ROBUST MULTI-MODAL GROUP ACTION RECOGNITION IN MEETINGS FROM DISTURBED VIDEOS WITH THE ASYNCHRONOUS HIDDEN MARKOV MODEL

Marc Al-Hames, Claus Lenz, Stephan Reiter, Joachim Schenk, Frank Wallhoff, and Gerhard Rigoll

Technische Universität München
Institute for Human-Machine Communication
Arcisstrasse 21, 80333 München, Germany
{alh, len, res, joa, rigoll}@mmk.ei.tum.de

ABSTRACT

The Asynchronous Hidden Markov Model (AHMM) models the joint likelihood of two observation sequences, even if the streams are not synchronised. We explain this concept and how the model is trained by the EM algorithm. We then show how the AHMM can be applied to the analysis of group action events in meetings from both clear and disturbed data. The AHMM outperforms an early fusion HMM by 5.7% recognition rate (a rel. error reduction of 38.5%) for clear data. For occluded data, the improvement is in average 6.5% recognition rate (rel. error red. 40%). Thus asynchronity is a dominant factor in meeting analysis, even if the data is disturbed. The AHMM exploits this and is therefore much more robust against disturbances.

***Index Terms*—** Meetings, Video signal processing, Robustness, Hidden Markov models, Multimedia communication

1. INTRODUCTION

In a recent study [1] the participants were asked to select emotion terms that they thought would be frequently perceived in a meeting: Two third of the participants named "boring", nearly one third mentioned "annoyed" as a frequently perceived emotion. On the other hand meetings, lectures and conferences can consume large parts of our working days and are mandatory for the information flow in companies and other organisational structures. The *Intel Corporation* – for example – schedules around 3 million meeting hours and another 5.7 million hours of audio bridge conferences each year [2]. Even more interestingly: Intel spends 56 thousand hours each year only on teaching their employees how to hold an effective meeting.

Thus there is a huge discrepancy between the importance of meetings in organisational structures on the one hand and the participants perception about these meetings on the other hand. Projects like the ICSI meeting project [3], Computers in the Human Interaction Loop [4], or Augmented Multi-party Interaction [5] therefore investigate how computers can be used to make meetings and lectures more effective, and how to automatically analyse them.

A first step for the automatic analysis of the meetings is a segmentation into meeting group action events like discussion or presentation [6]. This structuring can then be used to produce an agenda and a summarisation of the meeting. Different automatic methods for this structuring have been introduced [6, 7, 8, 9] and successfully applied to the analysis of recorded meeting data sets.

However, in real meetings the data can be disturbed in various ways: events like slamming of a door or background babble can mask the audio channel. The visual channel can be (partly) masked by persons standing or walking in front a camera, or a laptop computer can be placed in front of the meeting participants. All these realistic conditions influence the behaviour and therefore generally decrease the performance of the proposed meeting analysis methods. In [10] a graphical model, based on a multi-modal mixed-state dynamic Bayesian network (DBN), was proposed to handle occlusions in meeting data. The proposed model was successfully applied to both clear and occluded meeting data and it was shown that the recognition performance for the disturbed data only slightly decreased. However the mixed-state DBN is computational very complex and therefore computational infeasible in (near) real-time.

In this work we therefore propose to apply the Asynchronous Hidden Markov Model (AHMM) to the analysis of disturbed meeting data. The AHMM [11, 12] can model the joint likelihood of two observation streams, even if they are not synchronised. This is the main advantage of the AHMM compared to other multi-modal Markov models, like coupled, multi-stream, or early fusion HMMs. Previously the AHMM has therefore been successfully applied to audio-visual speech recognition [11, 12], person identification [13], the fusion of speech and gestures [14], and – in a two-layer version – for meeting analysis [15]. We will show how the AHMM can be learned from data and then used for the classification of meeting group action events. In an experimental section we will evaluate the model on real meeting data.

2. MEETING ROOM AND DATA SET

The data for this work was recorded in the IDIAP smart meeting room [16], which is equipped with a table, a whiteboard, and a projector with screen. The corpus consists of 60 videos with a length of approximately 5 minutes. Each meeting has 4 participants and is recorded with 3 cameras. All participants have a lapel and a headset-microphone attached and a microphone array is placed on the table.

