MULTI-MODAL ACTIVITY AND DOMINANCE DETECTION IN SMART MEETING ROOMS

Benedikt Hörnler and Gerhard Rigoll

Technische Universität München
Institute for Human-Machine-Communication
80290 Munich, Germany
{hoernler,rigoll}@mmk.ei.tum.de

ABSTRACT

In this paper a new approach for activity and dominance modeling in meetings is presented. For this purpose low level acoustic and visual features are extracted from audio and video capture devices. Hidden Markov Models (HMM) are used for the segmentation and classification of activity levels for each participant. Additionally, more semantic features are applied in a two-layer HMM approach.

The experiments show that the acoustic feature is the most important one. The early fusion of acoustic and global-motion features achieves nearly as good results as the acoustic feature alone. All the other early fusion approaches are outperformed by the acoustic feature. More over, the two-layer model could not achieve the results of the acoustic features.

Index Terms— Machine Learning, Human-Machine Interaction, Activity Detection, Meeting Analysis, Multi-modal Low Level Features

1. INTRODUCTION

In every face-to-face meeting – even if the participants do not know each other – an order of dominance is established after a short period of time. However, not only a dominance level will be found in the meeting, also the activity of the different participants is observable. These social signals are connected to each other, for example if the participant is talking more then the others, this one will be recognized as more active and also as more dominant than the other participants. Still it is possible that a person who is not talking at all has a high level of dominance, because he is shaking the head for a short moment in the ongoing discussion and so his level of activity will be low. Not only an order of dominance also a hierarchical ranking is observable in a face-to-face meeting. This ranking is highly correlated with the role each participant has in the meeting, for example the project manager will normally have the highest rank. Furthermore, a relation between the dominance and the hierarchical rank is given.

Previous work on dominance detection mostly use high level features as speech transcriptions [1, 2]. The main problem with these approaches is the high latency and the high real-time factor. In this work, a system is described which can detect the dominance/activity of the participants in meetings from low level features. This system will use video- and audio-data from cameras and microphones, which are available for remote participants as well as for people seated in a smart meeting room. Thus the system can be used in meetings as well as video conferencing.

In this work, low level acoustic and visual features are extracted from the audio and video streams which are captured in a meeting room. These features and the combinations of them are used for an early fusion using Hidden Markov Models. A second approach with more semantic features which contains information about the current status of the group and the people is also evaluated. In this approches, a two-layer HHM system is used and at the first layer these semantic informations are directly derived from the low level features.

The next section gives an overview of the data set and the annotation, which are needed for the training of the models. Section 3 describes the used acoustic, visual and semantic features. The pattern recognition models used in this work are presented in Section 4. In Section 5 the results from the experiments are shown and finally the conclusion is drawn in Section 6.

2. DATA SET

The AMI corpus [3], which is publicly available, is used for this work. A subset of 36 meetings with a duration of five minutes each, was created and four participants are always located somewhere in the IDIAP smart meeting room [4] during these meetings. The meeting room is equipped with seven cameras, 22 microphones, a projector screen and a whiteboard. This work uses only four close talking microphones and does not take into account the installed microphone arrays

This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

(a) Camera Left (b) Camera Center (c) Camera Closeup

Fig. 1. Sample shots of three cameras from the IDIAP smart meeting room: left camera (L), centre view (O) of the room and a closeup (C_1) of participant one.

for the far field recordings. For the video capturing, seven cameras are installed: one closeup camera for each participant ($C_1 - C_4$). An overview camera (O) that records the table, the whiteboard and the projector screen. Two additional cameras are located at the left (L) and right (R) wall and capturing two participants and the half table in front of them. Three example shots of these cameras are shown in figure 1.

2.1. Annotation

As novel pattern recognition techniques were applied to the activity detection task, annotation is needed for the whole set of data. Therefore, all 36 five minutes meetings have been annotated by using the labels: absent, not active, little active, active and most active. An additional label called decision making is added at special points of the meeting when one participant made a decision. This label contains compared to the others more semantic information of the ongoing meeting. For each meeting, a sequence of short segments has been created and annotators selected the according labels. The inter-annotator-agreement of two annotators is 60.9%, which is a moderate agreement and therefore the annotation seems to be quiet robust and consistent.

3. FEATURES

In this work, three different modalities of features are used: acoustic, visual and semantic. The first two modalities are low level features and are derived directly from the audio- and video streams. The semantic features contain more related information of the occurrences in the ongoing meeting. In the following paragraphs the features are described.

