AN AUDIO-VISUAL FUSION FRAMEWORK WITH JOINT DIMENSIONALITY REDUCTION

Ming Liu, Yun Fu, and Thomas S. Huang*

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{mingliu1, yunfu2, huang}@ifp.uiuc.edu

ABSTRACT

By combining audio and visual modalities, the speech recognition systems achieve higher performance and robustness. The fusion strategies to this point are mainly three types: feature level fusion, model level fusion, and decision level fusion. In this paper, we present a novel audio-visual fusion framework, in which a joint dimensionality reduction approach is used to project the audio and visual features into more compact subspaces. With correlation preserving criteria, the representations of projected audio and visual features will be able to preserve the correlation conveyed in the original audio and visual feature space. At the same time, the better model efficiency is achieved in the more compact feature spaces. The experiments on audio-visual person verification demonstrate the efficiency and effectiveness of the proposed fusion framework.

***Index Terms*—** Audio-visual fusion, dimensionality reduction, canonical correlation analysis, audio-visual person verification.

1. INTRODUCTION

Fusion of multimodal information is an important topic for modern pattern recognition systems. Due to the the increasing availability of multimodal data, more and more pattern recognition systems are fusing different modalities to achieve better performance/robustness, such as audio-visual speech recognition, audio-visual speaker verification and audio-visual person tracking, etc. As a special type of fusion, audio-visual fusion is particular interesting because of these two modalities convey the most important information for human computer communication. Researchers from computer vision, multimedia and speech processing have given intensive efforts on this problem. Audio-visual speech recognition has been shown superior performance over the conventional speech recognizer. Visual speech recognition is originally proposed to help acoustic speech recognizer in the scenario of Automatic Speech Recognition (ASR) by Petajan [1]. Although significant progress has been made on ASR in recent decades, the performance of the state-of-the-art system is still beyond the practical requirement. Most state-of-the-art ASR systems only use the acoustic signal for recognition, which makes it susceptible to acoustic noise [2]. Visual speech, on the other hand, is not affected by acoustic noise and it provides partial information about the place of articulation (visibility of tongue, teeth and lips). Since Petajan [1], the visual speech information has been successfully adopted by ASR system to achieve better accuracy and robustness [3, 4, 5, 6].

In spite of the success of visual speech recognition, there are two major open problems. One is to normalize the visual feature for different speakers. In current appearance based method, there is no mechanism to compensate the appearance difference between speakers which results in poor generalization ability to recognize the visual speech for an unseen person. The other problem is the feature extraction of visual and acoustic data are done separately, which enlarges the possibility that detailed correlation structure may not be picked up with current independent feature extraction procedure. Among all the proposed fusion schemes in the literature, there are mainly three types. The first is early fusion, so called feature level fusion, which simply concatenates the different modality features together. This type of fusion often suffers inferior performance compared to the other fusion methods. The late fusion is also called model-based fusion which combines two modalities by fusion two single-modality statistical models to form a hybrid (multimodal) statistical model. This multi-modality statistical model has two different types of observations which are fused in the model level. The third fusion strategy is decision level fusion which basically integrates the output of two single-modal statistical models.

The main contribution of this paper is the fusion scheme based on joint dimensionality reduction. The detailed A-V correlation can be efficiently captured in a more compact feature space. Compared with conventional A-V fusion methods, the proposed method has more capability to fine tune the detailed A-V correlation. Experimental results show that this method successfully captures the correlation between A-V modalities. Also, it achieves the best fusion benefit over than conventional fusion methods.

2. AUDIO-VISUAL FUSION METHODS

There are lots of fusion methods proposed in the literature. Generally, they belong to three types: early fusion so called feature level fusion, model level fusion and decision level fusion. The last two fusion methods are also called late fusion methods. These three fusion types are summarized as follows.

Feature Level Fusion. Feature level fusion is one of the simplest methods to integrate audio and visual information [3, 4, 7]. Basically, it concatenates the audio feature A and visual feature V at the same time to a larger feature vector $O = (A, V)$. Then based on this new observation, the probabilistic model is learned in this new feature space to compute the likelihood $P(O|\lambda) = P(A, V|\lambda)$ of the observation, where λ is the model parameter. Despite of the simplicity of feature level fusion, this method often suffers inferior fusion performance compared to state level and decision level fusion. The main reason is that the information conveyed in audio stream and visual stream are often not equal and even time varied, so the equally weighted concatenation is a suboptimal solution.

