Liveness detection using cross-modal correlations in face-voice person authentication

Girija Chetty and Michael Wagner

Human Computer Communication Laboratory
School of Information Sciences and Engineering
University of Canberra, Australia
g.chetty@student.canberra.edu.au

Abstract

In this paper we show the potential of two new features as powerful anti-spoofing measures for face-voice person authentication systems. The features based on latent semantic analysis (LSA) and canonical correlation analysis (CCA), enhance the performance of the authentication system in terms of better anti-imposture abilities and guard against video replay attacks, which is a challenging type of spoof attack. Experiments conducted on 2 speaking-face databases, VidTIMIT and UCBN, show around 42% improvement in error rate with CCA features and 61% improvement with LSA features over feature-level fusion of face-voice feature vectors.

1. Introduction

Audio-visual (AV) person authentication systems, which process information on both the face and voice of a person, are potentially less vulnerable to replay attack than single-mode systems because of the inherent differential difficulty of "spoofing" both a person's voice and synchronously the image of the person's speaking face [1]. However, in most current AV authentication systems [2,3], the face and voice information is processed independently and face-voice synchrony is neither measured nor preserved, making the systems vulnerable to replay attacks with a combination of pre-recorded audio and a still photo of the person. Checking liveness guards against this type of attack by ensuring that the biometric data is captured from an authorized, live person present at the time. Systems that check the liveness of face-voice data include *Identix* [4], which uses a liveness algorithm quantifying head movements from 3D face data, and a face-voice recognition system [5], where liveness checking is done by matching the lip movement between voice and face. Even systems that detect lip movements in the facial image are not entirely safe against replay attack, since it has recently become possible to animate still photos with lip movements, eye blinking and other facial expressions using face generation and lip synching software [6]. Such attacks cannot be thwarted by simply detecting lip movements. One defence is the detection and subsequent preservation of audio-visual synchrony by analysing face-voice data in a cross-modal space.

In this paper, we propose latent semantic analysis (LSA), based on feature extraction in the joint face-voice feature space, and canonical correlation analysis (CCA), based on optimising cross-correlations in a rotated audio-visual subspace for liveness detection. An earlier paper dealt with feature-level fusion of face and voice features extracted separately with a promising defence against pre-recorded audio combined with a still photo [7]. In this paper, we address video replay attacks with fake video clips created from still photos. The proposed CCA and LSA features presented in this paper address the liveness detection part of the framework. We present the LSA method in Section 2 and the CCA method in Section 3. The experimental data are described in Section 4 and experimental results are discussed in Section 5. Conclusions and further suggestions are found in Section 6.

2. Latent semantic analysis

Latent semantic analysis is a powerful tool used in text information retrieval to discover underlying semantic relationships between different textual units [8]. The LSA technique achieves three goals: dimension reduction, noise removal and the uncovering of the semantic and hidden relation between different objects such as keywords and documents. In our current context, we used LSA to uncover the synchronism between image and audio features in a video sequence. The method consists of four steps: construction of a joint multimodal feature space, normalisation, singular value decomposition and semantic association measurement.

First we build a matrix for the sequence of audio and video frame vectors. With n visual features (such as image principal components) and m audio features (such as mel-frequency cepstrum coefficients) at each of the t AV frames, the joint feature vectors can be expressed as:

$$X = [V_1, V_2, \dots, V_n, A_1, A_2, \dots, A_n] \quad (1)$$

$$V_i = [v_i(1), v_i(2), \dots, v_i(t)]^T \quad (2)$$

$$A_i = [a_i(1), a_i(2), \dots, a_i(t)]^T \quad (3)$$

Features are normalised with respect to their range over time as:

$$X_{it} \leftarrow X_{it} / \max_t (|X_{it}|) \quad (4)$$

All normalized matrix elements therefore have values between –1 and 1. Singular value decomposition (SVD) yields

$$X=S\ V\ D^T \quad (5)$$

where S and D are composed of the eigenvectors of XX^T and V is the diagonal matrix of eigenvalues. Normally, S, V and D must all be of full rank. The strength of SVD, however, lies in that it allows a simple strategy for an optimal approximate fit using smaller matrices.

If we order the eigenvalues in V in descending order and keep the first k elements, we can represent X by $X \approx X^{(k)}=S^{(k)}V^{(k)}D^{(k)T}$, where $V^{(k)}$ consists of the first k ele-

ments of V, $S^{(k)}$ consists of the first k elements of S, and $D^{(k)}$ consists of the first k elements of D. It can be shown that $X^{(k)}$ is an optimal representation of X in a least-squares sense [8].

