VIDEO BASED PERSON AUTHENTICATION VIA AUDIO/VISUAL ASSOCIATION

Ming Liu, Thomas S. Huang

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
405 N. Matthew Ave., Urbana, IL, 61801
{mingliu1, huang}@ifp.uiuc.edu

ABSTRACT

Multi-modal person authentication systems can achieve higher performance and robustness by combining different modalities. The current fusion strategies of different modalities are mainly based on the output of individual modalities. However, there are detail structures between facial movement and speech signal. In this paper, Audio/Visual association, a lower level fusion, is proposed to fuse the information between lip movement and speech signal. The experimental results indicate that this type of fusion strategy improve the performance of multi-modal person authentication system.

1. INTRODUCTION

In recently years, more and more attention are drawn to Multimodal person authentication system for better security and more robust biometric system. It is clearly shown that the combination of different modalities do improve the performance and robustness of identity authentication systems. [4] adopts Bayesian supervisor to fuse face and speech cues. [1] compare SVM, minimum cost Bayesian classifier,Fisher Discriminant, C4.5 decision tree and MLP for speech/face fusion. Among these methods, SVM and Bayesian classifiers have the best performance.

Instead of two modalities fusion, [7] investigate the multimodal biometric system by fusing face, fingerprint and speech. On a small database, the fused system achieve very good performance. Although, there are fairly large amount of research has been conducted in multi-modal person identification/authentication systems. The fusion strategies of different modalities are mainly based on the combination of decision outputs of individual modalities. However, in video sequence, the lip movement and speech signal are highly correlated. There are detail structures between lip movement and speech signal. In this sense, the fusion of these two modalities should start at lower level to incorporate the correlation between lip movement and speech signal. [9][8] are try to integrate the lip movement and speech for speaker verification task. The experimental results show significant performance boost on low SNR conditions by fusing lip movement. The facial features, excluding lip region, are less correlated with speech signal. However, these features are more efficient in describing person characteristic than lip region. In order to achieve high performance for person authentication, these features are very crucial cues. Moreover, the person dependent lip shape, texture and movement will impair the generic modelling between lip movement and speech signal. They are good for person authentication. In this paper, an video based person authentication system is proposed. By learning audio visual association of video sequence, the better description of video data is achieved which will lead to more accurate biometric system.

The paper is organized as following. Section 2 describes the concept of video based person authentication. The speaker authentication subsystem is described in section 3. Section 4 give a face authentication subsystem. Audio/Visual association is illustrated in section 5. Experiments and results are shown in section 6. Conclusions are in section 7.

2. VIDEO BASED PERSON AUTHENTICATION

The emphasis of video based person authentication is focusing on the detail structure between speech signal and lip movement inside face region. By modelling this type of correlation between these two modalities, a more robust and secure biometric system can be obtained. Further more, the lip movement will change the face appearance. By modelling the correlation of speech and lip movement, the system may predict the appearance change in face region, thus lead to more robust face authentication.

The difference between video based person authentication and audio/visual speech recognition is the speaker dependence of the model between lip movement and speech signal. In audio/visual speech recognition, the model is trained to capture the the correlation between lip movement and speech signal for all speaker. Thus it ignored the speaker dependent lip shape, appearance and movement. Any type of speaker specific lip shape, appearance and movement are disturbance for a/v speech recognition. However, for person authentication system, the speaker specific lip shape, appearance and

movement are good features to differentiate the true speaker from imposter speakers. Thus, the constraint on speaker dependent lip modelling is reasonable and essential step for video based person authentication.

3. SPEAKER AUTHENTICATION

There are two type of tasks in Speaker Authentication. Among them, text-dependent(TD) authentication achieve better performance and require shorter training/evaluating utterances. The price for those advantages is more constraints. The TD authentication requires that the same phonetic sequences have to present in enrollment and test utterances. The most adopted methods for text-dependent speaker authentication are Hidden Markov Model(HMM) based methods [5][3][10]. In spite of many advantages, the limited enrollment data will degrade the estimation of model parameters via EM algorithm. The solution to this problem could be adaptation technique which generate the speaker model via adaptation from background model.