State Level Fusion. Instead of concatenation in feature level fusion, state level fusion is combining the state likelihood of prob-

*This research was funded in part by the U.S. Government VACE program, and in part by the NSF Grant CCF 04-26627. The views and conclusions are those of the authors, not of the US Government or its Agencies.

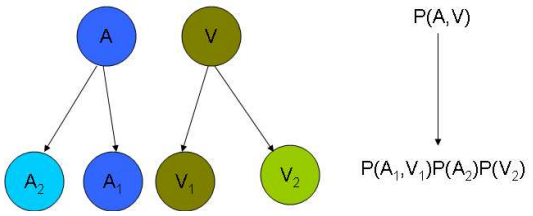

Fig. 1. Joint dimensionality reduction for audio-visual fusion.

abilistic models of audio stream and visual stream[3, 8, 9, 6, 10]. The likelihood of the observation on an A-V state $S = (S_A, S_V)$ is given by $P(A, V|S, \lambda) = P(A|S_A, \lambda)^{\beta} P(V|S_V, \lambda)^{1-\beta}$, which is the weighted combination of the likelihood of individual modality. In this procedure, the A-V features are normalized by each state parameters. In this sense, the state fusion is more preferable than feature level fusion. Also, the asynchrony of A-V streams can be modeled by specific topology of the probabilistic model, such as coupled hidden Markov model [8]. The weighting of each modalities can be easily adopted by weighted summation of the log likelihood score of audio streams and visual streams. So far, state level fusion is the state-of-the-art for A-V speech recognition.

Decision Level Fusion. Unlike the state level fusion, decision level fusion is to combine the final likelihood of audio and visual probabilistic models, e.g. $P(A, V|\lambda) = P(A|\lambda)^{\alpha} P(V|\lambda)^{1-\alpha}$. Its scheme is much easier than state level fusion methods. However, it is often slightly worse than state level fusion.

3. AUDIO-VISUAL FUSION BASED ON JOINT DIMENSIONALITY REDUCTION

The three methods mentioned above treat audio and visual feature as a whole vector. And the feature extraction of audio and visual modality are done independently. However, the feature extracted from each modality might be equally correlated to each other. For example, the lip region within the face region is more related to speech signal than other facial parts. So we can reduce the original audio and visual feature to a more compact feature space while still preserving the correlation between A-V feature in the original feature space. This procedure is essentially a joint dimensionality reduction which tries to preserve the correlation after projection. After dimensionality reduction, we can only model the joint distribution in this compact A-V feature space which leads to better model efficiency and less parameters. Figure 1 illustrates the basic idea of the proposed fusion framework. The audio feature A and visual feature V are simultaneously projected into more compact feature subspaces. And the joint distribution $P(A, V|\lambda)$ in the original spaces can be approximated by the factorized distribution in the projected spaces $P(A, V|\lambda) \sim P(A_1, V_1|\lambda)P(A_2|\lambda)P(V_2|\lambda)$, where A_1 and V_1 are the most correlated A-V features. A_2 and V_2 are the least correlated features.

3.1. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) turns out to be one of this kind of joint dimensionality reduction technique. CCA was originally developed by H. Hotelling [11]. It tries to find two sets of bases simultaneously for two multidimensional random variables X and Y. After projection on these two set of bases, the correlations between the projected variables are mutually maximized. Therefore, this dimensionality reduction is to preserve the correlation conveyed in the original features. CCA has enjoyed popularity in statistics, economics, medical studies and meteorology. [12, 13, 14].

Without losing generality, let us consider only one pair of bases w_x in X and w_y in Y space respectively. The bases pair associated with the largest canonical correlation can be solved by the following optimization problem.

$$\begin{pmatrix} w_x^* \\ w_y^* \end{pmatrix} = \arg\max_{w_x, w_y} \frac{E[w_x^T X Y^T w_y]}{\sqrt{E[w_x^T X X^T w_x] E[w_y^T Y Y^T w_y]}} \quad (1)$$

$$= \arg\max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}}, \quad (2)$$

where,

- the maximum of the correlation $\frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}}$ with respect to w_x and w_y is the maximum canonical correlation projections.
- $C_{xx} = E[XX^T]$ and $C_{yy} = E[YY^T]$ are within-sets covariance matrices.
- $C_{xy} = E[XY^T] = C'_{yx}$ are the cross-sets covariance matrices.