By keeping only the first and most important eigenvalues, we derive an approximation of X with reduced feature dimensions, where the correlations (semantic information) between audio and visual features is mostly preserved and irrelevant noise is greatly reduced.

3. Canonical correlation analysis

Canonical correlation analysis is a multivariate statistical technique, which attempts to find a linear mapping that maximizes the cross-correlation between two features sets [9]. It finds the transformation that can best represent (or identify) the coupled patterns between features of two different subsets.

A set of linear basis functions, having a direct relation to maximum mutual information, is obtained by CCA in each signal space, such that the correlation matrix between the signals described in the new basis is diagonal [14]. The basis vectors can be ordered such that the first pair of vectors w_{x1} and w_{y1} maximizes the correlation between the projections ($x^T w_{x1}$, $y^T w_{y1}$) of the signals x and y onto the two vectors respectively. A subset of vectors containing the first k pairs defines a linear rank-k relation between the sets that is optimal in a correlation sense. In other words, it gives the linear combination of one set of variables that is the best predictor and at the same time the linear combination of another set which is most predictable. It has been shown that finding the canonical correlations is equivalent to maximizing the mutual information between the sets if the underlying distributions are elliptically symmetric [9].

Given two random variables x and y, from a multi-normal distribution [14]:

$$\begin{pmatrix} x \\ y \end{pmatrix} \cong N\left(\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} \right), \tag{6}$$

where $C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}$ is the covariance matrix. C_{xx} and C_{yy} are non-singular matrices and $C_{xy} = C_{yx}^T$. Consider the linear combinations, $x = w_x^T(x - x_0)$ and $y = w_y^T(y - y_0)$ of the two variables respectively. The correlation between x and y is

$$\rho = \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}} \tag{7}$$

and the description of the canonical correlations is

$$\begin{bmatrix} C_{xx} & [0] \\ [0] & C_{yy} \end{bmatrix}^{-1} \begin{bmatrix} [0] & C_{xy} \\ C_{yx} & [0] \end{bmatrix} \begin{pmatrix} \hat{w}_x \\ \hat{w}_y \end{pmatrix} = \rho \begin{pmatrix} \lambda_x \hat{w}_x \\ \lambda_y \hat{w}_y \end{pmatrix} \tag{8}$$

where ρ, λ_x, $\lambda_y > 0$ and $\lambda_x \lambda_y = 1$. Eqn (8) can be rewritten as

$$\begin{cases} C_{xx}^{-1} C_{xy} w_y = \rho \lambda_x w_x \\ C_{yy}^{-1} C_{yx} w_x = \rho \lambda_y w_y \end{cases} \tag{9}$$

Solving (9) gives K solutions $\{\rho_k, w_{xk}, w_{yk}\}$, k=1…K, where K is the minimum of the input dimensionality and the output dimensionality. The linear combinations, $x_k = w_{xk}^T x$ and $y_k = w_{yk}^T y$ are the canonical variates and ρ_k are the correlations. An important aspect in this context is that the canonical correlations are invariant to affine transformations of x and y. Also note that the canonical variates corresponding to the different roots of (9) are uncorrelated, implying that:

$$\begin{cases} w_{xk}^T C_{xx} w_{xm} = 0 \\ w_{yk}^T C_{yy} w_{ym} = 0 \end{cases} \quad if \ \ n \neq m \tag{10}$$

4. Speaking-face data and liveness detection

The speaking face data from two different databases, VidTimit and UCBN were used for liveness detection. The VidTIMIT multimodal person authentication database [10], 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 with corresponding audio provided as a 16-bit 32-kHz mono PCM file.

Figure 1: Faces from (a) VidTimit, (b) UCBN

The UCBN database is an Australian free-to-air broadcast news database being developed by authors. We have developed an omni-face detection scheme for detecting faces against often complex backgrounds [11]. The database consists of 20-40 second video clips of 5 female and 5 male anchor persons and newsreaders. The training data comprises 15 clips and the test data comprises 5 additional clips of the same speakers. Each video sample is a 25 fps MPEG2-encoded stream with a resolution of 720×576 pixels and corresponding 16-bit, 48-kHz PCM audio. Figure 1a and Figure 1b show sample speaking-face data from VidTIMIT and UCBN.

The liveness detection subsystem of the proposed face-voice person authentication framework contains an audiovisual cross-modal feature extractor, which computes LSA and CCA feature vectors for low-level visual and audio features. The visual features are 20 PCA (eigenface) coefficients, and the audio features are 12 MFCC coefficients. The cross modal feature extractor computes the LSA and CCA features from the PCA and MFCC vectors and, based on our experimental results, fewer than 10 LSA and CCA features are normally

sufficient to achieve good performance. This is a significant reduction of feature dimension compared with the 32-dimensional audio-visual feature vector formed by concatenating 20 PCA and 12 MFCC vectors in the feature-level fusion methods reported earlier [7,11].

