Robust Meeting Event Recognition with a Multi-Modal Mixed-State Graphical Model

Marc Al-Hames and Gerhard Rigoll

Technische Universität München
Institute for Human-Machine Communication
Arcisstrasse 16, 80333 München, Germany
{alh,rigoll}@mmk.ei.tum.de

Introduction

Meetings are social events, were people exchange information. Often a summarization of the meeting is necessary, for example for people not attending the meeting or to fix decisions. A first step for the automatic analysis of the meetings is a segmentation into meeting group action events like discussion or presentation [4]. This structuring can then be used to produce an agenda and a summarization of the meeting. Different approaches for this structuring, based on hidden Markov models (HMMs) [4] and dynamic Bayesian networks (DBNs) [5, 3] have been introduced for clear data sets. However, in real meetings the data can be disturbed in various ways: events like slamming of a door may mask the audio channel or background babble may appear; the visual channel can be (partly) masked by persons standing or walking in front of a camera. In this work we present a novel multi-modal mixed-state dynamic Bayesian network (DBN) [1, 2] for robust meeting event classification from disturbed videos.

Features

The model uses information from the audio and the visual channel to structure meetings into segments: four MFC coefficients from the lapel microphones per speaker, a binary speech and silence segmentation obtained with the SRP-PHAT measure [4] from the microphone array, and global motions [6] as visual feature have been used.

The multi-stream mixed-state DBN model

The multi-stream mixed-state DBN is a multi-stream HMM coupled with a linear dynamical system (LDS) to compensate disturbances in the visual channel. Thereby the HMM is used as driving input for the LDS. This can be described as a graphical model, as shown in in Fig. 1. Each row represents one time slice. Arrows pointing down represent the dependencies between subsequent time slices. Horizontal arrows represent dependencies between hidden and observed variables within one time slice. Hidden variables are white, observed variables shadowed. Squares mark discrete probability distributions and circles denote continuous Gaussian nodes $\mathcal{N}(\boldsymbol{x}, \boldsymbol{\mu}, \Sigma)$. It is this special structure, that allows the model to handle noise and occlusions in the video.

From: MLMI 2005, Edinburgh, UK

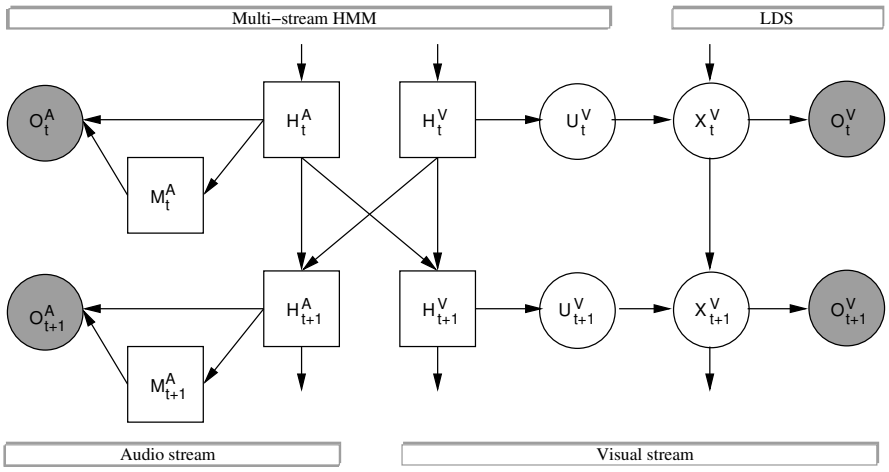

Fig. 1. Multi-stream mixed-state DBN: The HMM is driving input for the LDS.

Results

The model has been compared to three single modal HMMs (MFCC, microphone array, and visual), to a multi-modal early fusion HMM, and a multi-stream HMM. Experimental results on real meeting data show that the new model is highly preferable to single-stream (audio, resp. video) HMM approaches. Furthermore, compared to the baseline multi-modal HMMs, the new DBN is up to 3.5%, respectively up to 6.1% better for clear and visual disturbed data, this corresponds to a relative error reduction of 23.6%, respectively 29.9%.

References

[1] Marc Al-Hames and Gerhard Rigoll. A multi-modal graphical model for robust recognition of group actions in meetings from disturbed videos. In *Proc. IEEE ICIP*, September 2005.

[2] Marc Al-Hames and Gerhard Rigoll. A multi-modal mixed-state dynamic Bayesian network for robust meeting event recognition from disturbed data. In *IEEE ICME*, July 2005.

[3] A. Dielmann and S. Renals. Dynamic Bayesian networks for meeting structuring. In *Proc. IEEE ICASSP*, Montreal, Canada, 2004.

[4] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *Proc. IEEE ICASSP*, Hong Kong, April 2003.

[5] S. Reiter and G. Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proc. IEEE ICASSP*, Philadelphia, USA, March 2005.

[6] F. Wallhoff, M. Zobl, and G. Rigoll. Action segmentation and recognition in meeting room scenarios. In *IEEE ICIP*, October 2004.