3.1. Background Modelling

In general MAP adaptation of speech recognition system, the background model is a set of HMMs trained via EM. It is also considered as a speaker-independent model or world model. Although HMM-based background model is widely used in speech recognition literature, GMM-based background model is dominant in text-independent speaker verification literature.

GMM-based adaptation is different from HMM-based adaptation. The initial HMM for digits is not defined in GMM-based adaptation. In this case, the background model is used to specify the structure and the initial parameters of the HMM. This initial of HMM also encode the speaker characteristics. An Index Transformation is performed to initialize the HMM. Given the UBM model and training utterance, the index transformation is to find the best Gaussian index sequence for the training utterance. The procedure is illustrated as following equations.

$$[i_1, i_2, ..., i_T] = F([x_1, x_2, ..., x_T]|\lambda) \quad (1)$$

$$i_t = \arg \max_{1 \le m \le M} w_m N(x_t; \mu_m, \Sigma_m) \quad (2)$$

Where $F(\cdot)$ is the index transformation, x_t is the acoustic feature vector at frame t, w_m and $N(:, \mu_m, \Sigma_m)$ are the parameters of a Gaussian in the UBM model. i_t is the best index at frame t. After index transformation, the training utterance X_1^T is converted into a integer sequence I_1^T.

Considering each Gaussian in the UBM model as a state, the UBM model can be treated as a HMM except the transition probability and the initial probability is not defined. The Gaussian index sequence is used to count the initial probability $\Pi(i)$ and transition probability $P(j|i)$. These probabilities define the topology of the HMM.

After defining the initial HMM structure and parameters, the MAP adaptation[6] is performed as following.

$$\hat{\mu}_i = \frac{\tau}{\tau + \gamma_i}\mu_i + \frac{\gamma_i}{\tau + \gamma_i}\bar{\mu}_i \quad (3)$$

where γ_i is the occupation soft-count at state i. μ_i is the mean of background model at state i. $\bar{\mu}_i$ is the mean of the training data.

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} \gamma_i(t) x_t}{\sum_{t=1}^{T} \gamma_i(t)} \quad (4)$$

Thus, if the occupation soft-count is small, the MAP estimation is close to the value of background model. In the following experiments, the factor $\tau = 0$.

4. FACE AUTHENTICATION

The Face authentication method in this paper is based on an extension of GMM-based face verification[2]. The face is represented by a set of feature vectors extracted from a rectangular grid on the face. In [2], a global GMM is built for background modelling. However, in order to model the appearance in spatial dependent way, we build local Gaussian Mixture Model for each grid node. The position information is encoded in the background modelling step. The speaker face model is built by adaptation from this background model.

Figure 4 illustrates the diagram of the face authentication system. After partition of the face region into small patches, 2D discrete transform is performed to reduce the dimensionality of appearance vector. In this paper, we set face region to be 60x60 square, the patch window is 12x12, and the scan step is 5. After 2D DCT, only the most significant coefficients are kept (28 dimension). For each grid position, we build a GMM via EM algorithm.

A weight adjusted MAP adaptation is used to generate the face model for a speaker. Assume the observed training images are I_1^N, the corresponding features are V_1^N where V_i is the assembly of all patches $(v_i^j), j = 1, ..., S$. The local GMM at each grid position j is represented as $\lambda_j = (w_j^m, \mu_j^m, \Sigma_j^m), m = 1, ..., M$, where w_j^m is the weight of m-th component of grid j, μ_j^m is the mean vector of m-th component, Σ_j^m is the covariance matrix of m-th component. The adaptation of weight is given as following.

$$\hat{w}_s^m = \frac{1}{N}\sum_{i=1}^{N} \frac{w_s^m N(v_i^s : \mu_s^m, \Sigma_s^m)}{\sum_k w_s^k N(v_i^s : \mu_s^k, \Sigma_s^k)} \quad (5)$$

$$(6)$$

The speaker face model then is generated as $\lambda_j = (\hat{w}_j^m, \mu_j^m, \Sigma_j^m), m = 1, ..., M$ for each grid position.

