AUTOMATED VIDEO EDITING FOR MEETING SCENARIOS APPLYING MULTIMODAL LOW LEVEL FEATURE FUSION

Dejan Arsić , Benedikt Hörnler and Gerhard Rigoll

Institute for Human-Machine-Communication
Technische Universität München, Germany
{arsic,b,rigoll}@tum.de

1. INTRODUCTION

Most of the employees dislike business meetings because of the effort, the duration and the low efficiency. The AMIDA-project [1] attempts to increase the efficiency by the use of modern machine-learning techniques. One of the ideas of AMIDA is that a camera selection could be performed in smart-meeting rooms, which are equipped with several cameras, so that the most relevant information is shown in the output video. This video could be used for a video conference, even on small devices as a cell phone, for catching up missed parts of an currently ongoing meeting or for storing of summaries of the past meetings. The important camera is selected by using a support vector machine and easy and fast computable low level features.

2. THE AMI MEETING DATABASE

For this task a subset of the AMI-Corpus [2] with a total length of three hours, recorded at the IDIAP-Smart-Meeting-Room, was created. The subset contains 36 meetings with a length of five minutes. The IDIAP-room is equipped with seven cameras, 22 microphones, a whiteboard and a projector with a screen. Four of these cameras record closeup views of the four participants, two cameras show the left, respective the right side of the table and one camera captures the whiteboard and the projector screen. Eight microphones capture close-talking audio which is used of the video-editing task.

3. LOW LEVEL FEATURE EXTRACTION

A major issue for on line streaming of meetings is the need for real time capable feature extraction and subsequent classification. Additionally the features have to be extracted robustly from every frame. Facial feature points as defined by the MPEG7 standard [3] have been discarded, as they are not visible in all frames. Therefore we decided to focus on global motion features [4] extracted from various parts of the image based on a simple difference image: $d(x,y,t) = I(x,y,t) - I(x,y,t+1)$ First the center of motion $m = [m_x, m_y]$ can be computed both in x and y direction:

$$m_x = \frac{\sum_{x,y} x * d(x,y,t)}{\sum_{x,y} d(x,y,t)},\ m_y = \frac{\sum_{x,y} y * d(x,y,t)}{\sum_{x,y} d(x,y,t)}$$

Since the behavior is independent of the passenger's location, only changes in the direction of movement and their value is used: $\delta m_{x/y} = m_{x/y}(t) - m_{x/y}(t-1)$
To distinguish between motions of large or small parts of the body the mean absolute deviation $\sigma = [\sigma_x, \sigma_y]$ is computed with:

$$\sigma_x = \frac{\sum_{x,y} d(x,y,t)|x - m_x|}{\sum_{x,y} d(x,y,t)}\ \sigma_y = \frac{\sum_{x,y} d(x,y,t)|x - m_y|}{\sum_{x,y} d(x,y,t)}$$

Furthermore the changes within a series of variance are considered: $\delta\sigma_x = \sigma_x(t) - \sigma_x(t-1)$ and $\delta\sigma_y = \sigma_y(t) - \sigma_y(t-1)$. Additionally the so called intensity of motion

$$i = \frac{\sum_{x,y} d(x,y,t)}{\sum_{x,y} 1}$$

is taken into account, which describes the changes in the entire image.

Figure 1 illustrates the various areas features are extracted

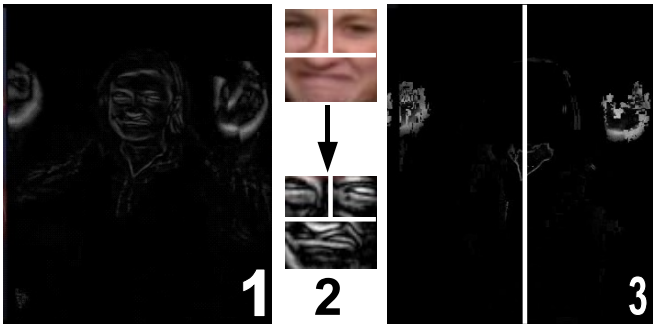

Fig. 1. Visualization of applied features

This work has partially been funded by the European Union within the FP6 AMIDA project www.amiproject.org and FP7 PROMETHEUS Project www.prometheus-FP7.eu.

