A MULTI-MODAL GRAPHICAL MODEL FOR ROBUST RECOGNITION OF GROUP ACTIONS IN MEETINGS FROM DISTURBED VIDEOS

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 from disturbed videos. The model uses information from the audio and the visual channel 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 visual channel. Thereby the HMM is used as driving input for the LDS. Thus the model can handle noise and occlusions in the video. 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 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%.

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 an agenda and a summarization of the meeting. Different approaches for this structuring, based on hidden Markov models (HMMs) [2] and dynamic Bayesian networks (DBNs) [3, 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. The model uses audio information to drive a linear dynamical system, that compensates disturbances in the visual channel.

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 disturbances to the recognition performance, the evaluation data was cluttered: The audio data from the lapel microphones and the microphone array was disturbed with a background-babble with 10dB SNR. To simulate a person standing (or walking) between the camera and the recorded persons, the video data was occluded with a grey bar covering one third of the image at different positions (left, middle, and right third of the picture). For another evaluation set, a grey cross, covering 5/9 of the video was added. In a final set, a 10dB SNR Gaussian noise was added to the images. Fig. 1 shows a typical video snapshot of the meeting data and added occlusions.

For this work 30 clean videos were used for the training of the models. For the evaluation, the remaining 30 unknown videos have been cluttered with one or a combination of disturbances.

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]:

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

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].

From: ICIP 2005, Genova, IEEE

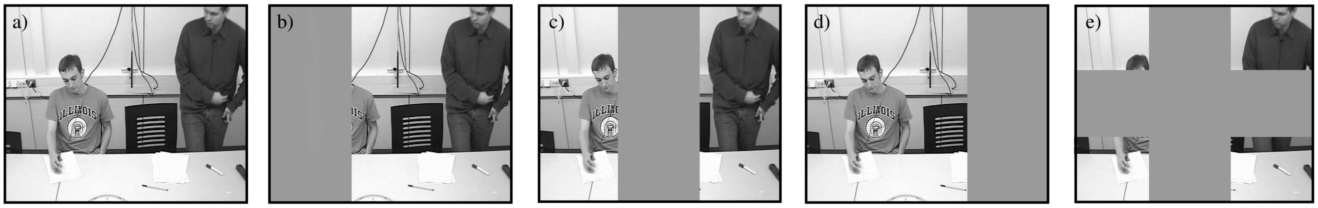

Fig. 1. Typical video snapshot of a meeting (a) and the same image with different kinds of occlusions added (b-e).

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. A visual and an audio feature stream has been formed; altogether 68 features from three modalities: microphone array, lapel microphone, and visual information have been used.

4.1. 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:

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

and

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

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.

4.2. 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 [2] from the microphone array. This results in a six-dimensional discrete vector $\vec{x}_{BSP}(t)$ containing position dependent information. The speaker- and the position-dependent vectors have been concatenated

$$\vec{x}_A(t) = [\vec{x}_S(t), \vec{x}_{BSP}(t)] \quad (7)$$

resulting in the final audio feature vector.

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) [7] or linear dynamical systems (LDS) [8] can be described within the DBN-framework.

Mixed-state DBNs are an HMM coupled with a LDS, they have been introduced and applied to recognizing human gestures in [9]. Here, this approach is extended to a novel multi-stream DBN for meeting event recognition.