5. Spoof attack experiments

To evaluate the potential of LSA and CCA features in spoof protection, different sets of experiments to investigate replay attacks were conducted. Eight LSA and CCA features were extracted from 20 eigenface vectors and 12 MFCC vectors using 30 millisecond Hamming windows.

In the training phase, a 10-Gaussian mixture model of each client's LSA and CCA feature vectors in the cross-modal space was built by constructing a gender-specific universal background model (UBM) and then adapting each UBM by MAP adaptation [12]. Both text-dependent and text-independent experiments were conducted with VidTIMIT and UCBN corpus data. For the VidTimit database in text dependent mode, there were 48 client trials (24 clients × 2 utterances per client) and 1104 impostor trials (24×23×2) for male subjects and 38 client trials (19×2) and 684 impostor trials (19×18×2) for female subjects. In text-independent mode there were 144 client trials (24×6) and 3312 impostor trials (24×23×6) for male subjects, and 114 client trials (19×6) and 2052 impostor trials (19×18×2) for female subjects. For the UCBN database, there were 50 client trials (5×10) and 200 impostor trials (5×4×10) for male as well as for female subjects, in both text dependent and text-independent mode. In the test phase, clients' live test recordings were evaluated against a client's model λ by determining the log likelihoods $log\ p(X|\lambda)$ of the time sequences X of audiovisual feature vectors in cross-modal space. A Z-norm based approach [13] was used for score normalization.

Data	EF	LSA	CCA
DB1TDMO	2.03	0.25	0.40
DB1TDFO	2.56	0.53	0.90
DB1TIMO	2.93	0.38	0.75
DB1TIFO	3.02	0.79	1.15
DB2TDMO	2.54	0.26	0.81
DB2TDFO	2.20	0.35	0.08
DB2TIMO	2.71	0.41	0.73
DB2TIFO	2.82	0.52	0.87

Table 1: Equal-error rates for Type-1 replay attacks

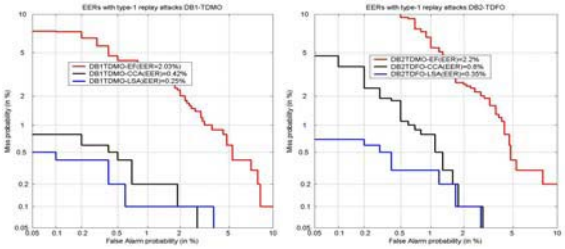

Figure 2 : DET curves for Type-1 TD tests, (a) male subjects in VidTimit, (b) female subjects in UCBN

For testing replay/spoof attacks, two types of replay-attack experiments were conducted. For *Type-1* replay attacks, a number of "fake" 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. Such a fake sequence represents an attack on the authentication system, which is carried out by replaying an audio recording of the 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.

For *Type-2* replay attacks, a video clip was constructed from a still photo of each speaker. This represents a scenario of a spoof attack with an impostor presenting a fake video clip constructed from pre-recorded audio and a still photo of the client animated with facial movements and voice-synchronous lip movements.

Data	EF	LSA	CCA
DB1TDMO	3.02	1.01	1.92
DB1TDFO	3.46	1.78	2.76
DB1TIMO	4.5	1.35	2.11
DB1TIFO	4.64	2.47	3.65
DB2TDMO	3.14	1.45	2.43
DB2TDFO	3.47	1.68	2.75
DB2TIMO	3.01	2.22	2.31
DB2TIFO	3.62	2.30	2.81

Table 2: Equal-error rates for Type-2 replay attacks male subjects, (b) female subjects

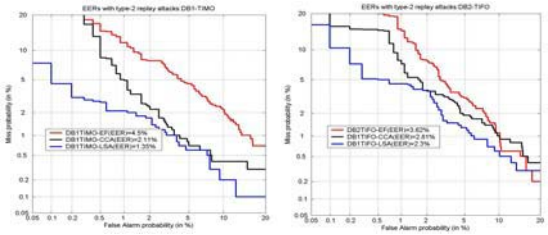

Figure 3 : DET curves for Type-2 TI tests. (a) male subjects in VidTimit, (b) female subjects in UCBN

The final set of experiments conducted were to test the training data size required for the LSA/CCA features to learn the audio-visual synchrony, by varying the length of training. All experiments included examining the performance of the feature-level fusion technique with concatenated audio-visual feature vectors (20 PCA+12 MFCC features) for baseline comparison.