To investigate the influence of disturbances to the recognition performance, the evaluation data was cluttered: The audio data was disturbed with a background-babble with 10 dB 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). For another evaluation set, a grey cross, covering 5/9 of the video was added. In a final set, a 10 dB SNR Gaussian noise was added to the video. Fig. 2 shows a snapshot of a meeting and the 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.

This work is supported by the European IST Programme Project FP6-0033812.

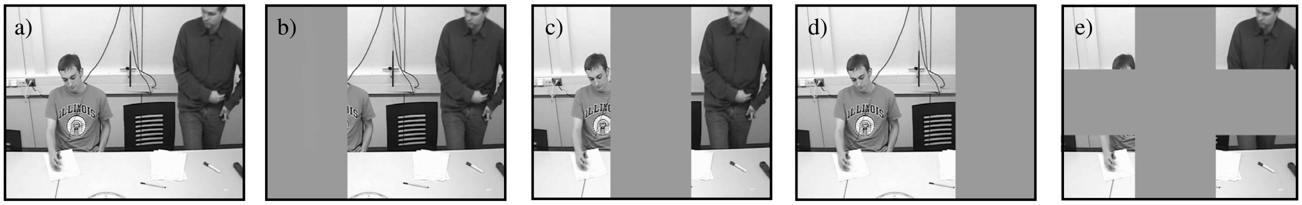

Fig. 1. Video snapshots of a meeting in the smart meeting room (a) and the same image with different kind of occlusions added (b-e)

3. GROUP ACTION MEETING EVENTS

For a first structuring of the meeting the eight different group actions $E = \{E_{\text{D}}, E_{\text{M},1}, E_{\text{M},2}, E_{\text{M},3}, E_{\text{M},4}, E_{\text{N}}, E_{\text{P}}, E_{\text{W}}\}$ are widely used [6, 7, 8, 9, 10]. The events E_j are

E_{D}: Two or more persons are talking with each other.
$E_{\text{M,Id}}$: 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 then be modelled 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 [6], e. g. in a meeting browser [17].

4. FEATURES

Visual features: In the meeting room the persons are usually at one of six locations: one of four chairs, the whiteboard, or at a presentation position: $L = \{C_1, C_2, C_3, C_4, W, P\}$. For each location L 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 [18] are derived from this image sequence: The centre of motion is calculated for the x- and y-direction:

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

and

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

The changes in motion express the dynamics of movements:

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

and

$$\Delta m_y^L(t) = m_y^L(t) - m_y^L(t-1) \tag{2}$$

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

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

and

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

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

$$i^L(t) = \frac{\sum_{(x,y)} |I_d^L(x, y, t)|}{\sum_{(x,y)} 1} \tag{4}$$

These seven features are concatenated for each frame in the location dependent vector $\vec{x}^L(t) = [m_x^L, m_y^L, \Delta m_x^L, \Delta m_y^L, \sigma_x^L, \sigma_y^L, i^L]^T$. With this motion vector the video stream is reduced to a seven dimensional vector, but it preserves the major characteristics of the observed motion. Concatenating the motion vectors from each of the six positions $\vec{x}^L(t)$ leads to the final visual feature vector $\vec{x}_{\text{V}}(t)$ that describes the overall motion in the room with 42 features.

Audio features: For each speaker four Mel frequency cepstral coefficients (MFCCs) and the energy were extracted from the lapel-microphones. This results in a 20-dimensional vector $\vec{x}_{\text{S}}(t)$ with speaker-dependent information. A binary speech and silence segmentation (BSP) for each of the six locations L in the room was extracted with the SRP-PHAT measure [6] from the microphone array, resulting in a six-dimensional vector $\vec{x}_{\text{BSP}}(t)$ containing position dependent information. The speaker- and the position-dependent vectors have been concatenated $\vec{x}_{\text{A}}(t) = [\vec{x}_{\text{S}}(t), \vec{x}_{\text{BSP}}(t)]$ resulting in the final audio feature vector. The feature frequency of the audio signal was four times higher than the video feature frequency.

