A MULTI-MODAL MIXED-STATE DYNAMIC BAYESIAN NETWORK FOR ROBUST MEETING EVENT RECOGNITION FROM DISTURBED DATA

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

ABSTRACT

In this work we present a novel multi-modal mixed-state dynamic Bayesian network (DBN) for robust meeting event classification. The model uses information from lapel microphones, a microphone array and visual information to structure meetings into segments. Within the DBN a multi-stream hidden Markov model (HMM) is coupled with a linear dynamical system (LDS) to compensate disturbances in the data. Thereby the HMM is used as driving input for the LDS. The model can handle noise and occlusions in all channels. Experimental results on real meeting data show that the new model is highly preferable to all single-stream approaches. Compared to a baseline multi-modal early fusion HMM, the new DBN is more than 2.5%, respectively 1.5% better for clear and disturbed data, this corresponds to a relative error reduction of 17%, respectively 9%.

1. 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. Nowadays these summarizations are mainly written by a person attending the meeting. This process is both time demanding and error-prone.

Thus it would be good, if meetings could be summarized automatically. Projects like the ICSI meeting project [1] and "Augmented Multi-party Interaction (AMI)" deal with this topic of automatic speech transcription, analysis of videos, and summarization of meetings.

A first step for the automatic analysis of the meetings is a segmentation into meeting group action events like discussion or presentation [2]. This structuring can then be used to produce a low level agenda and a summarization of the meeting. Different approaches for this structuring, based on hidden Markov models (HMMs) [2, 3] and dynamic Bayesian networks (DBNs) [4] 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, or a laptop computer may stand in front of the persons.

In this work we present a novel multi-modal approach for meeting event recognition, based on mixed-state DBNs, that can handle noise and occlusions in all channels.

2. MEETING DATA

The data for this work was collected in the IDIAP smart meeting room [5]. The corpus consists of 60 videos with a length of approximately 5 minutes. Each meeting has four participants and is recorded with three cameras. All participants have a lapel microphone attached and a microphone array is placed on the table. Thus, the corpus provides high quality audio-visual recording of the meetings.

To investigate the influence of disturbance to the recognition performance, the evaluation data was cluttered: the video data was occluded with a black bar covering one third of the image at different positions. The audio data from the lapel microphones and the microphone array was disturbed with a background-babble with 10dB SNR.

In this work 30 undisturbed videos were used for the training of the models. The remaining 30 unknown videos have been cluttered for the evaluation.

3. GROUP ACTION MEETING EVENTS

In the recorded corpus each meeting has four participants:

$$S = \{S_1, S_2, S_3, S_4\}$$

For a first structuring of the meeting the following eight different group actions are widely used [2, 3, 4]:

$$E = \{E_D, E_{M,1}, E_{M,2}, E_{M,3}, E_{M,4}, E_N, E_P, E_W\}$$

From: ICME 2005, Amsterdam, IEEE

where the events E_j are

E_D: Two or more persons are talking with each other.
$E_{M,Id}$: The person Id is talking without being interrupted.
E_N: All persons write something down.
E_P: One person in front of the room gives a presentation.
E_W: One person writes on the whiteboard.

Each meeting can now be modeled as a sequence of these group actions E_j. In average each meeting in the corpus consists of five action segments. This sequence of actions can then be used as a rough structuring of the meeting [2].

4. FEATURES

Feature vectors have been extracted from the audio-visual stream. In the meeting room the four persons are expected to be at one of six different locations: one of four chairs, the whiteboard, or at a presentation position:

$$L = \{C_1, C_2, C_3, C_4, W, P\}$$

This information has been used to extract position dependent audio- and visual-features. The signals from the lapel-microphones have been used to add speaker dependent audio features. Altogether 68 features from three modalities: microphone array, lapel microphone, and visual information have been used.

4.1. Audio features

For each of the speakers four MFC coefficients and the energy were extracted from the lapel-microphones. This results in a 20-dimensional vector $\vec{x}_S(t)$ containing speaker-dependent information. A binary speech and silence segmentation (BSP) for each of the six locations in the meeting room was extracted with the SRP-PHAT measure [3] from the microphone array. This results in a six-dimensional discrete vector $\vec{x}_{BSP}(t)$ containing position dependent information.