The EER results in Table 1 and Figure 2 show the potential of the proposed LSA and CCA features over early fusion (EF) for *Type-1* replay attacks. DB1 is VidTimit, DB2 is UCBN, TD and TI are the text-dependent and text-independent tests, MO and FO are the male-only and female-only cohorts. An improvement of 80% with 8-dimensional LSA features and 60% with 8-dimensional CCA features is achieved over 32-dimensional feature-level fusion.

Table 2 and Figure 3 show the improvement in error rates achieved for *Type-2* replay attacks. Approximately 43% improvement in EERs with 8-dimensional LSA features and 22% with 8-dimensional CCA features is achieved. This is a

remarkable improvement in EERs, due to ability of LSA and CCA features to detect mismatch in synchrony in video replay attacks.

Data	EF	LSA	CCA
DB1- TDMO	9.55	4.42	6.34
DB1- TDFO	12.98	4.85	6.74
DB1- TIMO	13.05	5.04	7.02
DB1- TIFO	13.65	5.62	7.54
DB2-TDMO	14.45	3.88	5.08
DB2- TDFO	14.66	3.95	5.27
DB2- TIMO	13.35	4.46	6.35
DB2- TIFO	13.68	4.77	6.62

Table 3: Equal-error rates for 2-second training data

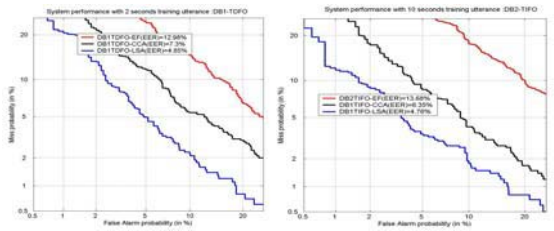

Figure 4: DET curves for female subjects. (a) TD with 2s training in VidTimit, (b) TI with 10s training in UCBN

The influence of training data size on the performance of LSA and CCA features is examined in Table 3 and Figure 4. Here the training utterance length was kept as 2s for VidTIMIT and 10s for UCBM. In general there is a drop in performance for shorter enrolment. However, an improvement of 55% with LSA features and 38 % with CCA features is achieved compared to early fusion, when utterance length is reduced to 2s and 10s from the normal utterance length of 4s and 20s for the two databases. This is due to the reduction in dimension of the LSA and CCA features as well as to their ability to detect synchrony mismatch in cross-modal space.

6. Conclusion

The potential of two new cross-modal features for reducing video-replay spoof attacks has been shown in this study. About 42% overall improvement in error rate with canonical correlation analysis features and 61% improvement with latent semantic analysis features is achieved as compared to feature-level fusion of image-PCA and MFCC face-voice feature vectors.

7. References

[1] J.B. Millar, F. Chen, I. Macleod, S. Ran, H. Tang, M. Wagner, X. Zhu, Overview of speaker verification studies towards technology for robust user-conscious secure transactions, Proc. 5th Austr. Int. Conf. on Speech Sc. and Techn., SST-94, Perth, pp. 744-749, 1994.

[2] N. Poh and J. Korczak, "*Hybrid biometric person authentication using face and voice features*," in Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification, Halmstad, Sweden, June 2001, pp. 348--353.

[3] Conrad Sanderson and Kuldip K. Paliwal, "Identity verification using speech and face information", Digital Signal Processing, Volume 14, Issue 5, Pages 397-507 (September 2004)

[4] http://www.identix.com

[5] C.C.Brown, X.Zhang, R.M.Mersereau and M.Clements, " *Automatic speechreading with application to speaker verification",* Proc. ICASSP, Orlando, May 2002.

[6] M.M Cohen and D.Massaro, *"Synthesis of visible speech,"* Behaviour research methods, instruments and computers, Vol.22, no.2, pp.260-263, April 1990.

[7] Chetty, G. and Wagner, M., "*Liveness' Verification in Audio-Video Authentication*", Proc. Int Conf on Spoken Language Processing ICSLP-04, Paper Spec3603p6.

[8] Deerwester, S., Dumais, S.T., Frunas, G.W., Landauer, T.K., and Harshman, R. Indexing by Latent Semantic Analysis, Journal American Society for Information Sci., 41(6),391-407.

[9] J.Kay, *"Feature discovery under contextual supervision using mutual information"*., International Joint Conference on neural networks, volume 4, pp. 79-84, IEEE,1992.

[10] Sanderson, C. and K.K. Paliwal (2003), "*Fast features for face authentication under illumination direction changes"*, Pattern Recognition Letters 24, 2409-2419.

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

[12] 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.

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

[14] M.Borga, H.Knutsson, "An adaptive stereo algorithm based on canonical correlation analysis", Proceedings of the Second IEEE International Conference on Intelligent Processing Systems, pp.177-182, August, 1998.