5. THE ASYNCHRONOUS HMM

The AHMM is used to model the joint likelihood $p(\vec{x}, \vec{y})$ of two observation sequences $\vec{x}$ with length T, and $\vec{y}$ with length S. Without loss of generality it is assumed that $S \leq T$ (if $T > S$ a simple extension is necessary). The joint likelihood can of course not be calculated directly, as it is intractable.Therefore two hidden variables are introduced: the first variable $q_t = 1 \ldots N$ is synchronised with the stream $\vec{x}$ and identical to the state in standard HMMs. The total number of states in the model is denoted as N. It is assumed, that a state always emits a symbol from the stream $\vec{x}$ at each time step t. Furthermore each state $q_t = i$ emits with the probability $\epsilon(i, t)$ at the same time a second symbol from the stream $\vec{y}$. The hidden variable $\tau_t = 0 \ldots S$ models the alignment between $\vec{x}$ and $\vec{y}$. Whenever a state emits a symbol from stream $\vec{y}$, the alignment variable τ_t is incremented, until all symbols from $\vec{y}$ have been emitted and $\tau_t = S$.

This can be represented in a three dimensional trellis, as shown in Fig. 2. The two axis time t and state q_t are identical to those from HMMs. The axes τ_t represents the alignment between the streams.

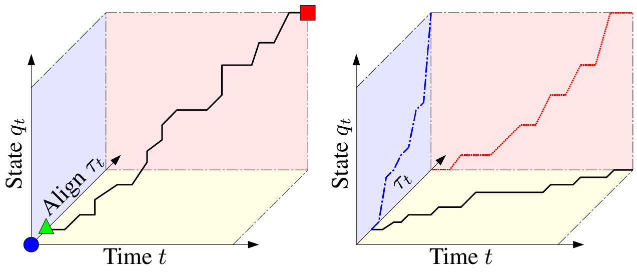

Fig. 2. 3-dimensional AHMM trellis (left) and projection (right).

$$Q(\lambda,\lambda') = \Big\langle \log p(\vec{x},\vec{y},\vec{q},\vec{\tau}|\lambda)\Big\rangle_{\vec{x},\vec{y},\lambda'} = \sum_{\vec{q}\in Q}\sum_{\vec{\tau}\in\Delta}\log(\pi_{q_1})\cdot\frac{p(\vec{x},\vec{y},\vec{q},\vec{\tau}|\lambda')}{p(\vec{x},\vec{y}|\lambda')} + \sum_{\vec{q}\in Q}\sum_{\vec{\tau}\in\Delta}\Big[\sum_{t=2}^{T}\log(a_{q_{t-1},q_t})\Big]\cdot\frac{p(\vec{x},\vec{y},\vec{q},\vec{\tau}|\lambda')}{p(\vec{x},\vec{y}|\lambda')}$$
$$+\sum_{\vec{q}\in Q}\sum_{\vec{\tau}\in\Delta}\Big[\sum_{t=1}^{T}\log\Big(p(x_t|q_t)\cdot(1-\epsilon_{q_t})\cdot\delta(\tau_t-\tau_{t-1})+p(x_t,y_{\tau_t}|q_t)\cdot\epsilon_{q_t}\cdot\delta(1-\tau_t+\tau_{t-1})\Big)\Big]\cdot\frac{p(\vec{x},\vec{y},\vec{q},\vec{\tau}|\lambda')}{p(\vec{x},\vec{y}|\lambda')} \tag{11}$$

In each time step t one symbol from $\vec{x}$ is emitted. This corresponds to move one step right in the trellis. If a state also emits a symbol from $\vec{y}$, the movement is one step right and one step in the τ_t-direction. The model can of course in each step jump to any of the states from the state axes (for an ergodic model). The projections to the three planes are shown in Fig. 2 (right). The most interesting one is the (t-τ_t)-plane (yellow): The projection of the path to this dimension represents the alignment between the $\vec{x}$- and the $\vec{y}$-stream. For comparison, an early fusion HMM always emits both a symbol from the $\vec{x}$- and the $\vec{y}$-stream, thus the alignment would be a straight line from the origin to the right-upper corner of the alignment plane.

