BOOSTING MULTI-MODAL CAMERA SELECTION WITH SEMANTIC FEATURES

Benedikt Hörnler and Dejan Arsić and Bjön Schuller and Gerhard Rigoll

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

ABSTRACT

In this work semantic features are used to improve the results of the camera selection. These semantic features are group action, person action and person speaking. For this purpose low level acoustic and visual features are combined with high level semantic ones. After the feature fusion, a segmentation and classification are performed by Hidden Markov Models. The evaluation shows that an absolute improvement of 6.5% can be achieved. The frame error rate is reduced to 38.1% by using acoustic and all semantic features. The best model using only low level features achieves a frame error rate of 44.6%, which is the best one reported on this data set.

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

1. INTRODUCTION

Today's business world is full of meetings and of travel to meetings around the world. Video conferencing [1] is a successful approach to reduce costs for companies. At the beginning of online conferencing only one video stream was exchanged between two locations. Nowadays different locations with multiple cameras are connected and a new problem arises: Which camera should be shown? Which cameras could be ignored?

Not only for online video conferences, but also for previously recorded meetings, it is an interesting topic to show a selected camera, which contains the most relevant informations from the meeting. For the playback of past meetings, a meeting browser [2] can be used. These are the main usage scenarios of the system which is described in this work.

Previous work concentrates on two different approaches for these usage scenarios. The first approach [3] uses high level features, as speech transcripts and person movements, and the camera selection process is based on rules. For the evaluation, people have been watching the created video and judged the quality. Therefore, it is impossible to compare it with others. The second approach uses low level features and two different models for the camera selection. The first one is based on thresholds [4] and the second one uses Hidden Markov Models [5].

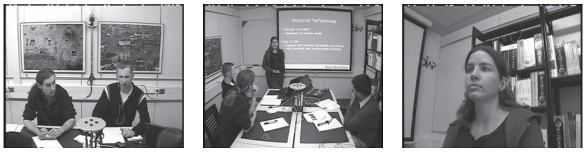
(a) Camera Left (b) Camera Center (c) Camera Closeup

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.

In this work we combine the best of both approaches to achieve better results. In the first step low level features are extracted from the audio and video sources. Additionally to these features, high level ones, such as group action, person action and person speaking, are used for the camera selection task. The second step is concatenating the features on feature level. After that, a segmentation and classification is done by Hidden Markov Models (HMM) [6]. Different combinations of features and settings for the models have been evaluated.

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 [7], which is publicly available, is used for this work. A subset of 24 meetings with a duration of five minutes each, was created and four participants are always located somewhere in the IDIAP smart meeting room [8] dur-

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.

978-1-4244-4291-1/09/$25.00 ©2009 IEEE

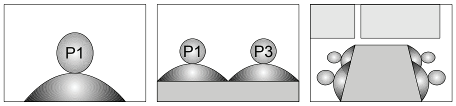

Fig. 2. Sketches of three available video modes in the IDIAP smart meeting room: closeup (C_1) of participant one, left camera (L) and centre view (O).

ing 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 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 are 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

Annotation of the whole data set is needed for two reasons: First, it is necessary for the training of the Hidden Markov Models and second the ground truth is used for the evaluation of the results. Thus 24 five minutes meetings have been annotated. The annotators have to decide which of the seven video modes represents the meeting best. In this work a video mode is defined as one of the seven cameras. This leads to seven different modes which contain four closeups views, one for each participant, a left and a right view, which records two people at one side of the table, and a centre view, which covers the whole meeting room. In figure 2 three sketches are shown of the available video modes. In the future additional video modes, as the recorded slides from the projector, are planned to be annotated.
The task of annotating video modes is very subjective as the low average of inter-annotator agreement ($\kappa = 0.3$) shows. It is highly depending on the taste of the different annotators. Therefore, only one annotator labels the whole corpus to achieve a consistent annotation. Moreover, the inter-annotator agreement of a single annotator, doing the same meeting twice, raises to $\kappa = 0.6$. Even though the shot boundaries are on a frame base and no gray array is allowed around the shot change.

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) [9] 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 [10, 11] 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

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

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

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

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

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

and

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

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

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

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

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

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 [12] 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 interesting because of the close relation between what a person or the group is doing and which camera is important. The features, which have been applied are group action, person action and person speaking.

The group action has been deeply investigated in the research community over the last couple of years [13, 14]. The systems are working directly on audio and video streams and achieve reliable 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 [10, 11]. 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 person should be shown even though he is not speaking.

The last semantic feature which is currently used is the person speaking. It is a four dimensional vector which contains binary information for each participant and each time frame. The bits are set to one if a person is speaking.

