTIMIT with original audio recorded in a noisy office environment and clean visual environment, and UCBN with clean audio and visual environments, but complex backgrounds. This allows the robustness of feature-level fusion to be examined accurately in our study. Figure 1(a) and 1(b) show sample speaking-face data from the VidTIMIT and UCBN databases.

Fig. 1: Faces from (a) VidTimit (b) UCBN

3. AUDIOVISUAL FUSION

3.1 Acoustic feature extraction

The mel frequency cepstral coefficients (MFCC) as derived from the cepstrum information were used for extracting acoustic features. The pre-emphasized audio signal was processed using a 30ms Hamming window with one-third overlap, yielding a frame rate of 50 Hz, to obtain the MFCC acoustic vectors. An acoustic feature vector was determined for each frame by warping 512 spectral bands into 30 mel-spaced bands, and computing the 8 MFCCs. Cepstral mean normalization was performed on all MFCCs before they were used for training, testing and evaluation. Before extracting MFCCs, the audio files from the two databases were mixed with factory noise (Factor1.wav) from the Noisex-92 database [5] at a signal-to-noise ratio of 6 dB. Channel effects with a telephone line filter were then added to the noisy PCM files to simulate the channel mismatch.

3.2 Visual feature extraction

Before the facial features can be extracted, faces need to be detected and recognised. The face detection for video was based on the approach of skin colour analysis in red-blue chrominance colour space, followed by deformable template matching with an average face, and finally verification with rules derived from the spatial/geometrical relationships of facial components, [6]. The lip region was determined using derivatives of the hue and saturation functions, combined with geometric constraints. Figures 2(a) to 2(e) show some of the results of the face detection and lip feature extraction stages. The details of the scheme are described in [6].

Similarly to the audio files, the video data in both databases were mixed with artificial visual artefacts such as addition of Gaussian blur and Gaussian noise, using a visual editing tool [Adobe Photoshop]. The "Gaussian Blur" of Photoshop was set to 1.2, and "Gaussian Noise" of Photoshop to 1.6.

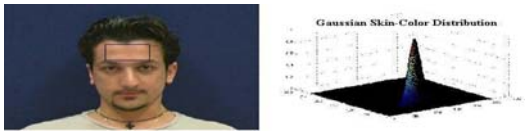

Fig. 2(a): A skin colour sample and the Gaussian distribution in red-blue chromatic space

Fig. 2(b): Lip region localisation using hue-saturation thresholding and detection of lip boundaries using pseudo-hue/luminance images

Fig. 2(c): Illustration of geometric features and measured key points for different lip openings

3.3 Joint Audiovisual Feature Vector

To evaluate the power of the feature-level fusion strategy in preserving the audiovisual synchrony, and hence verification of liveness, experiments were conducted with both feature fusion (also referred to as early fusion) and late fusion of audiovisual features. In case of feature fusion, the audiovisual fusion involved a concatenation of the audio features (MFFCs-8) and visual features (eigen-lip projections(10)+ lip dimensions(6)), and the combined feature vector was then fed to a GMM classifier. The audio features acquired at 50 Hz, and the visual features acquired at 25Hz were appropriately rate interpolated to obtain synchronized joint audiovisual feature vectors. For late fusion, audio and visual features were fed to independent GMM classifiers and the weighted scores (β) [10] from each stage, were fed to a weighted-sum fusion unit.

3.4 Likelihood normalization

The use of normalized likelihoods together with a global threshold in addition, leads to improvements in performance as well as robustness of person authentication systems [4]. The Universal Background Model (UBM) approach [9] is the most popular normalized likelihood approach when utilizing Gaussian Mixture Model (GMM) classifier.

For VidTIMIT, the data from 24 male and 19 female clients were used to create separate gender specific universal background models. The background models were then adapted to speaker models using MAP adaptation [9]. The first two utterances for all speakers in the corpus being common were used for text dependent experiments and 6 different utterances for each speaker allowed text independent verification experiments to be conducted. For text independent experiments, four utterances from session 1 were used for training and four utterances from session 2 and 3 were used for testing.

For the UCBN database, the training data for both text dependent and text independent experiments contained 15 utterances from 5 male and 5 female speakers, and 5 utterances for testing, each recorded in a different session. The utterances were of 20-second duration for text dependent experiments, and of 40-second duration in text independent mode. Similarly to VidTIMIT, separate UBMs for the male and female cohorts were created for UCBN data.

4. REPLAY ATTACK EXPERIMENTS

The replay attack experiments were conducted in two phases, training and testing. In the training phase, a 10-mixture gaussian mixture model λ of a client's audiovisual feature vectors was built, reflecting the probability densities for the combined phonemes and visemes in the audiovisual feature space.

