

Clip Retrieval using Multi-modal Biometrics in Meeting Archives

Himanshu Vajaria, Sudeep Sarkar and Rangachar Kasturi
Dept. of Computer Science & Engineering,
University of South Florida, Tampa, FL, F33620, USA
{hvajaria,sarkar,r1k}@cse.usf.edu

Abstract

We present a system to retrieve all clips from a meeting archive that show a particular individual speaking, using a single face or voice sample as the query. The system incorporates three novel ideas. One, rather than match the query to each individual sample in the archive, samples within a meeting are grouped first, generating a cluster of samples per individual. The query is then matched to the cluster, taking advantage of multiple samples to yield a robust decision. Two, automatic audio-visual association is performed which allows a bi-modal retrieval of clips, even when the query is uni-modal. Three, the biometric recognition uses individual-specific score distributions learnt from the clusters, in a likelihood ratio based decision framework that obviates the need for explicit normalization or modality weighting. The resulting system, which is completely automated, performs with 92.6% precision at 90% recall on a dataset of 16 real meetings spanning a total of 13 hours.

1 Introduction

Lately, many organizations have begun recording and archiving their meetings for reference. The growing size of such archives necessitates an efficient system to retrieve relevant clips in response to a user's query. Here, we present a system to handle one particular kind of query - to find all clips in the archive where a particular person spoke. This requires (a) segmenting each meeting into such clips, (b) finding the speaker's face, and (c) comparing the query face-voice sample to samples from these clips. Although these problems are well addressed in the literature, we propose novel approaches to each of these tasks.

Clustering speech from different individuals in a recording is commonly referred to as the *speaker diarization* problem. In contrast to most solutions that use only the speech signal [11] to solve this task, this work also incorporates video information.

Various approaches based on Mutual Information (MI) [5, 6, 9] have been proposed to associate the speaker's face and voice in short clips where the speakers are fac-

ing the camera. However, in meeting scenarios, where the speaker's face is often non-frontal or occluded, the instantaneous synchrony assumption on which MI approaches are based does not hold. This system uses a novel eigen-analysis approach [13], that exploits long-term speech-movement co-occurrences to find the image region corresponding to the speaker and detect the face from this region.

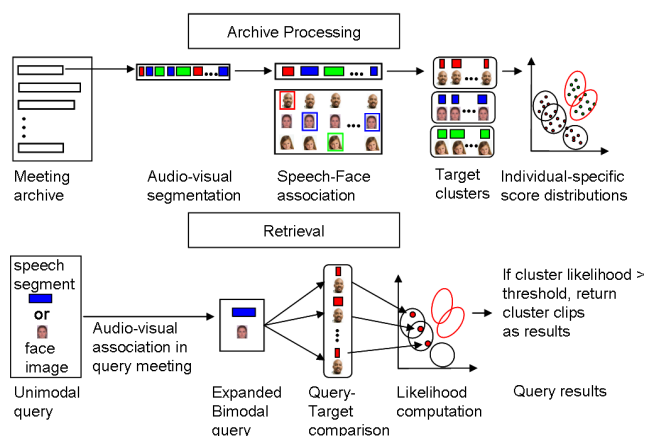


Figure 1. System schema

Once face and speech samples from a clip are found, they are compared with the query face/voice sample. However, since clips from the same individual are clustered together, the query can be matched to multiple samples of the cluster for a robust decision. Such multi-sample, multimodal biometric recognition has been addressed in works such as [1, 4, 10], where individual-specific modality weights, polynomial discriminant functions or support vector machines (SVMs) are employed. The likelihood ratio based decision framework [7], which was found superior to other approaches [12], is used here and adapted to individual-specific distributions. In addition, unlike other works that use a separate training set to learn individual-specific distributions, the distributions are automatically learnt from the clusters.

The system follows a query by example framework as illustrated in Figure 1. Meetings in the archive are first parti-

tioned into clips such that only a single person is speaking in each clip. Audio-visual association and clustering of these clips yields a cluster for each individual that contains all of the individual’s speech segments and associated face images. For each cluster, a bi-variate match score distribution is generated by comparing all face-speech samples within the cluster and similarly a non-match score distribution is generated by matching samples of the cluster with samples in other clusters. A user browsing a particular meeting, might choose a person’s face image or a speech sample to request all clips from the archive where that person is speaking. Audio-visual association is used to find a sample from the missing modality from the query meeting, expanding the uni-modal query into a bi-modal query. The query sample is matched to each face-voice sample in the cluster and a likelihood ratio framework is used to determine if the cluster matches the query.