5.1. Parameters

The AHMM is parameterised with five distributions λ:

- The initial state distribution: $\pi_i = p(q_1 = i)$
- The state transition distribution: $A_{ji} = p(q_{t+1} = i|q_t = j)$
- The probability of emitting two symbols in a state: $\epsilon_i = p(\tau_t = s|\tau_{t-1} = s-1, q_t = i)$
- The emission distributions for a single symbol $p(\vec{x}_t|q_t = i)$ and for a pair of symbols $p(\vec{x}_t,\vec{y}_t|q_t = i)$.

As with standard HMMs, the emission distributions can be modelled discretely or continuously (e. g. with a mixture of Gaussians). Furthermore the distribution for a pair of symbols $p(\vec{x}_t,\vec{y}_t|q_t = i)$ can be modelled in various ways: jointly, independently, or conditional on each other. This allows a flexibility, that is not trivially possible in other multi-modal Markov models (e. g. early fusion HMMs, where the output is always modelled jointly).

5.2. Likelihood computation

To compute the joint likelihood $p(\vec{x},\vec{y}|\lambda)$ of two streams, a forward procedure has been developed in [11] and slightly extended in [14]. The forward path variable is *defined* as:

$$\alpha(i,s,t) = p(q_t = i, \tau_t = s, \vec{x}_t, \vec{y}_s) \tag{5}$$

The model can start with either emitting one or two symbols in the first step. The *initialisation* step therefore is:

$$\alpha(i,0,1) = [1-\epsilon_i]\cdot\pi_i\cdot p(\vec{x}_1|q_t = i) \tag{6}$$

$$\alpha(i,1,1) = \epsilon_i\cdot\pi_i\cdot p(\vec{x}_1,\vec{y}_1|q_t = i) \tag{7}$$

for all $1 \le i \le N$. In Fig. 2 the initialisation is plotted as a blue circle, resp. green triangle for a model that can only start in the first state. As long as none of the symbols from $\vec{y}$ have been emitted ($s = 0$), the *induction* step is:

$$\alpha(i,0,t+1) = \tag{8}$$

$$[1-\epsilon_i]\cdot p(\vec{x}_{t+1}|q_{t+1} = i)\cdot\sum_{j=1}^{N}p(q_{t+1} = i|q_t = j)\,\alpha(j,0,t)$$

for all $1 \le i \le N$ and $1 \le t \le T-S$. If a symbol from $\vec{y}$ has already been emitted ($s > 0$), the *induction* step becomes:

$$\alpha(i,s+1,t+1) = \tag{9}$$

$$\epsilon_i\cdot p(\vec{x}_{t+1},\vec{y}_{s+1}|q_{t+1} = i)\cdot\sum_{j=1}^{N}p(q_{t+1} = i|q_t = j)\,\alpha(j,s,t)$$

$$+[1-\epsilon_i]\cdot p(\vec{x}_{t+1}|q_{t+1} = i)\cdot\sum_{j=1}^{N}p(q_{t+1} = i|q_t = j)\,\alpha(j,s+1,t)$$

for all $1 \le i \le N$, $1 \le t \le T$, and $\max\{0; t-(T-S)\} \le s \le \min\{S; t\}$. Finally the *termination* with the likelihood of the observation is:

$$p(\vec{x},\vec{y}|\lambda) = \sum_{j=1}^{N}\alpha(j,S,T) \tag{10}$$

In Fig. 2 the termination point is plotted as a red square for a model that has to end in the last state N. This procedure calculates the likelihood of an observation in $\mathcal{O}\big(N^2[TS - S^2 + T]\big)$.

Replacing the summations in Eq. (8 – 10) with maximisations leads to a *Viterbi-algorithm*. Then the best state-sequence and the best alignment between the two streams can be derived. In Fig. 2 (right) the alignment path is shown through the yellow (t-τ_t)-plane.

5.3. EM Training

To learn the parameters λ of an AHMM, an EM training procedure can be derived [11]. A backward variable is defined as $\beta(i,s,t) = p(x_{t+1}, y_{s+1}|q_t = i, \tau_t = s)$ and can be calculated analogous to the forward path. Furthermore for the learning of the output distributions, we need two auxiliary forward path variables, $\alpha^0(i,s,t)$ and $\alpha^1(i,s,t)$. They represent those parts of α where a state emits only one, resp. a pair of symbols. They can easily be calculated by only considering the single symbol-emitting, resp. pair of symbol-emitting parts of Eq. (6 – 10). As a matter of fact: in a practical implementation one would first calculate both α^0 and α^1 and then sum the two components to derive α.

