MULTIMODAL MEETING EVENT RECOGNITION FUSING THREE DIFFERENT TYPES OF RECOGNITION TECHNIQUES

Stephan Reiter and Gerhard Rigoll

Institute of Human-Machine-Communication
Technische Universitaet Muenchen
Arcisstr. 21, 80290 Muenchen, Germany
email: {reiter, rigoll}@ei.tum.de

Abstract

In this paper we introduce a new method for recognizing meeting events. In the present case the boundaries of the meeting segments are a priori known. The recognition task is performed using a classifier fusion technique that combines the three different used approaches for meeting event recognition. The results show that by classifier fusion a more stable result can be achieved than using only one classifier.

For our experiments with meeting segmentation and meeting event recognition we make use of special scripted meetings that were recorded in the IDIAP Smart Meeting Room [1]. These recorded meetings consist of a set of predefined meeting events in a specific order. The appearing events were: Monologue1 to Monologue4 (one of the four participants (1 to 4) speaks continuously), discussion (all participants engage in a discussion), note-taking (all participants write notes), white-board (one participant at front of room, makes notes on the white board), presentation (one participant at front of room makes a presentation), consensus (all participants express consensus) and disagreement (all participants express disagreement). Thus we differentiate 10 different events.

The classification task is performed by three different approaches, that are combined via Late Semantic Fusion: A static approach using simulated results of specialized recognizers, a dynamic approach using the audio files, and a dynamic approach using the transcriptions. The basic idea of the first approach is to take advantage of the results of various specialized recognizers, like gesture recognizers, person trackers and so on. As long as these recognizers only exist at a developing level we have to simulate them. This has been done for all scripted meetings by annotating the various actions that each person is performing e.g. talking, writing, pointing, standing etc. For each period of time global statistics are calculated that tell the percentage of an action in that period. Carefully selected items are then put together into a feature vector. The classification of an unknown feature vector is performed by a multi-class Support Vector Machine.

The second approach uses only the audio files. From the four lapel files (one of each participant) the Mel-Frequency-Cepstral-Coefficients (MFCCs) are calculated [2]. Here we use twelve cepstral coefficients plus the energy. These thirteen features are calculated for each participant and then concatenated. So we get our feature vector with 52 dimensions. The MFCCs are extracted every 10 milliseconds. All these features are then trained by an Hidden-Markov-Model with six states and continuous Gaussian mixture outputs.

Our third approach for recognizing meeting events is based on the transcriptions that are available for a couple of these scripted meetings. Following suggestions made by [3] each word is assigned a probability. Also for each word the conditioned probability that it belongs to a specific class is calculated. Then with the use of the Bayes' Rule the conditioned probability for a sequence of words belonging to a specific class is derived.

A comparison of the three different classificaton techniques gives an interesting result. If only the video was available without audio signal and the single actions of the participants could be extracted perfectly, then the recognition rate is already farely high with 82.79%. If only the audio files were used, the recognition rate decreases significantly by about 14 percent. This is probably due to the loss of the local information that was needed to distinquish e.g. monologues from presentation events. The even worser result by using the transcriptions could be explained by the lack of the information, who is saying what, because there only the words themselves are considered, but not who uttered them. So a combination of all of this three methods should give better results because the classifiers use complementary information.

Each instance of the three specified meeting recognition techniques produces an output, in which the most likely meeting event is reported. In addition a score is delivered that declares how reliable this result can be. To reach better results than each of the classifiers alone a simple fusion technique is used: If two or more of the instances deliver the same class then the fused result is that class. Only if all three classifier say different things, then the one with the highest score is considered best. With this fusion technique a gain of about three percent in the recognition rate was obtained.

Classifier	Annotations	MFCCs	Transcripts	Fused
Recognition Rate	82.79 %	68.03 %	44.44 %	86.07 %

Although this rather simple fusion technique improves the overall recognition rate it could be possible to get a better result by using Early Signal Fusion instead of the proposed Late Semantic Fusion. The main challenge is to combine feature streams with highly different time rates. For this task we think of using hierarchical HMMs for the combination of the audio features and the transcriptions. To incorporate the global statistics from the annotations we consider using some kind of graphical model that can be represented by Dynamic Bayesian Networks (DBNs) with an adapted topology that allows to combine feature streams with different time rates. On our poster presentation we are planning to present first results using this new promising approach.

References

[1] D. Moore, "The idiap smart meeting room," IDIAP-COM 07, IDIAP, 2002.

[2] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland, *The HTK Book,* Cambridge University Engineering Department, 2002.

[3] F. Walls, H. Jin, S. Sista, and R. Schwartz, "Topic detection in broadcast news," in *Proceedings of the DARPA Broadcast News Workshop*, 1999, pp. 193–198.

From: MLMI 04 JOINT AMI/PASCAL/IM2/M4