2 Audio-Visual association and clustering

The first processing step is to split the meeting into clips that contain speech from only a single person using the joint audio-visual stream. The motivation for using both audio and video features to determine speaker change-points stems from the observation that speech and movement are coexpressive [14]. Typically, since a speaker exhibits more movement than a listener, a change in speaker is also accompanied by a change in the image region where movement occurs.

Mel-frequency cepstral coefficients are extracted at 30 Hz using 32 filters with the bandwidth ranging from 166 Hz to 4000 Hz. The video features, which intend to capture motion, are obtained using image differences (three frames apart), thresholded to suppress jitter noise, dilated by a 3x3 circular mask, downsampled from the original size of 480x720 to 48x72, and finally vectorized. The audio/video features are then projected onto PCA subspaces to reduce their dimensionality, and a joint audio-visual subspace is obtained by concatenating the coefficients using Equation 1.

$$X(t) = \begin{bmatrix} \alpha A(t) \\ V(t) \end{bmatrix} \quad (1)$$

Here $A(t)$ and $V(t)$ are PCA projections of the audio and video signal, respectively. The scaling factor $\alpha = |\Sigma_V|/|\Sigma_A|$ is used to make the variance of the audio features equal to that of the video features.

The task of finding clip boundaries in X , is cast into the model selection framework. We want to determine whether the feature set $X_C = (X_1, X_2 \dots X_N)$ is better represented by a single model M_C or whether there exist two models M_1 and M_2 that can better represent feature sets $X_{W_1} = (X_1, X_2 \dots X_i)$ and $X_{W_2} = (X_{i+1}, X_{i+2} \dots X_N)$ respectively.

Defining ΔBIC to be $BIC(M_1; M_2) - BIC(M_C)$, we have

$$\Delta BIC(i) = \frac{i}{2} \log |\Sigma_{X_{W_1}}| + \frac{N-i}{2} \log |\Sigma_{X_{W_2}}| - \frac{N}{2} \log |\Sigma_{X_C}| - \frac{i}{2} \lambda \left(d + \frac{d(d+1)}{2} \right) \log N \quad (2)$$

Here, $|\Sigma_{X_C}|$ is the determinant of the covariance of the entire feature set and $|\Sigma_{X_{W_1}}|, |\Sigma_{X_{W_2}}|$ are the determinant of the covariance of the two subdivisions of the feature sets. The last term induces a penalty on complex models proportional to the dimensionality of the model d , and λ is an empirical constant, usually set to 1. A positive $\Delta BIC(i)$ indicates that the data is better represented by two models. The frame i for which $\Delta BIC(i)$ is maximum represents the best split of the feature set, indicating a clip boundary. The implementation follows a sliding window approach, where X_C spans six seconds and is progressively shifted along the entire length of X .

The next goal is to generate target clusters, one for each speaker that contain all speech segments for that speaker and find the associated face images. Clips from a meeting in which the same person is speaking are first grouped by performing temporal clustering of the audio signal. This is followed by spatial clustering in the corresponding video frames using a novel eigen-analysis method [13] to find the dominant modes of motion, which are different image regions belonging to the different participants. Since a speaker exhibits more movement than the listeners, the dominant motion mode will be the one associated with the speaker. Thus an association between the person’s speech and location (image region) is found. A Haar face detector is run on this region to find the speaker’s face. The extracted speech and face samples from a cluster’s clips are now compared to the query sample as described in the next section.

3 Modeling Individual Score Distributions

Significant effort has been devoted to learning user-specific modality weights and thresholds [1, 10]. Most of the focus has been on learning user-specific discriminative models and using SVMs for classification. Recently, a generative framework, that uses a likelihood ratio based decision was proposed [7], where score distributions are modeled by a mixture of Gaussians. Amongst other advantages, the framework can handle arbitrary score ranges and distributions of different matchers, and correlation between their scores. In this work, we extend their likelihood ratio based fusion framework to learning individual-specific models, to provide better discrimination between individuals and hence better retrieval performance.

Ideally, each cluster contains all clips from a meeting where the same individual is speaking. Of the face images associated with each clip’s audio, we randomly chose one face image as the individual’s face sample and the entire

audio portion of the clip as the speech sample. Thus each cluster now contains multiple face-speech samples (one per clip) for each user. Bi-variate match scores are generated by matching all face-speech samples within a cluster to each other. Similarly, non-match scores are generated by matching a cluster’s samples to samples from all other clusters of the same meeting. Gaussian Mixture models (GMMs), θ_c^M and θ_c^{NM} , are then learned [2] from the match and non-match score distributions of cluster c .