3.1. Acoustic Features

Mel frequency cepstral coefficients (MFCC) [5] are widely used in the automatic speech recognition domain. The feature can be calculated in real time with only a latency of one window. Therefore, it seems to be a good idea to use MFCCs as an acoustic feature in the activity detection. For each close talking microphone, which a participant was carrying, the energy plus twelve cepstral coefficients and the first and second derivations are extracted.

3.2. Visual Features

Global motions (GM) have been successfully applied to various meeting tasks [6, 7] and can be calculated in real-time. First the meeting room is split into six locations L. Each of the four closeup cameras represents one location. From the centre view camera, we extract the projection board and the whiteboard location. Then, a difference image sequence $I_d^L(x,y)$ of two subsequent frames is calculated for each location. The seven global motion features are derived from the image sequence, again for each location. The centre 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 centre 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:

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

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

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

Concatenating the motion vectors from each of the six positions leads to the final motion vector.

The second visual features used are skinblobs which are derived from each of the cameras. In [8] various approaches of face detection are deeply investigated and one of these is a skin color look-up-table. To the regions extracted by the approach, a dilation filter is applied and then by taking into account the proportions, a bounding box for head and hand blobs are found. The positions, the size and the movement of these boxes from each camera are concatenated to the final vector.

3.3. Semantic Features

Not only acoustic and visual low level features are applied to the detection task, but also features that contain more semantic information are used. These features are interesteing because of the close relation between what a person or the group is doing and the level of activity. The features, which have been applied are group action, person action and person movement.

The group action has been deeply investigated in the research community over the last couple of years [9, 10, 11]. The systems are working directly on audio and video streams and achieve reliably results, but they are currently not real time capable. The meeting is segmented into a sequence of labels like monologue participant one to four, discussion, presentation, whiteboard and note taking.

Moreover, a person action detection system has been developed [6, 7]. These systems create a sequence of actions for each of the participants, thus four features for each time frame are available. The labels used, are similar to the group actions but contain some more classes: sitting down, standing up, nodding, shaking the head, writing, pointing, using a computer, giving a presentation, writing on the white-board, manipulation of an important item and idle. Idle for example is used if the person is speaking or listening to the meeting. The classes nodding or shaking should help to find points in the meeting where a decision is made or a person is active.

The last semantic feature which is currently used is the person movement. It describes what each person is currently doing in the meeting as the person action does, but only the labels off camera, sit, other, move, stand whiteboard, stand screen and take notes are available. Thus, it should improve the results for the activity detection, as it contains information about what the participants are currently doing.

4. ACTIVITY DETECTION MODELS

In this work, Hidden Markov Models (HMM) [12] are applied to the previous described pattern recognition problem. It can be used for classification and in combination with the Viterbi algorithm [13] also for segmentation of feature streams. For the training of the HMMs the EM-algorithm [14] is used. For each class k a model with the parameters $\lambda_k = (\mathbf{A}, \mathbf{B}, \vec{\pi})$ is trained. The model parameter $\mathbf{A}$ is the transition matrix, $\mathbf{B}$ models the output distribution using Gaussians mixtures and $\vec{\pi}$ denotes the initial state distribution.

Two different types of HMMs are used in the evaluation: single- and multi-stream HMMs. The main difference between these two types is the possibility to group different modalities of feature into several weighted streams D by using mutli-stream HMMs. The transition matrix ($\mathbf{A}$) and the initial state distribution ($\vec{\pi}$) are unchanged but for each stream a different output distribution ($\mathbf{B} = B_1, \dots B_D$) is defined. The observation of stream d is produced statistically independent from all other streams. The joint probability of the observation is similar to the single stream model.

4.1. Two-layer model

The approach adds semantic information to the single layer model. In this the first layer semantic features as group action, person action or person movement are classified. This is done by using the similar HMMs as for the single-layer model which have been trained with the annotated semantic information. The second layer of the model is also similar to the single-layer model, but additional semantic features are added to the input feature stream for the training as well as for the decoding. For the training the annotated semantic information is used again, but the decoding is performed differently. For the decoding the first layer is decoded and the output is added to the observation of the second layer and then this layer is decoded.