4.2. Visual features

For each of the six locations L in the meeting room a difference image sequence $I_d^L(x, y)$ is calculated by subtracting the pixel values of two subsequent frames from the video stream. Then seven global motion features [6] are derived from the image sequence: The center of motion is calculated for the x- and y-direction according to:

$$m_x^L(t) = \frac{\sum_{(x,y)} x \cdot |I_d^L(x, y, t)|}{\sum_{(x,y)} |I_d^L(x, y, t)|}$$

and

$$m_y^L(t) = \frac{\sum_{(x,y)} y \cdot |I_d^L(x, y, t)|}{\sum_{(x,y)} |I_d^L(x, y, t)|} \quad (1)$$

The changes in motion are used to express the dynamics of movements:

$$\Delta m_x^L(t) = m_x^L(t) - m_x^L(t-1)$$

and

$$\Delta m_y^L(t) = m_y^L(t) - m_y^L(t-1) \quad (2)$$

Furthermore the mean absolute deviation of the pixels relative to the center of motion is computed:

$$\sigma_x^L(t) = \frac{\sum_{(x,y)} |I_d^L(x, y, t)| \cdot \left(x - m_x^L(t)\right)}{\sum_{(x,y)} |I_d^L(x, y, t)|}$$

and

$$\sigma_y^L(t) = \frac{\sum_{(x,y)} |I_d^L(x, y, t)| \cdot \left(y - m_y^L(t)\right)}{\sum_{(x,y)} |I_d^L(x, y, t)|} \quad (3)$$

Finally the intensity of motion is calculated from the average absolute value of the motion distribution:

$$i^L(t) = \frac{\sum_{(x,y)} |I_d^L(x, y, t)|}{x \cdot y} \quad (4)$$

These seven features are concatenated for each time step in the location dependent motion vector

$$\vec{x}^L(t) = [m_x^L, m_y^L, \Delta m_x^L, \Delta m_y^L, \sigma_x^L, \sigma_y^L, i^L]^T \quad (5)$$

With this motion vector the high dimensional video stream is reduced to a seven dimensional vector, but it preserves the major characteristics of the currently observed motion. Concatenating the motion vectors from each of the six positions $\vec{x}^L(t)$ leads to the final visual feature vector

$$\vec{x}_V(t) = [\vec{x}^{C_1}, \vec{x}^{C_2}, \vec{x}^{C_3}, \vec{x}^{C_4}, \vec{x}^W, \vec{x}^P]^T \quad (6)$$

that describes the overall motion in the meeting room with 42 features.

5. DYNAMIC BAYESIAN NETWORK MODEL

A Bayesian network (BN) is a graphical model that describes statistical dependencies between a set of variables. The variables are marked as nodes and the dependencies between them with edges. Dynamic Bayesian networks (DBNs) are a generalization of BNs, they are used to describe time series: One BN represents one time slice. Additionally edges describe the dependencies of variables between subsequent time slices. For a given observation O with length T the DBN is "unrolled": The time slices are repeated T-times and connect through their inter-edges. Different learning and inference methods are known for DBNs. Well known models, like Hidden Markov Models (HMMs) or linear dynamical systems (LDS) [7] can be described within the DBN-framework.