To make a query-cluster match decision, each query face-voice sample is matched to all face-voice pairings in a target cluster. A commercial face recognition algorithm is used to generate distances between face samples and a KL2-GMM speaker recognition algorithm [8] is used for computing distances between speech samples. Let $S_{qc_i} = [F_{qc_i} V_{qc_i}]$ represent the generated bivariate score obtained by matching the query q with the i^{th} sample from a target cluster c . The log likelihood ratio of the cluster belonging to the query LL_{qc} is computed using 3 and if it exceeds a threshold, all clips of c are considered to match the query.

$$LL_{qc} = \sum_{\forall i \in c} \log \frac{p(S_{qc_i} | \theta_c^M)}{p(S_{qc_i} | \theta_c^{NM})} \quad (3)$$

For evaluating global bi-modal models, we replace θ_c^M and θ_c^{NM} with θ_G^M and θ_{NG}^M respectively, which are GMMs learnt on the agglomeration of match and non-match scores from all clusters. Similarly, multi-sample uni-modal performances are evaluated using global and individual-specific GMMs learnt from the marginal face and voice score distributions.

4 Results

The query system is tested on 16 meetings from the NIST pilot meeting room archive [3]. Each meeting is recorded from four different camera views and the audio channel consists of a gain normalized mix from multiple distant microphones. The video frame-rate is 29.97 Hz with a spatial resolution of 720 x 480 and the audio data is sampled at 44 kHz and has a resolution of 16 bits per sample. The number of participants in each meeting varies from three to nine and the total number of unique participants in the dataset is 50. The number of different meetings that a person participates in varies from one to six.

The clip boundary detection performance using only audio (A) is compared to that using both audio and video (AV_x) in Table 1. Four joint audio-visual streams are generated by combining the audio channel with each of the four camera feeds independently. A consolidated boundary is generated if clip boundaries occur in two or more of the audio-visual streams within a two second window. This eliminates spurious boundaries occurring in the individual audio-visual streams reducing the false detection

rate (FDR), while simultaneously reducing the missed detection rate (MDR) by detecting boundaries missed in single streams. A forgiveness collar of 0.5 seconds was used when computing the MDR/FDR to account for imprecisions in the ground-truth. The high FDR implies that clips are often incorrectly over-segmented, but since clips are clustered in the later stages, this is acceptable.

Table 1. Clip boundary detection performance (%).

	A	AV_1	AV_2	AV_3	AV_4	AV_C
MDR	13.91	7.31	5.63	6.16	5.23	2.17
FDR	39.34	34.95	35.71	37.5	33.93	23.20

Each clip is characterized by its speech and a randomly chosen face image from the speaker’s location in the video. For unimodal samples, the precision-recall curves shown in Figure 2 are generated by considering each sample as the query and matching it to the remaining samples. The precision of the face and voice systems at 90% recall is 64% and 71%, respectively. Z-normalization followed by sum rule fusion of the two modalities results in a precision of 80% at 90% recall, which highlights the improvement due to multiple modalities.

Figure 3 shows the performances of the uni-modal and bi-modal systems using global and individual-specific models. The precision for face and voice using global models is 79.1% and 78.8%, respectively at 90% recall, which shows how multiple samples can improve the performance of single modalities. Using individual-specific models results in better class discrimination, improving the performance to 83.7% and 85.5%, respectively at the same recall. The bi-modal performance using global models is 88.4% precision at 90% recall. The best performance is obtained by using individual-specific models and both modalities, which exploits multiple samples, multiple modalities and individual specific score modeling. The precision in this case is 92.6% at 90% recall.

Figure 4 shows sample clips retrieved using either the face or the speech as a query for two of the participants. Of the approximately 250 retrieved clips, samples from only four clips are shown. One of the strengths of this system is that it can retrieve clips in which the speaker’s face is non-frontal even when the query is a frontal face. It does this by exploiting the speech associated with the query face.

5. Conclusions

This paper presents a fully automatic system for querying meeting archives to find speech segments from the same person. Novel solutions have been proposed for the three problems involved. To segment the meeting into clips containing speech from a single speaker, both audio and video are used, unlike previous works that rely only on audio.