4. CAMERA SELECTION MODELS

In this work, Hidden Markov Models (HMM) [6] are applied to the previous described pattern recognition problem. It can be used for classification and in combination with the Viterbi algorithm [15] also for segmentation of feature streams. For the training of the HMMs, the EM-algorithm [16] is used. For each class k a model with the parameters $\lambda_k = (\mathbf{A}, \mathbf{B}, \vec{\pi})$ is trained. The model parameter **A** is the transition matrix, **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 (**A**) and the initial state distribution ($\vec{\pi}$) are unchanged but for each stream a different output distribution ($\mathbf{B} = B_1, \ldots B_D$) 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.

Table 1. Evaluation of different modality combinations. The number of states per model varies over the different combinations of modalities. MS indicates that a multi-stream model achieves this result. AER means action error rate, FER is the frame error rate and RR stands for recognition rate.

Model	AER	FER	RR
Audio (A)	158.7	50.1	47.6
Global Motion (GM)	177.5	64.3	34.8
Skinblob (SK)	600.3	78.6	16.8
Group Action (GA)	84.8	61.0	26.2
Person Action (PA)	72.2	62.8	28.2
Person Speaking (PS)	62.2	51.5	48.3
Audio & GA	63.1	49.2	48.3
Audio & PA	60.2	42.5	51.5
GA & PA & PS	58.3	39.6	54.8
A & GA & PA & PS (MS)	56.2	38.1	53.9
Audio & GM (MS)	60.8	44.6	52.9

5. EXPERIMENTS

For all the experiments, a six-fold cross validation with person disjoint test and training sets were performed. Three different measurements are used for the evaluation: recognition rate (RR), action error rate (AER) and frame error rate (FER). High rates of RR are good and in the case of AER and FER lower values are better.

The experiments consist of two different tasks: classification and combined segmentation and classification. For the first one, the class boundaries are given and only a classification of these segments is performed. The results of this task are measured as RR and it is equal to the number of correct classified segments divided by the total number of segments. The second experiment is the combined process of finding the right class boundaries and classify these segments correctly. This is the real task of the system and the measurements for that are the FER and AER. The FER counts all the correct detected frames and divides them by the total number. Thus, the FER takes into account the correct position of the boundaries. The AER consideres only the correct sequence of segments.

In table 1, first the results of all single modalities are presented. The low level audio features achieve a FER of 50.1%, as the best single modality. Only the person speaking features performs nearly comparable. The visual features alone are not enough for the camera selection task, because most of the time the person who is speaking is important. The high AER of the low level features means that too many shot changes have been added to the video.

The first idea was combining acoustic features, as audio or person speaking, with visual hints, as global motion or person actions. The fusion of audio and group actions improves the results slightly. The use of person actions reduces the FER about 7.6% to a rate of 42.5%. This is already better than the best low level feature result of this work (44.6%) and all evaluated fusions in [5] (47.9%). The FER can be further reduced by combining all semantic features to a rate of 39.6%. The best results achieves a multi-stream HMM by using audio features and all high level features with a FER of 38.1%. For the RR and AER the picture is very similar, only for the RR the best model uses all semantic features only.

6. CONCLUSION

In this work we presented the combination of low level and semantic features for camera selection. The system performs a feature fusion using single and multi-stream HHMs. There is an reduction of 6.5% for the FER from the best low level feature model to the best combination. The integration of semantic features, as group action, person action and person speaking, into the system is successful.

In the future we plan to detect the semantic features from low level features and combine all the systems to one stand alone camera selection system. Further work will be conducted in the field of late fusion and more complex graphical models.

7. REFERENCES

[1] S. Sabri and B. Prasada, “Video conferencing systems,” *Proceedings of the IEEE*, vol. 73, no. 4, pp. 671 – 688, 1985.

[2] P. Wellner, M. Flynn, and M. Guillemot, “Browsing recorded meetings with ferret,” in *Proceedings of the 1st Joint Workshop on MLMI*, S. Renals and S. Bengio, Eds. 2004, Springer Verlag.

[3] S. Sumec, “Multi camera automatic video editing,” in *Proceedings of the ICCVG*. 2004, pp. 935–945, Kluwer Verlag.

[4] M. Al-Hames, B. Hörnler, C. Scheuermann, and G. Rigoll, “Using audio, visual, and lexical features in a multi-modal virtual meeting director,” in *Proceedings of the 3rd Joint Workshop on MLMI*. 2006, Springer Verlag.

[5] M. Al-Hames, B. Hörnler, R. Müller, J. Schenk, and G. Rigoll, “Automatic multi-modal meeting camera selection for video-conferences and meeting browsing,” in *Proceedings of the 8th ICME*, 2007.

[6] 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.

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

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

[9] 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.

[10] 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.

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

[12] 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.

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

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

[15] 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.

[16] 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.