For testing, the clients' live test recordings were evaluated against the client's model λ by determining the log likelihoods $log\ p(X|\lambda)$ of the time sequences X of audiovisual feature vectors under the usual assumption of statistical independence of successive feature vectors.

For testing replay attacks, a number of "fake" or synthetic recordings were constructed by combining the sequence of audio feature vectors from each test utterance with ONE visual feature vector chosen from the sequence of visual feature vectors and keeping that visual feature vector constant throughout the utterance [3]. Such a synthetic sequence represents an attack on the authentication system, carried out by replaying an audio recording of a client's utterance while presenting a still photograph to the camera. Four such fake audiovisual sequences were constructed from different still frames of each client test recording. Log-likelihoods $log\ p(X'|\lambda)$ were computed for the fake sequences X' of audiovisual feature vectors against the client model λ. In order to obtain suitable thresholds to distinguish live recordings from fake recordings, detection error tradeoff (DET) curves and equal error rates (EER) were determined.

For all experiments, the threshold was set using data from the test data set. Table 2 shows the number of client trials and replay attack trials conducted for determining the EERs. The first row in Table 2 for example, refers to experiments with the VidTIMIT database in text dependent mode for a male-only cohort, comprising a total of 48 client trials (24 clients × 2 utterances per client) and 192 replay attack trials (24 clients × 2 utterances × 4 fake sequences per client). A convenient notation is used here for referring to the experiments in a particular mode (Table 1). A simple Z-norm based approach as proposed in [9] was used for the normalization of all scores.

Notation	**True description**
EER	Equal Error Rate
LF(0.25)	Late fusion with fusion weight β=0.25
FF	Feature Fusion
DB1	VidTIMIT database
DB2	UCBN database
TDMO	Text dependent male only cohort
TDFO	Text dependent female only cohort
TIMO	Text independent male only cohort
TIFO	Text independent male only cohort

Table 1 : Notation for different experiments with Speaking face data

Different sets of experiments were conducted to evaluate the performance of the system in terms of DET curves and equal error rates (EER). The results of only two types of data, that is DB1TIMO (VidTIMIT database text-independent male-only cohort) and DB2TDFO (UCBN database text-dependent female-only cohort) experiments are reported here. All late fusion experiments had varying combination weights 'β' for combining audio and visual scores. β is varied from 0→1 with β increasing for increasing visual scores.

Speaking face data	Client Trials	Replay Attack Trials
DB1TDMO	48(24×2)	192(24×2×4)
DB1TDFO	38(19 × 2)	152(19×2× 4)
DB1TIMO	144 (24 × 6)	384(24×2×2×4)
DB1TIFO	114(19 × 6)	304 (19×2×2×4)
DB2TDMO	100 (5 × 10)	200 (5×4×10)
DB2TDFO	100(5 × 10)	200 (5×4×10)
DB2TIMO	100 (5 × 10)	200 (5×4×10)
DB2TIFO	100 (5 × 10)	200 (5×4×10)

Table 2: Number of client and replay attack trials

For the first set of experiments, original data from VidTIMIT and original files from UCBN database were used. The DET curves for the baseline performance of the system with the original data, with feature fusion and late fusion are shown in Figure 3. As can be seen in Figure 3, the baseline EER achieved is 3.65% for DB1TIMO and 2.55% for DB2TDFO, as compared to 8.1% (DB1TIMO) and 6.8% (DB2TDFO), achieved for late fusion with β=0.75. In Figure 4, the behavior of the system is shown when subjected to different types of environmental degradations as is the EER sensitivity to variations in training data size. Once again, feature level fusion outperforms late fusion for acoustic and visual degradations. When mixed with acoustic noise (Factory noise at 6 dB SNR + channel effects), feature fusion allows a performance improvement of the order of 38%

compared to late fusion (β=0.25), and 18% for late fusion (β=0.75). When mixed with visual artefacts, the improvement in performance achieved with feature fusion is about 30.40% as compared to LF (β=0.75), and 18.9% with LF (β=0.25). Table 3 and Figure 4 show the baseline EERs achieved and EERs achieved with inclusion of visual artefacts, acoustic noise and shorter training data. The table also shows a drop in performance due to late fusion and feature fusion.