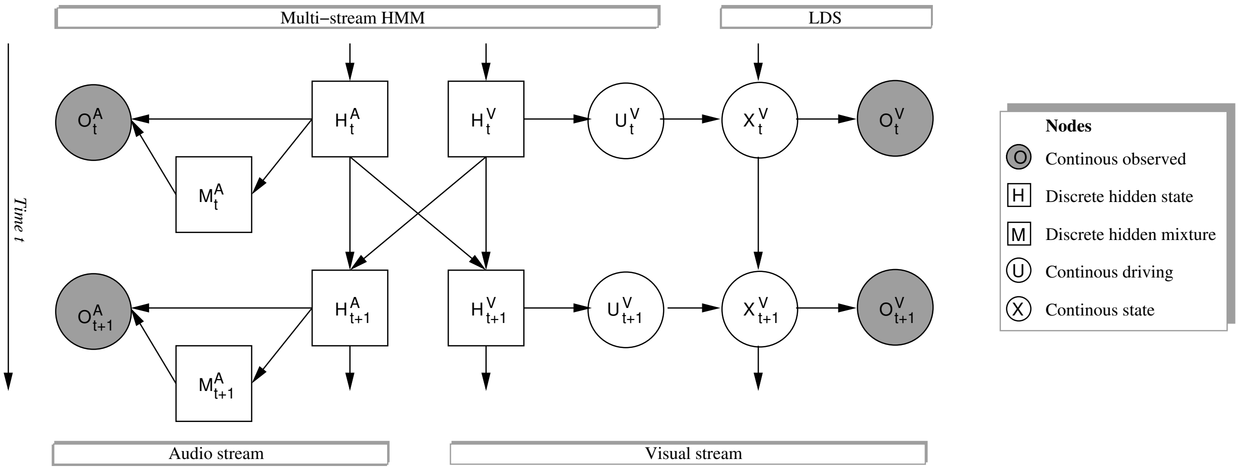

Fig. 2. Multi-stream mixed-state dynamic Bayesian network model: The HMM is driving input for the LDS.

A LDS is described by three state-space equations:

$$\vec{x}_0 = B\vec{u}_0 + \vec{v}_0 \quad (8)$$

$$\vec{x}_t = A\vec{x}_{t-1} + B\vec{u}_t + \vec{v}_t \quad (9)$$

$$\vec{o}_t = C\vec{x}_t + \vec{w}_t \quad (10)$$

where $\vec{x}_t$ ist the hidden state, $\vec{u}_t$ the driving input, and $\vec{o}_t$ the observation of the system. A,B, and C are the state transition, the input, and the observation matrices; $\vec{v}_t$ and $\vec{w}_t$ are noise terms. If the Gaussian distribution is defined as

$$\mathcal{N}(\vec{x},\vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} \exp\Big(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\Big)$$

the LDS can be described with probability distributions as well:

$$P(\vec{x}_0|\vec{u}_0) = \mathcal{N}(B\vec{u}_0,\vec{\mu},\Sigma) \quad (11)$$

$$P(\vec{x}_t|\vec{x}_{t-1},\vec{u}_t) = \mathcal{N}(\vec{x}_t - A\vec{x}_{t-1} - B\vec{u}_t,\vec{\mu},\Sigma) \quad (12)$$

$$P(\vec{o}_t|\vec{x}_t) = \mathcal{N}(\vec{o}_t - C\vec{x}_t,\vec{\mu},\Sigma) \quad (13)$$

For the new DBN model, the output of a multi-stream HMM is used as driving input $\vec{u}_t$ for this LDS. This can be described as a graphical model, as shown in in Fig. 2. 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 $\mathcal{N}(\vec{x},\vec{\mu},\Sigma)$.

The observed audio- and visual-features are modeled in separate streams. This streams correspond to a multi-stream HMM, where each stream has a separate representation for the observations. However, the visual stream is directly connected to a LDS, resulting in a mixed-state DBN. The LDS is implemented as four Gaussian nodes, in Fig. 2 represented by the two columns on the right (X_t^V, O_t^V). Thus the LDS uses information from the audio and the visual stream 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 HMM-LDS system can be described by a joint stream probability distribution. Therefore all HMM transition and mixture matrices, and the prior distributions have to be defined as discrete probability distributions [7]. The probability distributions for the LDS need to be defined according to Eq. 11 - 13. Then the probability P_A of the audio stream is:

$$P_A = P(H_0^A)\prod_{t=1}^{T-1} P(H_t^A|H_{t-1}^A,H_{t-1}^V) \prod_{t=0}^{T-1}\Big(P(O_t^A|M_t^A,H_t^A)P(M_t^A|H_t^A)\Big) \quad (14)$$

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

$$P_V = P(H_0^V)\prod_{t=1}^{T-1} P(H_t^V|H_{t-1}^V,H_{t-1}^A)\prod_{t=0}^{T-1} P(U_t^V|H_t^V) P(X_0^G|U_0^V)\prod_{t=1}^{T-1} P(X_t^V|X_{t-1}^V,U_t^V)\prod_{t=0}^{T-1} P(O_t^V|X_t^V) \quad (15)$$

