AN INVESTIGATION OF DIFFERENT MODELING TECHNIQUES FOR MULTI-MODAL EVENT CLASSIFICATION IN MEETING SCENARIOS

Marc A. Al-Hames and Gerhard Rigoll
Technische Universität München
Institute for Human-Machine Communication
Arcisstrasse 21, D-80333 München, Germany
{alh, rigoll}@mmk.ei.tum.de

In this work a hidden Markov model (HMM) and a multi-stream HMM are compared with a new dynamic Bayesian network (DBN) approach for multi-modal event classification in meeting scenarios. A set of 60 meetings - each with four participants - has been recorded at IDIAP [1]. Given segments of these meetings have been categorized to one of ten different states: consensus, disagreement, discussion, monologue (for each of the four speaker), note-taking, whiteboard, and presentation.

The classification was performed with 68 features: for each speaker four MFCC coefficients and the energy were calculated from the speech signal. A speech and silence segmentation for each of the six locations in the meeting room was extracted with the SRP-PHAT measure [2]. Furthermore seven global motion features [3] (center of motion, wideness of motion, etc.) were obtained for each location from the visual recording. For the HMM approaches the audio feature rates were adjusted to 12.5 Hz to match the visual feature stream.

For the first approach a single HMM has been trained with all features in a single vector. This corresponds to an early fusion process.

In the second approach a HMM with three streams for the audio, the binary speech profile and the visual stream has been used. The three streams were combined at specific temporal points, resulting in a single probability for the event [2]. Thereby different stream weights have been evaluated. However no significant improvements have been achieved with the multi-stream HMM.

The new approach presented in this work uses DBNs for the classification [4]. The model is shown in figure 1. The binary speech profile stream has discrete observation states (O_i^{Stream}) and the speech and the visual stream have Gaussian mixture observation states (M_i^{Stream}). With this approach each observation stream can be modeled separately and show a different frame rate (here the video stream has half the frame rate than the audio streams). However, the three streams are still connected and exchange information over their hidden states (H_i^{Stream}).

Evaluations show promising preliminary results for the DBN approach compared to the HMM approaches. The final results will be presented on the poster.

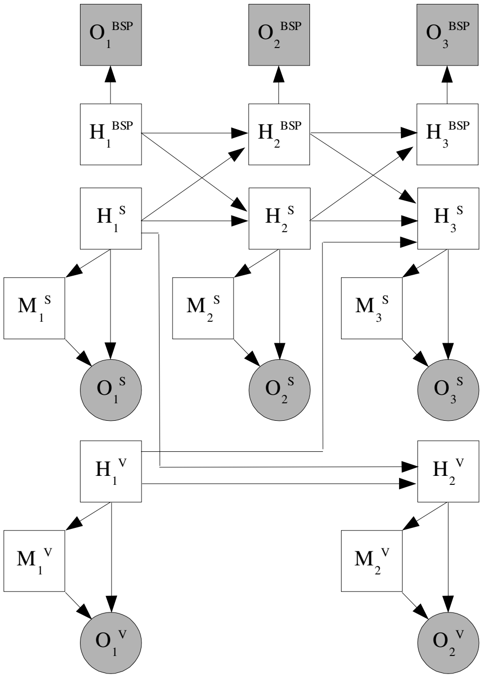

Fig. 1. A DBN with three connected streams.

References

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

[2] I. McCowan et. al., "Automatic analysis of multimodal group actions in meetings," IDIAP-RR 27, IDIAP, 2003.

[3] M. Zobl, F. Wallhoff, and G. Rigoll, "Action recognition in meeting scenarios using global motion features," in *Proceedings Fourth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ICVS)*, J. Ferryman, Ed., 2003, pp. 32–36.

[4] M. I. Jordan, Ed., *Learning in Graphical Models*, MIT Press, 1998.

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