We can then derive an EM Q-function, as shown in Eq. (11), on top of this page. In this function all parameters are separated. Introducing Lagrangian multipliers and derivation of the parameters leads to the update equations for the learning of the model parameters:

$$\pi_i = \frac{\sum_{s=0}^{1}p(\vec{x},\vec{y},q_1 = i|\tau_1 = s,\lambda')}{p(\vec{x},\vec{y},\tau_1 = s|\lambda')} \tag{12}$$

$$a_{ij} = \frac{\sum_{t=1}^{2}\sum_{s=0}^{t}p(\vec{x},\vec{y},q_t = j|q_{t-1} = i,\tau_t = s,\lambda')}{\sum_{t=2}^{T}\sum_{s=0}^{t}p(\vec{x},\vec{y},q_{t-1} = i|\tau_t = s,\lambda')} \tag{13}$$

$$\epsilon_i = \frac{\sum_{t=1}^{T}\sum_{s=0}^{T}p(\vec{x},\vec{y},\tau_t = s|\tau_{t-1} = s-1,q_t = i,\lambda')}{\sum_{t=1}^{T}\sum_{s=0}^{t}p(\vec{x},\tau_t = s|q_t = i,\lambda')} \tag{14}$$

The re-estimation equations for the output probabilities $p(x_t|q_t)$ and $p(x_t,y_s|q_t)$ depend on the output probability modelling and are not shown here. However their derivation is identical to the output distributions in standard HMMs.

Test Set	*Single-Modal*		*Multi-Modal*		
	Audio	Visual	HMM	DBN	AHMM
a) Clear data	83.1	67.2	85.2	88.7	90.9
b) Left occ.		40.9	82.6	87.8	89.9
c) Middle occ.		44.3	83.5	76.5	87.9
d) Right occ.		52.2	85.2	86.1	91.9
e) Cross occ.		33.0	79.1	81.7	88.9
f) Gauss. noise		42.6	84.4	87.8	89.9
I) Audio noise	61.1		80.9	87.0	80.8

Table 1. Recognition rates in percent (%) for the different models.

6. EXPERIMENTS

The AHMM was evaluated on the IDIAP meeting corpus (see Sec. 2) and compared to single-modal audio and visual HMMs, an early fusion HMM, and the DBN proposed in [10]. Each single-stream HMM was trained and evaluated with only one modality. For the early fusion HMM the frame rates of the two streams were adjusted and concatenated. For the DBN and the AHMM this is not necessary, because each stream is handled separately. The models were trained with clear data from 30 videos and tested with clear and cluttered data from the remaining 30 unknown videos. In test set (a), the audio and visual channel had no disturbances. Three 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 set (e), a grey cross was used (Fig. 2). In set (f), Gaussian noise with 10 dB SNR was added. For sets (b - f) the audio was not disturbed. For comparison an audio disturbed set (I) was included: a background-babble with 10 dB SNR was added to the audio channel.

Tab. 1 shows the recognition rates (RR) for all models. The audio stream has a good RR (83.1%) for clear data (a), while the visual stream alone provides less information (67.2%). All tested multimodal systems outperform the single-modal HMMs. The AHMM reaches the best RR of 90.9% and therefore outperforms the early fusion HMM by 5.7% absolute RRs – a relative error reduction of 38.5%. For clear data the AHMM also outperforms the DBN (19.5% rel. error reduction). Asynchronity in the data seems to be a dominant factor, which is best exploited by the AHMM.