from. In the first place the global motion features are computed from the entire image for each frame in a video sequence. Second we perform real time face tracking based on an initialization with a Multi Layer Perceptron proposed by Rowley [5] combined with the condensation algorithm [6]. After the face is located motion features can be extracted from both the left and right eye, which could indicate eye blinks. Furthermore a feature extraction is performed in the lower half of the face, which will model movements of the mouth. In a third stage hands are detected by applying a simple skin detection algorithm, based on a so called skin locus [7]. As the position of the face is already known it can be assumed, that remaining skin parts are representing hands and arms. Motion features are computed for separately for the right and left arm, by simply splitting the video stream in the middle. This way a total of 81 features can be extracted from every frame in each field of view. These features are extracted from the 4 close up views and from the left and right view camera. As there are 2 persons visible in the left and right cam view, the video is split into two parts by a vertical line in the middle of the image. In total $8 * 81 = 648$ features are extracted for each time frame.

The acoustical features simply describe who of the 4 participants is actively speaking in each video frame, resulting in a 4 dimensional feature vector with 1 for person speaking and 0 for being quiet.

4. RECOGNITION OF THE RELEVANT VIEW

In [8] we already presented an SVM [9] based approach for behavior analysis. Each video sequence is segmented by a window with a constant length of 25 frames without overlap. The minimum length of a shot is set to $1s$ this way. In contrast to a frame based decision this way the motion changes over time can be modeled in a finer way. Features are extracted from every segment, and the resulting vector, with a constant size of $N = 25 * num\_features$, is then classified by a Support Vector Machine.

A couple of combination of video modes and modalities has been tested for the recognition of the 7 classes. In the first place only visual features have been used training models for just the close ups *C*, the Left/Right view *LR* and all views together *All*. Secondly we added acoustic features to each group of views *+s*.

Table 1 illustrates recognition results for the various feature sets after a 4-fold evaluation. Training and test set have been chosen manually, which guaranteed, that actors would not appear in both sets. This way a person independent classification could be performed. As it can be obviously seen in all cases the acoustic features boost the performance drastically more than 20% in average. It is also remarkable that the smaller feature and easier to handle Close Up feature set performs even better as the entire set. Reasons might be the drastically smaller amount of data and the weak performance of the LR view in general, which might even disturb the classification of the entire feature set.

type:	C	C+S	LR	LR+s	All	All+s
#feat	8100	8104	8100	8104	16200	16204
Acc.%	38.45	61.04	32.5	55.34	39.18	60.83

Table 1. Recognition results on video data with with the different feature sets

5. CONCLUSION AND OUTLOOK

In the present treatise we have shown a simple method for a video editor in meeting scenarios with promising results of apx. 61%. Due to the huge feature dimension and the small amount of training data it seems reasonable to perform an aggressive feature selection in order to provide a stable training environment. Additionally the classifier output has to be further analyzed. It is noticable, that the correct classification predominates in longer sequences with some more or less random insertions. These mostly result from rapid movements in one of the other views. By applying dynamic programing it could be possible to remove some of the insertions and recieve a by far better over all performance.

6. REFERENCES

[1] M. Al-Hames et al., "Audio-visual processing in meetings: Seven questions and current AMI answers," in *Proceedings MLMI*, 2006.

[2] J. Carletta et al., "The AMI meetings corpus," in *Proceedings of the Measuring Behavior symposium on Annotating and measuring Meeting Behavior*, 2005.

[3] J. Ostermann, "Animation of synthetic faces in mpeg-4," *Computer Animation*, pp. 49–51, 1998.

[4] M. Zobl, F. Wallhoff, and G. Rigoll, "Action recognition in meeting scenarios using global motion features," in *Proceedings Fourth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ICVS), Graz Austria*, Mar. 2003, pp. 32–36.

[5] H. Rowley, S. Baluja, and Takeo Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, Jan. 1998.

[6] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29(1), pp. 5–28, 1998.

[7] B. Martinkauppi M. Soriano, S. Huovinen and M. Laaksonen, "Skin detection in video under changing illumination conditions," in *Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain*, 2000, pp. 839–842.

[8] D. Arsić, B. Schuller, and G. Rigoll, "Suspicious behavior detection in public transport by fusion of low-level video descriptors," in *Proceedings 8th International Conference on Multimedia and Expo ICME 2007, Beijing, China*, June 2007.

[9] B. Schoelkopf, "Support vector learning," *Neural Information Processing Systems*, 2001.