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

$$E_j = \{H^A, M^A, H^V, U^V, X^V\}$$

Given an observation O and the model parameters E_j, the joint probability of the model is: $P(O,E_j) = P_A \cdot P_V$

The model parameters are learned for each of the eight event classes j with the EM-algorithm during the training phase. In [10] an EM-algorithm based on variational inference was introduced, that can be applied to mixed-state DBNs [9]. 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:

$$\operatorname*{argmax}_{E_j \in E} P(O|E_j) \quad (16)$$

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		Multi-modal	
Evaluation set	Audio	Visual	HMM	DBN
a) Clear test data	83.1%	67.2%	85.2%	88.7%
b) Left occluded		40.9%	82.6%	87.8%
c) Middle occluded		44.3%	83.5%	76.5%
d) Right occluded		52.2%	85.2%	86.1%
e) Cross occluded		33.0%	79.1%	81.7%
f) Gaussian noise		42.6%	84.4%	87.8%
I) Audio disturbed	61.1%		80.9%	87.0%
II) A-V disturbed			80.0%	81.7%

Table 1. Meeting event recognition performance.

6. EXPERIMENTS AND RESULTS

The multi-stream mixed-state DBN was evaluated on the IDIAP meeting corpus (see Sec. 2) and compared to an audio and a visual single-stream HMM, and to a multi-modal 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 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 test set (a), the audio and the visual channel had no disturbances. Three test sets had the visual channel partly occluded: A grey bar covering one third of the image was added at the left (b), the middle (c), and the right (d). For another set (e), a grey cross was used (Fig. 1). In a final visual disturbed set (f), Gaussian noise with 10dB SNR was added to the visual channel. For sets (b-f) the audio channel wasn't disturbed. For comparison an audio disturbed evaluation set (I) was included: a background-babble with 10dB SNR was added to the audio channel. Finally, a set (II) with both the audio channel and a third of the visual channel occluded was evaluated.

Table 1 shows the recognition results for all models. The audio stream has a good recognition rate (83.1%) for clear data (a), while the visual stream alone does provide less information (67.2%). However after the senor fusion the visual stream improves the recognition rate by 4.6% to 88.7% for the DBN model compared to the audio-stream HMM. This effect is even stronger when the audio channel has background babble (I). Then the recognition rate for the audio HMM drops by 22%, while the DBN model only drops by 1.7%. Thus, in a multi-modal system, the visual channel provides significant information about the meeting status. When the visual channel is partly occluded (b-e) or disturbed by Gaussian noise (f), the recognition rate of the visual HMM drops in average by 24.6% compared to the clear channel (a). In comparison, the rate for the DBN model drops by only 4.7% in average.

The multi modal DBN outperforms the multi modal HMM in all except one case (c). For clean data the DBN improves the recognition performance from 85.2% (HMM) to 88.7% (DBN), this is a relative error reduction of 23.6%. For the disturbed audio data (I), the DBN reduces the relative error by 31.9% (absolute 6.1%) compared to the HMM; for an occluded visual channel the error reduction can be up to 29.9% (absolute 5.2%, b). In average the DBN improves the recognition rate by absolute 2.1% compared to the early fusion HMM. As expected from the theory, these results show, that the coupled LDS-HMM structure compensates disturbances much better, then the early fusion HMM.

7. CONCLUSIONS

In this work a new multi-modal mixed-state DBN for robust meeting event recognition from clear and disturbed data has been presented. The audio and the visual channel are fused in a multi-stream HMM. Within the graphical model this HMM is coupled to a LDS. The LDS uses both streams as driving input, to smooth the visual stream. Thus the model can compensate visual occlusions.

The DBN was compared to single-stream HMMs and an early fusion HMM. The DBN shows a significantly higher recognition performance than the single-modal HMMs. Compared to an multi-modal HMM, the novel DBN has a relative error reduction of 23.6% for clear and up to 29.9% for visual disturbed data. In average the DBN improves the recognition rate by absolut 2.1%.

The proposed model is not limited to the recognition of meeting events, but could be used for all applications where different channels could be used to improve the visual channel.

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-71).

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] 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.

[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] L.R. Rabiner, “A tutorial on Hidden Markov Models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, February 1989.

[8] 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.

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

[10] 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.