It is clear that scale up w_x and w_y will not change the correlation in Eq.1. The original optimization problem can be transformed into a constrained optimization problem.

$$\begin{pmatrix} w_x^* \\ w_y^* \end{pmatrix} = \arg \max_{\substack{w_x^T C_{xx} w_x = 1 \\ w_y^T C_{yy} w_y = 1}} w_x^T C_{xy} w_y. \quad (3)$$

By the lagrange multiplier method, the final solution is based on eigen-value equations

$$\begin{cases} C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX} \hat{w}_X = \rho^2 \hat{w}_X \\ C_{YY}^{-1} C_{YX} C_{XX}^{-1} C_{XY} \hat{w}_Y = \rho^2 \hat{w}_Y \end{cases}, \quad (4)$$

where ρ^2 is the squared canonical correlation and $\hat{w}_X$ and $\hat{w}_Y$ are the normalized canonical correlation basis vectors.

Canonical correlation analysis is closely related to mutual information maximization procedure. In fact, the correlation is just one special type of mutual information between two random variables. If there is no higher order statistics between two random variables, which is true if the two random variables are Gaussian distributed, the canonical correlation analysis will be equal to maximizing the the mutual information between the projected x and y. Also, there is close relationship between CCA and LDA. If one random variable of CCA is actually a discrete random variable, it can be shown that CCA is equal to LDA. Furthermore, [15] shows that Principal Component Analysis (PCA), Partial Least Squares (PLS) and Multivariate Linear Regression (MLR) and CCA can be unified in a general eigenvalue decomposition formulation.

4. AUDIO-VISUAL FUSION WITH CCA

The acoustic feature used in this paper is the widely adopted Mel Frequency Cepstral Coefficient (MFCC). And the visual feature is the PCA projection of the whole face images. Although, CCA can learn the correlation between the original whole face image with

speech signal. The PCA analysis will increase the numerical stability of the CCA computation. Therefore, we apply PCA before CCA in this paper.

Since we are evaluating the proposed fusion method on an text-independent A-V speaker verification task. The Gaussian Mixture Model (GMM) [16] is used as modeling method to capture the characteristic of each speaker.

An M-mixture GMM is defined as a weighted sum of M component Gaussian densities

$$p(\bar{x}|\lambda) = \sum_{m=1}^{M} w_m \mathrm{N}(\bar{x}|\bar{\mu}_m, \Sigma_m), \tag{5}$$

where $\bar{x}$ is a D-dimensional feature vector, w_m is the m^{th} mixture weight, and $\mathrm{N}(\bar{x}|\bar{\mu}_m, \Sigma_m)$ is a multivariate Gaussian density, with mean vector $\bar{\mu}_m$ and covariance matrix Σ_m. Note that $\sum_{m=1}^{M} w_m = 1$.

A speaker model $\lambda = \{w_m, \bar{\mu}_m, \Sigma_m\}_{m=1}^{M}$ is obtained by fitting a GMM to a training utterance $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_T\}$ using the Expectation-Maximization (EM) algorithm. The log likelihood of a testing utterance $Y = \{\bar{y}_1, \bar{y}_2, ..., \bar{y}_T\}$ on a given speaker model λ is computed as follows,

$$LL(Y|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\bar{y}_t|\lambda), \tag{6}$$

where $p(\bar{y}_t|\lambda)$ is the likelihood of the t^{th} frame of the utterance. To identify an utterance as having been spoken by a person out of a group of N people, we compute its utterance scores against all N speaker models and pick the maximum

$$\hat{\lambda} = \arg\max_{\lambda_n} LL(Y|\lambda_n), \tag{7}$$

where λ_n is the model of the n^{th} speaker.

The GMM algorithm described above requires that every speaker model be trained independently with the speaker's training data. In the case when the available training data are limited for a speaker, the model is prone to singularity. To tackle this problem, the UBM-GMM algorithm [17], a different scheme, is adopted to train the speaker models. A single speaker-independent Universal Background Model (UBM) λ_0 is trained with a combination of the training data from all speakers, and a speaker model λ is derived by updating the well-trained UBM with that speaker's training data via Maximum A Posteriori (MAP) adaptation [17]. The final score of the testing utterance is computed by the log likelihood ratio between target model and background model.

$$LLR(Y) = LLR(\bar{y}_1^T) = \frac{1}{T} \sum_{t=1}^{T} \log \frac{P(\bar{y}_t|\lambda_1)}{P(\bar{y}_t|\lambda_0)}, \tag{8}$$

where $\bar{y}_1^T$ is the feature vector of the observed utterance—test utterance Y, λ_0 is the parameter of UBM and λ_1 is the parameter of target model. Essentially, the verification task is to construct a generalized likelihood ratio test between hypothesis H_1 (observation drawn from the target) and hypothesis H_0 (observation not drawn from the target).