5. EXPERIMENTS

The experiments in this work consist of two separate tasks. For the first one, the boundaries of each segment are known and therefore only a classification is performed. This result is shown in the tables as recognition rate (RR) and the rate is equal to the number of correct classified segments divided by the total number of segments. The second experiment is the real task of the system, because the boundaries and the labels have to be detected automatically. Furthermore, a action error rate (AER) and the frame error rate (FER) are used. AER gives an impression of the sequence of the labels but does not take into account the right boundaries. The FER describes the number of correct detected frames divided by the total length of the meeting, thus it is an adequate measure for the real task. For all evaluations, a nine fold cross validation with person disjoint test and training sets was performed. Low rates of FER and AER and in the case of RR higher ones are better. Table 1 shows all the results of the evaluation. For the classification task, the best features are the acoustic ones with an RR of 54.9%. Only the single stream model using acousitic features and global motions performs nearly as good. All the other evaluated combinations achieve very bad rates. The same tendencies are shown for the FER and the AER. The acoustic feature again performs best with a FER of 47.7% and the single stream model is with 48.6% almost comparable.

The results lead to the conclusion that the feature level fusion does not fit the problem of activity detection in meetings. One problem with the semantic features is that these features are constant for several seconds and so they only disturb the training of the single stream models. Moreover a problem with the two-layer model is that the classification rate of the first layer is about 43% for the person actions and 40% for the person movements.

Table 1. Evaluation of different modality combinations. The model has 20 states for all single modalities and the fusion models has 15 states per class. The two layer models always have 20 states per class. As additional semantic features in combination with the acoustic feature have been used: movement (M), person actions (P) and group actions (G). AER stands for action error rate, FER means frame error rate and RR is the recognition rate.

Model	AER	FER	RR
Audio (A)	47.2	47.7	54.9
Global Motion (GM)	64.4	63.7	36.6
Skinblob (SK)	66.6	71.3	28.6
Single stream (A&GM)	48.4	48.6	51.8
Multi stream (A&GM)	63.6	55.8	49.2
Multi stream (A,GM&SK)	60.7	57.6	44.1
Two layer (M)	87.5	64.2	34.2
Two layer (P)	87.7	65.4	34.0
Two layer (G&M)	87.5	64.3	34.3
Two layer (G&P)	87.7	65.3	34.0

6. CONCLUSION

In this work we proposed a system for automatic detection of activity levels in the meeting environment. The system extracts low level features from audio and video sources and performs low level feature fusion. The task can be formulated as a pattern recognition problem and therefore different types of Hidden Markov Models can be applied for the extraction of a sequence of activity levels. The sequence is created for each participant and thus it is possible to rank them.

The experiments showed that the low level fusion of the different modalities does not lead to an improvement of the results. The best result achieves a simple Hidden Markov Model by only using the audio features. The integration of more semantic features at the feature level increases the frame error rate by more than 15%.

In the future it is planned to use other features, for example eye movement or a detector of slide changes, which should help to improve the system. The fusion of modalities normally increases the recognition rates and therefore other types of feature fusions will be investigated. Moreover the use of graphical models is planed because of the possibility of creating multi layer fusion systems. Finally, more research has to be done in the field of robust recognition of person actions and person movements.

7. REFERENCES

[1] D. Zhang et al., "Learning influence among interacting markov chains," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., pp. 1577–1584. MIT Press, 2006.

[2] R. Rienks and D. Heylen, "Automatic dominance detection in meetings using easily obtainable features," in *Proceedings of the 2nd Workshop on MLMI*, 2006.

[3] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Proceedings of the 2nd Workshop on MLMI*. 2006, pp. 28–39, Springer-Verlag.

[4] D. Moore, "The IDIAP smart meeting room," Technical Report 07, IDIAP, 2002.

[5] Z. Fang, Z. Guoliang, and S. Zhanjiang, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[6] M. Zobl, F. Wallhoff, and G. Rigoll, "Action recognition in meeting scenarios using global motion features," in *Proceedings of the 4th IEEE International Workshop on PETS-ICVS*, J. Ferryman, Ed., 2003, pp. 32–36.

[7] F. Wallhoff, M. Zobl, and G. Rigoll, "Action segmentation and recognition in meeting room scenarios," in *Proceedings of the 11th ICIP*, 2004.

[8] M.-H. Yang, D.J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transasctions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[9] D. Zhang et al., "Modeling individual and group actions in meetings: a two-layer HMM framework," in *Proceedings of the Second IEEE Workshop on Event Mining: Detection and Recognition of Events in Video, in Association with CVPR*, 2004.

[10] M. Al-Hames et al., "Multimodal integration for meeting group action segmentation and recognition," in *Proceedings of the 2nd Joint Workshop on MLMI)*, 2006.

[11] S. Reiter, B. Schuller, and G. Rigoll, "Hidden conditional random fields for meeting segmentation," in *Proceedings of the 8th ICME*, 2007.

[12] 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, 1989.

[13] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260 – 269, 1977.

[14] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.