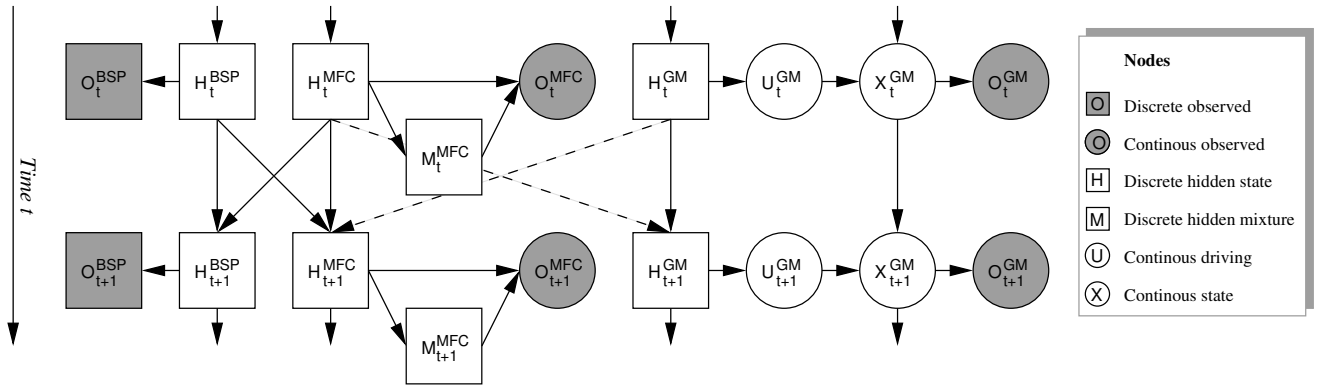

Fig. 1. Multi-stream mixed-state dynamic Bayesian network model

5.1. A novel multi-stream mixed-state DBN

Mixed-state DBNs are an HMM coupled with a LDS, they have been introduced and applied to recognizing human gestures in [8]. Here, this approach is extended to a novel multi-stream DBN for meeting event recognition.

The model is shown in Fig. 1. Each row represents one time slice. Arrows pointing down represent the dependencies between subsequent time slices. Arrows pointing to the right represent dependencies between hidden and observed variables within one time slice. Hidden variables are white, observed variables shadowed. Squares mark discrete probability distributions, circles denote continuous Gaussian distributions.

Each of the three observed features microphone array (**B**SP), lapel microphone (**M**FCC) and the visual global motion stream (**G**M) is modeled in a separate stream. The streams correspond to a multi-stream HMM, where each stream has a separate representation for the features. However, the visual stream is connected to a LDS, resulting in a mixed-state DBN. The LDS is implemented as four Gaussian nodes, in Fig. 1 represented by the two columns on the right (X_t^{GM}, O_t^{GM}). This LDS is a Kalman filter, using information from all streams as driving input, to smooth the visual stream. With this filtering, movements are predicted based on the previous time-slice and on the state of the multi-stream HMM at the current time. Thus occlusions can be compensated with the information from all channels.

With the DBN framework, this coupled HMM-LDS system can be described by the joint stream probability distributions. Thereby, the probability P_M of the lapel microphone stream is:

$$P_M = P(H_0^M) \prod_{t=1}^{T-1} P(H_t^M | H_{t-1}^M, H_{t-1}^B, H_{t-1}^G) \prod_{t=0}^{T-1} \Big(P(O_t^M | M_t^M, H_t^M) P(M_t^M | H_t^M) \Big) \tag{7}$$

the probability P_B of the microphone array stream can be calculated as:

$$P_B = P(H_0^B) \prod_{t=1}^{T-1} P(H_t^B | H_{t-1}^B, H_{t-1}^M) \prod_{t=0}^{T-1} P(O_t^B | H_t^B) \tag{8}$$

and the probability P_G of the coupled HMM-LDS-structure for the global motion stream:

$$P_G = P(H_0^G) \prod_{t=1}^{T-1} P(H_t^G | H_{t-1}^G, H_{t-1}^M) \prod_{t=0}^{T-1} P(U_t^G | H_t^G) \; P(X_0^G | U_0^G) \prod_{t=1}^{T-1} P(X_t^G | X_{t-1}^G, U_t^G) \prod_{t=0}^{T-1} P(O_t^G | X_t^G) \tag{9}$$

Each meeting event can now be described by a DBN with the model parameters

$$E_j = \{H^B, H^M, M^M, H^G, U^G, X^G\}$$

Given an observation O and the model parameters E_j, the joint probability of the model is:

$$P(O, E_j) = P_B \cdot P_M \cdot P_G \tag{10}$$

The model parameters are learned for each of the eight event classes j with the EM-algorithm during the training phase. In [9] an EM-algorithm based on variational inference was introduced, that can be applied to mixed-state DBNs [8]. This algorithm can be adapted to the multi-stream mixed-state DBN.