The advantages of the UBM-GMM over the GMM are two-fold. First, the UBM is trained with a considerable amount of data and is thus quite well-defined. A speaker model, obtained by adapting the parameters of the UBM with a small amount of new data, is expected to be well-defined, too. Hence, the UBM-GMM approach should be robust to limited training data. Second, during adaptation, only a small number of Gaussian components of the UBM are updated.

- Audio only system
$$P(A|\lambda) \tag{9}$$
- Visual only system
$$P(V|\lambda) \tag{10}$$
- Feature level fusion system
$$P(A, V|\lambda) \tag{11}$$
- Decision Level fusion system
$$P(A|\lambda)^{\alpha} P(V|\lambda)^{1-\alpha} \tag{12}$$
- CCA fusion system
$$P(A, V|\lambda) = P(A_1, V_1|\lambda) P(A_2|\lambda) P(V_2|\lambda) \tag{13}$$

Within the UBM-GMM framework, audio only speaker modeling is combing the acoustic features (MFCCs) of all the speakers to train a audio only UBM model. Then MAP adaptation is applied to generate the model for each speaker. Visual only speaker modeling is combing all the visual features to train a visual only UBM model. The visual feature is a PCA projected of the whole face region. The PCA space is learned from the face images of all speakers.

In order to fuse the two modalities, shown in Eq. 9 and 10, for better performance, an A-V fusion module is applied to combine these two modalities. Eq.11 shows the feature-level fusion, which mainly concatenates the features from different modalities as a single big feature vector. State level fusion is to fuse the observation likelihood of different modalities on the same state. However, in the text-independent speaker recognition task, it is difficult to encoding the strict temporal information from audio and visual streams. Hence, it is difficult to apply state level fusion methods. To circumvent this difficulties, we compare CCA based fusion Eq.13 with the decision level fusion in Eq.12.

5. EXPERIMENTS AND RESULTS

A set of experiments are conducted to evaluate the proposed method. The database used in our experiments contains 102 subjects. For each subject, half hour of video was recorded in a studio environment. The speech scripts include connected-digits and continuous sentences. For text independent speaker identification, we randomly select 30sec of speech as training data for each speaker and 10sec speech as testing data. The corresponding videos are then tracked and cropped before the visual feature extraction. The tracking algorithm applied in this paper in based on the work by J. M. Buenaposada [18].

The demonstration experiment is about the correlation structure between the whole face region with the speech signal. For a video sequence containing the whole face region for one subject, the CCA analysis is performed between every local patches around a pixel position in the face region with the speech signal. If the CCA can find the true relation between facial regions and speech signals, the lip region should be the most correlated local patches among all the patches sampled in the whole face region. Figure 2 shows the result of this experiment. At each pixel position, the largest canonical correlations between the local patches and speech signals are recorded as the values at this position. Therefore, the result is actually a 2D surface on the image plane. By searching the maximum point in this surface, we can locate the maximum correlated local patches with speech signals. The right column illustrates this maximum correlated local patches which are truly the lip regions.

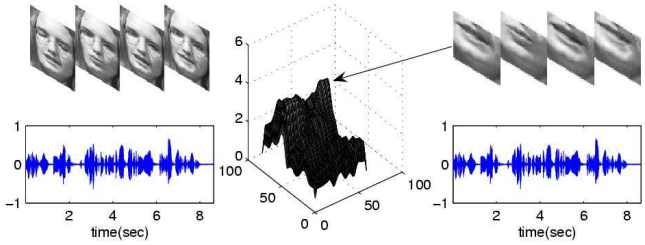

Fig. 2. CCA analysis is performed for the local patches covering the whole face region. The highest correlation score at each pixel position constructs the surface in the middle column. Clearly, lip patch is highlighted due to the high correlation between audio and visual streams.

Table 1. Experiments and Results of text-independent audio-visual speaker recognition. Each column denotes different component UBM models. Performance is measured in recognition accuracy. Row "A" shows the results of audio-only system. Row "V" shows the results of visual-only system; Row "A-V(feature)" shows the results of feature level fusion system; Row "A-V(decision)" shows the results of decision level fusion system; Row "A-V(CCA)" shows the CCA based fusion system.