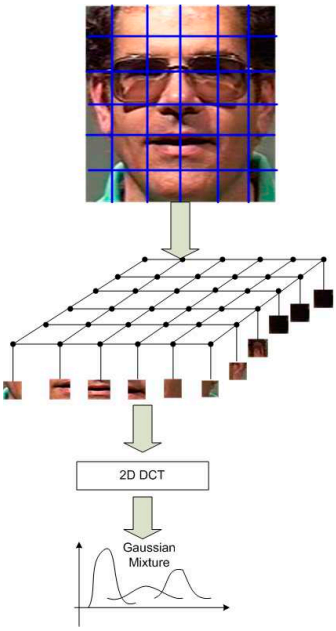

Fig. 1. Diagram of face authentication module

The log-likelihood ratio between speaker model and background model is used to give final decision based on fixed threshold.

$$LLR(V) = \frac{\prod P(v_s|\lambda_s^{spk})}{\prod P(v_s|\lambda_s^{bak})} \quad (7)$$

5. AUDIO/VISUAL ASSOCIATION

In order to model the fine structure between lip movement and speech signal, audio/visual association module is introduced in our framework. The basic function of audio/visual association is to compute the likelihood of co-appearance of video signal and speech signal. More specific, this association is speaker dependent which further encode the speaker characteristic in the system.

A Dynamic Bayesian network is shown in Figure 5, which is used to illustrate the dependence between the speech and face modalities. In the network, the shadow nodes are the hidden random variables, A corresponds to audio state, L corresponds to lip movement state, F corresponds to person identity. The white nodes are the observations: speech and faces. The L is set to represent the whole face region. Since the grid representation can be treated as a Markov Random Field, the L can be viewed as the assembly of all nodes in this Markov Random Field. According to Bayes formula and the conditional independence between O_A and O_V, we have, the likelihood of given observation sequence to one speaker S_F

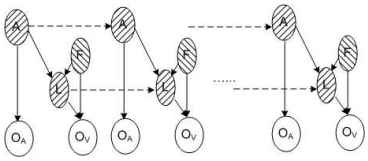

Fig. 2. Dynamic Bayesian Network for Audio/Visual Association

can be computed as following

$$P(O_A, O_V|S_F) = \sum_{\forall S_A, S_L} P(O_A, O_V, S_A, S_L|S_F)$$
$$= \sum_{\forall S_A, S_L} P(O_A|S_A, S_L)P(O_V|S_L, S_F)P(S_L, S_A|S_F)$$

Where, $P(S_L, S_A|S_F)$ is the audio/visual association likelihood for speaker S_F which can be decomposed as follows.

$$P(S_L, S_A|S_F) = P(S_L^0|S_A^0, S_F)P(S_A^0|S_F) \prod_{t=1}^{T} P(S_L^t|S_L^{t-1}, S_A^t, S_F)P(S_A^t|S_A^{t-1})$$

To simplify the audio/visual association model, we assume that

$$P(S_L^t|S_L^{t-1}, S_A^t, S_F) \propto P(S_L^t|S_L^{t-1}, S_F)P(S_L^t|S_A^t, S_F)$$

Where the conditional probability $P(S_L^t|S_A^t, S_F)$ describes the audio/visual association between speech signal and lip movement at each time instant.

6. EXPERIMENTS AND RESULTS

As mentioned in previous section, the shadow node L in Figure 5 can be viewed as the assembly of all nodes in a Markov Random Field which represent the whole face region. We can analysis the correlation between audio state A with all nodes in this Markov Random Fields, the idea output would be high correlation output for the nodes which represent lip regions. Figure 6 shows the correlation between the grids(10x10) with speech signal and the region corresponds to the maximum correlation gird position. This results indicates that the underline structure between lip movement and speech signal is properly captured by audio/visual association. Moreover, by thresholding these correlation output, it may be possible differentiate well articulate speaker and poor articulate speaker since poor articulate speaker will have less high light in the correlation map. This indicator will be very helpful to determine the weight for video stream.