During the classification an unknown observation O is presented to all models E_j. Then $P(O|E_j)$ is calculated for each model and O is assigned to the class with the highest likelihood:

$$\underset{E_j \in E}{\operatorname{argmax}} P(O|E_j) \tag{11}$$

Applying the Viterbi-algorithm to the model, leads to a meeting event segmentation framework. This is however not the scope of this work.

		Single modal HMMs			Multi modal	
		Audio	Array	Visual	HMM	DBN
I	Clear test data	83.10%	83.61%	67.24%	85.22%	87.83%
II	Lapel microphone disturbed	61.11%			80.87%	86.96%
III	Microphone array disturbed		75.41%		83.48%	86.96%
IV	Visual stream 1/3 occluded			40.90%	82.61%	83.48%
V	All three streams disturbed				80.00%	81.74%

Table 1. Meeting event recognition performance for clear and disturbed data.

6. EXPERIMENTAL RESULTS

The multi-stream mixed-state DBN was evaluated on the IDIAP meeting corpus (see Sec. 2) and compared to three single-stream HMMs and a baseline early fusion HMM. Each single-stream HMM was trained and evaluated with only one modality. For the early fusion HMM the frame rates of the three observation streams were adjusted and concatenated to one large stream.

The models were trained with clear data from 30 videos. For the evaluation clear and cluttered data from the remaining 30 unknown videos have been used. In the first evaluation set (I) clear data was used. In the second set (II) the lapel microphone data and in the third set (III) the microphone array was disturbed with a 10dB SNR babble. In the fourth set (IV) one-third of the visual stream was occluded with a black bar. The fifth set (V) has all three disturbances.

Table 1 shows the recognition results of all models. It can be seen, that the two audio streams have a good recognition rate for clear data. The visual stream alone does not provide very much information. However after the sensor fusion the visual stream improves the recognition rate, significantly. The new DBN reaches a recognition rate of 87.83% for clear (I), and 81.74% for completely disturbed data (V). Compared to the early fusion HMM, the multi-modal mixed-state DBN reduces the relative error by 17% (2.61% absolute), respectively 9% (1.74% absolute) for clear (I) and disturbed (V) data. Thus, the DBN model is highly robust against noise and occlusions in all channels, and outperforms all evaluated HMM approaches.

7. CONCLUSIONS AND FUTURE WORK

In this work a new multi-modal mixed-state DBN for robust meeting event recognition from clear and disturbed data has been presented. Three audio and visual modalities are fused in a multi-stream HMM. Within the graphical model this HMM is coupled to a LDS. This LDS uses information from all streams as driving input, to smooth the visual stream. The model is a mixed-state DBN that is robust against noise and occlusions in all streams.

The DBN was compared to single-stream HMMs and an early fusion HMM. The DBN shows a significantly higher recognition performance than all single-modal HMMs. Compared to an multi-modal early fusion HMM, the novel DBN has a relative error reduction of 17% for clear and 9% for disturbed data. In the future we plan to add higher semantic features, like detected persons to the model.

8. ACKNOWLEDGEMENT

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-64).

9. REFERENCES

[1] N. Morgan et al., "Meetings about meetings: research at ICSI on speech in multi-party conversations," in *Proc. IEEE ICASSP*, Hong Kong, April 2003.

[2] 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.

[3] I. McCowan, D. Gatica-Perez, S. Bengio, and G. Lathoud, "Automatic analysis of multimodal group actions in meetings," IDIAP-RR 27, IDIAP, 2003.

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

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

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

[7] Z. Ghahramani, "Learning dynamic Bayesian networks," in *Adaptive Processing of Sequences and Data Structures. Lecture Notes in Artificial Intelligence*, C.L. Giles and M. Gori, Eds., Berlin, 1998, pp. 168–197.

[8] V. Pavlovic, B. Frey, and T.S. Huang, "Time series classification using mixed-state dynamic Bayesian networks," in *Proc. IEEE CVPR*, 1999.

[9] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M.I. Jordan, Ed. 1998, pp. 105–161, MIT Press.