This behaviour is increased if visual occlusions are added. Now the visual HMM drops significantly in the RR (33.0% - 44.3% depending on the occlusion). The early fusion HMM and the DBN drop slightly in their RRs. The AHMM however remains nearly unaffected from the occlusions. For one occlusion (d) the RR is even increased compared to the clear data. We have not fully investigated this effect, but the same tendency can be found for both the HMM and the DBN: the rate does not drop much for occlusion (d). Thus some misleading motion might be covered in this set. The same effect might lead to the increase in RR for the AHMM between set (c) and (e). In average for the disturbed sets (b-f) the AHMM outperforms the HMM by 6.5% and the DBN by 5.7% RR (a rel. error red. of 40%, resp. 32.8%). Again this shows that the alignment between visual and acoustic information and the asynchronity is a dominant factor, even in disturbed data. By exploiting this asynchronity with the AHMM the system gets much more robust against occlusions.

In a final evaluation we used the audio disturbed data. Here, the AHMM reaches a RR comparable to the standard HMM, but worse than the DBN. The DBN uses a Kalman filter structure to heavily improve the visual channel, thus it uses much more information from the visual channel than the AHMM and the HMM have available. Thus the asynchronity becomes less dominant for audio noise.

7. CONCLUSIONS

In this work we proposed to apply the asynchronous HMM for the recognition of group actions in meetings from disturbed data. The AHMM exploits the audio-visual stream alignment and can model asynchronity between them. Experiments showed that compared to an HMM, the AHMM improves the recognition rate by 5.7% for clear and in average by 6.5% for occlusions in the visual channel – a relative error reduction of 38.5%, resp. 40%. Thus asynchronity seems to be one of the dominant factors in the multi-modal analysis of meetings, even if the channels are heavily disturbed.

However, we found that the AHMM performance strongly depends on a careful model initialisation. In the future we therefore like to investigate the influence of initial parameters on the training.

8. REFERENCES

[1] D. Heylen et al., “Determining what people feel and think when interacting with humans and machines: Notes on corpus collection and annotation,” in *Proc. RAEM*, 2006.

[2] C. House, “Visual lexicons: The quest for data-driven decision making,” *Invited Talk at MLMI, 2005*, recording available from http://groups.inf.ed.ac.uk/mlmi05/.

[3] A. Janin et al., “The ICSI meeting corpus,” in *ICASSP*, 2003.

[4] A. Waibel et al., “CHIL: Computers in the human interaction loop,” in *Proc. NIST ICASSP Meeting Recogn. Worksh.*, 2004.

[5] J. Carletta et al., “The AMI meetings corpus,” in *Proc. Symposium on Annotating and Measuring Meeting Behavior*, 2005.

[6] I. McCowan et al., “Modeling human interaction in meetings,” in *Proc. ICASSP*, 2003.

[7] S. Reiter and G. Rigoll, “Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach,” in *Proc. ICASSP*, 2005.

[8] A. Dielmann and S. Renals, “Dynamic Bayesian networks for meeting structuring,” in *Proc. ICASSP*, 2004.

[9] M. Al-Hames et al., “Multimodal integration for meeting group action segmentation and recognition,” in *P. MLMI*, 2006.

[10] M. Al-Hames and G. Rigoll, “A multi-modal graphical model for robust recognition of group actions in meetings from disturbed videos,” in *Proc. ICIP*, 2005.

[11] S. Bengio, “An asynchronous Hidden Markov Model for audio-visual speech recognition,” in *NIPS 15*, 2003.

[12] S. Bengio, “Multimodal speech processing using asynchronous Hidden Markov Models,” *Information Fusion*, vol. 5, no. 2, pp. 81–89, 2004.

[13] S. Bengio, “Multimodal authentication using asynchronous HMMs,” in *Proc. AVBPA*, 2003.

[14] M. Al-Hames and G. Rigoll, “Reduced complexity and scaling for asynchronous HMMs in a bimodal input fusion application,” in *Proc. ICASSP*, 2006.

[15] D. Zhang et al., “Modeling individual and group actions in meetings: a two-layer HMM framework,” in *Prc. CVPR*, 2004.

[16] D. Moore, “The IDIAP smart meeting room,” Research Report 07, IDIAP, 2002.

[17] P. Wellner, M. Flynn, and M. Guillemot, “Browsing recorded meetings with Ferret,” in *Proc. MLMI*, 2004.

[18] F. Wallhoff, M. Zobl, and G. Rigoll, “Action segmentation and recognition in meeting room scenarios,” in *Proc. ICIP*, 2004.