In order to evaluate the performance of the multi-modal person authentication system, we conduct a sets of experiments on the subset of our Internal Audio/Visual speech database

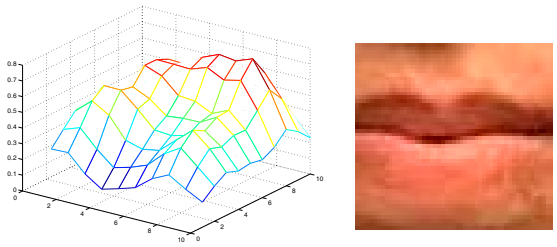

Fig. 3. Audio/Visual Association of grid positions and speech signal and the cropped grid patch of maximum associated position

Table 1. Comparison between four different person authentication system of different modalities and fusion strategies, the performance measure is Equal Error Rate(EER) in percentage

SNR	Audio	Video	AV(1)	AV(2)
0dB	26.59%	6.92%	8.04%	4.88%
5dB	18.45%	6.92%	7.49%	4.47%
10dB	14.92%	6.92%	5.31%	2.58%
15dB	10.72%	6.92%	4.35%	2.31%
25dB	5.02%	6.92%	2.59%	1.62%
clean	2.58%	6.92%	2.58%	1.51%

which contain 36 speakers(half male speakers). The continuous digits are recorded in a studio environment. For each digit, only 2 utterances are used to train speaker model, the rest 8 utterances of same digits are used as trials. 12 dimensional MFCC and ΔMFCC are used as features and Gaussian white noise is added into speech according to specified SNR.

Table 1 shows the results on person authentication. The audio-only system perform poorly under noisy condition. AV(1) is the audio/visual fusion based on decision and AV(2) is the audio/visual association fusion. It is clear that the level fusion outperform decision level fusion cross all conditions.

To verify the effectiveness of audio/visual association, we synthesis a set of "fake" data which combine video frames of one digit with speech signal of another digits of the same speaker. In this case, independently fusion of two modalities will not be able to differentiate the "fake" data from the true speaker data. However, since the association of video and audio stream are damaged by the generation, the fake data will have low association likelihood. Thus, fusing via audio visual association will able to detect the "fake" data out of true data. Table 2 shows the experimental results of four systems. The EER of audio-only, visual-only and AV fusion from decision are around 50% which make sense because each modality are really from true speaker. Since AV fusion from audio/visual association can detect the detail correlation between these two modalities, the EER of this system is 17.65%.

Table 2. Comparison between four different person authentication system on "fake" data

SNR	Audio	Video	AV(1)	AV(2)
clean	55.88%	50.0%	52.94%	17.65%

7. CONCLUSION

This paper propose a video-based person authentication framework for biometric system. With audio/visual association, the detail relationship between lip movement and speech signal is captured and used for better description of observed audio/video data. In this sense, the fusion based on audio/visual association is able to outperform traditional decision-level fusion, more over, it will able to detect the "fake" data by measuring the audio/visual association pattern in the "fake" data.

8. REFERENCES

[1] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Fusion of face and speech data for person identity verification. 10:1065–1074, 1999.

[2] F. Cardinaux, C. Sanderson, and S. Bengio. Face verification using adapted generative models. In *Proceeding of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 825–830, 2004.

[3] G. Doddington. Speaker recognition-identifying people by their voices. In *Proceedings of IEEE*, volume 73, pages 1651–1644, 1986.

[4] B. Duc, E. Bigun, J. Bigun, G. Maitre, and S. Fischer. Fusion of audio and video information for multi modal person authentication. 18(9):835–843, September 1997.

[5] S. Furui. Cepstral analysis technique for automatic speaker verification.

[6] J. Gauvain and C. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. 2:291–298, 1994.

[7] A. Jain, L. Hong, and Y. Kulkarni. A multimodal biometric system using fingerprint, face and speech. pages xx–yy, 1999.

[8] S. Lucy, T. Chen, S. Sridharan, and V. Chandran. Integration strategies for audio-visual speech processing: Applied to text-dependent speaker recognition. pages 495–506, 2005.

[9] T. Wark, S. Sridharan, and V. Chandran. Robust speaker verification via asynchronous fusion of speech and lip information. pages xx–yy, 1999.

[10] K. Yu, J. Mason, and J. Oglesby. Speaker recognition using hidden markov models, dynamic time warping and vector quantisation. 142:313–318, 1995.