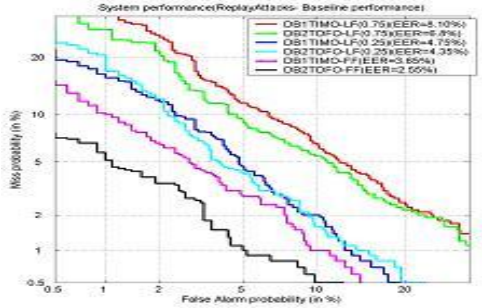

Fig 3: Baseline replay attack performance (late fusion vs. feature fusion)

Speaking Face Data	Base Line % EER	Acoustic Effects	Visual artefacts	Utterance Length Effects
DB1TIMO (LF-0.75)	8.1	9.98 (-23.22 %)	11.44 (-41.12%)	8.87 (-9.46 %)
DB2TDFO (LF-0.75)	6.8	8.18 (-20.22 %)	9.81 (-44.92%)	7.63 (-12.15%)
DB1TIMO (LF- 0.25)	4.75	6.76 (-42.33%)	6.19 (-30.22%)	6.01 (-26.57%)
DB1TDFO (LF- 0.25)	4.35	6.10 (-40.16%)	5.75 (-32.82%)	5.42 (-24.53%)
DB1TIMO (FF)	3.65	3.83 (-4.83%)	4.06 (-11.82%)	5.19 (-42.32%)
DB2TDFO (FF)	2.55	2.60 (-2.06%)	2.89 (-13.42%)	3.59 (-40.96%)

Table 3: Relative Performance with acoustic noise, visual artefacts and variation in training data size

The influence of training utterance length variation on system performance is quite remarkable and different as compared to other effects. The system is more sensitive to utterance length variation for feature fusion mode as compared to late fusion mode (Table 3).

The drop in performance is less (9.46% for late fusion (β=0.75)) and (26.57% for late fusion (β=0.25)) as compared to 42.32% drop for feature fusion for DB1TIMO, and likewise, the drop is 12.15% and 24.53% as compared to 40.96% drop for DB2TDFO. The utterance length is varied from 4 seconds to 1 second for DB1TIMO and from 20 seconds to 5 seconds for DB2TDFO data.

This drop in performance is because of a larger dimensionality of the joint audiovisual feature vectors used (8 MFCCs+10 eigenlips+6 lip dimensions), as well as the shorter utterance length, which seems to be not sufficient to establish the audiovisual synchrony.

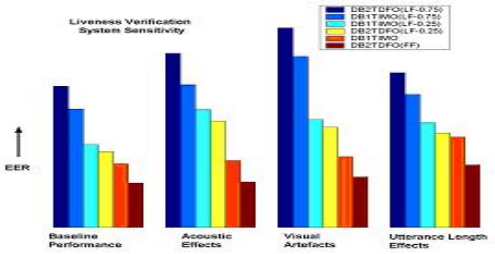

Fig 4: Liveness Verification System Sensitivity

Longer utterances and hence more training speech would allow the audiovisual correspondence to be learnt, and better liveness verification to be done.

5. CONCLUSIONS

In this paper we have shown that feature level fusion of audiovisual feature vectors substantially improves the performance of a face-voice authentication system for checking liveness and thwarting replay attacks. Also, the sensitivity of the feature fusion approach to variations in the size of the training data has been recognized.

6. REFERENCES

[1] Ross, A., and Jain, A.K., "Information fusion in biometrics", *Pattern Recognition Letters 24, 13 (Sept. 2003), 2115–2125.*

[2] J. Kittler, G. Matas, K. Jonsson, and M. S´anchez, "Combining evidence in personal identity verification systems," *Pattern Recognition Letters, vol.18, no.9,pp.845–852,Sept. 1997.*

[3] Chetty, G. and M. Wagner, "Liveness" verification in audiovideo authentication. *Proc. Int Conf on Spoken Language Processing ICSLP-04, pp 2509-2512.*

[4] Sanderson, C. and K.K. Paliwal, "Fast features for face authentication under illumination direction changes", *Pattern Recognition Letters 24, 2409-2419.*

[5] *Signal Processing Information Base (SPIB)* http://spib.rice.edu/spib/select_noise.html

[6] Chetty, G. and M. Wagner, "Automated lip feature extraction for liveness verification in audiovideo authentication", *Proc. Image and Vision Computing 2004, New Zealand, pp 17-22.*

[7] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing, Vol. 10,No. 1-3, 2000, pp.19-41.*

[8] M.C.Cheung, K.K. Yiu, M.W.Mak, and S.Y.Kung, "Multi-sample fusion with constrained feature transformation for robust speaker verification", *Proceedings Odyssey'04 Conference.*

[9] R.Auckenthaler, E.Paris, and M.Carey, "Improving GMM Speaker verification System by Phonetic Weighting", *ICASSP'99, pp. 1440-1444, 1999.*