Accuracy(%)	4	8	16	32
A	42.16	56.86	74.51	57.84
V	35.29	92.11	83.33	91.18
A-V(feature)	44.97	91.75	85.47	90.77
A-V(decision)	46.81	92.19	87.85	92.43
A-V(CCA)	**91.18**	**98.04**	**99.02**	**94.12**

To verify the CCA based fusion method, the text-independent speaker recognition experiments are conducted on a video database, which contains 102 subjects. The video is up sampled to 100 frames per sec to match the speech frame rate. The results are shown in Table 1. From these results, the CCA based fusion clearly achieves the best improvement by combing the audio and visual modalities. When the model is a 8-component GMM-UBM framework, the CCA based fusion achieves 98.04% accuracy. Notice, at the 4-component GMM-UBM setting, the audio-only and visual-only systems can only achieve less than 50% accuracy, while the fusion based on CCA boosts the performance to 91.18%. The A-V(CCA) achieves its highest performance by 99.02% in the 16-component GMM-UBM setting. These results clearly confirm that the CCA based fusion achieves much better modelling efficiency.

6. CONCLUSION AND DISCUSSION

In this paper, we proposed a novel A-V fusion framework based on joint dimensionality reduction. With this framework, the A-V correlation can be preserved in the feature extraction procedure, which makes the approach more flexible to find details of the A-V correlation. By formulating the joint dimensionality reduction problem in the framework of canonical correlation analysis [11], we obtained very efficient and stable joint dimensionality reduction technique for A-V fusion. The experimental results show that this method successfully capture the correlation between A-V modalities. In recognition experiments, CCA based fusion achieves the best fusion improvement over than conventional fusion methods. With this initial success of CCA based fusion technique, more experiments are expected in future to fully explore the properties of the proposed method.

7. REFERENCES

[1] E.D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. of Global Telecommunications Conference*, 1984, pp. 265–272.

[2] Y. Gong, "Speech recognition in noisy environment: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[3] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing*. MIT, 2004.

[4] M. T. Chan, Y. Zhang, and T. S. Huang, "Integrating visual and acoustic features for speech recognition.," in *Proc. of Army Research Laboratory Advanced Displays and Interactive Displays, Federated Laboratory Second Annual Symposium*, 1998, pp. 64–68.

[5] A. Adjoudani and C. Benoit, *Speechreading by Humans and Machines*, chapter On the integration of auditory and visual parameters in an HMM-based ASR, pp. 461–471, 1996.

[6] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, pp. 141–151, 2000.

[7] M. Liu, Z. Xiong, S. M. Chu, Z. Zhang, and T. S. Huang, "Audio visual word spotting," in *Proc. of IEEE Conf. on ICASSP'04*, 2004, pp. 785–788.

[8] S.M. Chu and T.S. Huang, "Audio-visual speech modeling using coupled hidden markov models," in *Proc. of IEEE Conf. on ICASSP'02*, 2002, pp. 2009–2012.

[9] S.M. Chu and T.S. Huang, "An experimental study of coupled hidden markov models," in *Proc. of IEEE Conf. on ICASSP'02*, 2002, pp. 4100–4103.

[10] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual lvcsr," in *Proc. of IEEE Conf. on ICME'01*, 2001, pp. 825–828.

[11] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[12] S. Becker, "Mutual information maximization: models of cortical self-organization network," *Computation in Neural Systems*, vol. 7, pp. 7–31, 1996.

[13] J. Kay, "Feature discovery under contextual supervision using mutual information," in *Proc. of International Joint Conference on Neural Networks*, 1992, vol. 4, pp. 79–84.

[14] P. W. Fieguth, W. W. Irving, and A. S. Willsky, "Multiresolution model development for overlapping trees via canonical correlation analysis," in *Proc. of IEEE Conf. on ICIP'95*, 1995, pp. 45–48.

[15] P. Li, J. Sun, and B. Yu., "Direction finding using interpolated arrays in unknown noise fields," *Signal Processing*, vol. 58, pp. 319–325, 1997.

[16] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[17] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[18] J.M. Buenaposada, E. Munoz, and L. Baumela, "Efficient appearance-based tracking," in *Proc. of IEEE Workshop on Articulated and Nonrigid Motion (with IEEE CVPR'04)*, 2004.