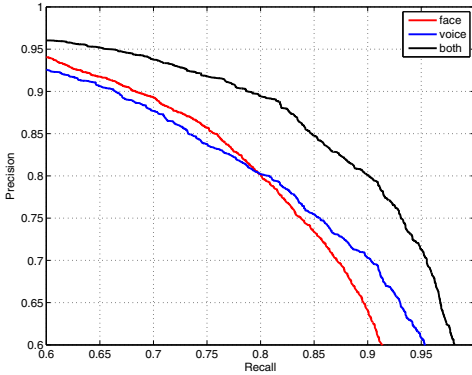


Figure 2. Uni-sample Retrieval Performance

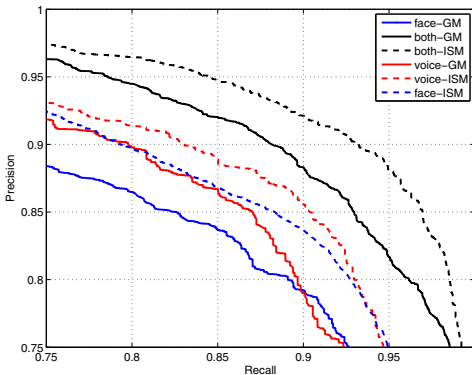


Figure 3. Multi-sample Retrieval Performance

The system uses a novel eigen-analysis approach to automatically associate face and speech samples, unlike previous work in biometric recognition, where this association is performed manually. Finally, a novel matching framework was introduced that exploits the nature of meeting videos by clustering multiple samples from an individual and learning individual-specific, joint face-voice distributions. By directly using the individual-specific models to generate likelihood ratios, the system performs implicit normalization and user-specific modality weighting. The resulting system has a precision of 92.6% at 90% recall when tested on an archive of 16 real meetings spanning a total of 13 hours.

References

- [1] J. Aguilar, D. Romero, J. Garcia, and J. Rodriguez. Adapted user-dependent multimodal biometric authentication exploiting general information. *PRL*, 26:2628–2639, 2005.
- [2] M. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3):381–396, 2002.
- [3] J. S. Garofolo, C. D. Laprun, M. Michel, V. M. Stanford, and E. Tabassi. The NIST meeting room corpus. *LREC*, 2004.
- [4] A. K. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. *Proceedings of the International Conference on Image Processing (ICIP)*, pages 57–60, 2002.
- [5] E. Kidron and Y. Schechner. Pixels that sound. *CVPR*, pages 88–95, 2005.

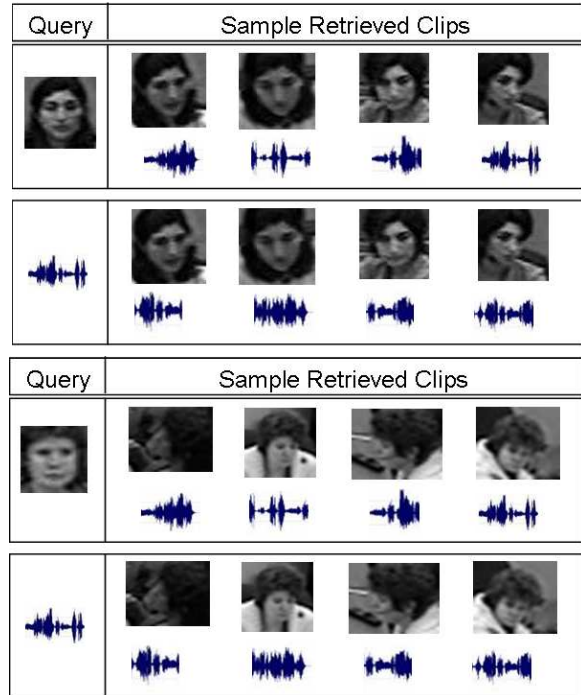


Figure 4. Sample Results for Clip Retrieval for two subjects. Associating speech allows the system to retrieve clips containing non-frontal faces even when the query is a frontal face.

- [6] G. Monaci, Òscar Divorra Escoda, and P. Vanderghyest. Analysis of multimodal sequences using geometric video representations. *Signal Processing*, pages 3534–3548, 2006.
- [7] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain. Likelihood ratio based biometric score fusion. *PAMI*, 2008.
- [8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41, 2000.
- [9] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. *NIPS 2000*, 14, 2001.
- [10] K. Toh, X. Jian, and W. Yau. Exploiting global and local decisions for multimodal biometrics verification. *IEEE Transactions on Signal Processing*, 52(10):3059–3072, 2004.
- [11] S. Tranter and D. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on audio speech and language processing*, 14(5):1557–1565, 2006.
- [12] B. Ulery, A. Hicklin, C. Watson, W. Fellner, and P. Hallinan. Evaluation of selected biometric fusion techniques. studies of biometric fusion. *NIST, Tech. Rep. IR 7346*, 2006.
- [13] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi. Audio segmentation and speaker localization in meeting videos. *ICPR*, 2:1150–1153, 2006.
- [14] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll. Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures. *Proceedings of European Signal Processing Conference*, pages 75–78, 2002.