Some Preliminary Results on Multimodal Recognition of Events in Smart Meeting Rooms

Gerhard Rigoll

Institute for Human-Machine Communication, Munich University of Technology,
D-80333 Munich, Germany

Abstract: This paper aims to present some novel ideas developed for the work on automatic meeting transcription in the framework of a European Community sponsored research project. The goal of this project is the development of a meeting browser supporting the multimodal analysis of and access to meetings held in smart meeting rooms. Due to this multimodal aspect, speech recognition as well as visual processing techniques have to be deployed in an integrated way, resulting into some novel audiovisual processing methods to be presented in this contribution.

INTRODUCTION

This paper presents some new audio-visual processing methods that our institution has developed in the framework of the M4 project (see [1]). The abbreviation M4 stands for **M**ulti**M**odal **M**eeting **M**anager, and represents a project that is sponsored by the European Union in the 5th framework programme, as part of the IST action. The project started in early 2002 and has a duration of 3 years. The basic idea of this project is the analysis and transcription of meetings recorded in so-called smart meeting rooms, equipped with multimodal sensors. Thus, the overall objective of the M4 project is the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analyzed meetings. For each meeting, audio, video, textual, and interaction information will be available. Audio information comes from close talking and distant microphones, as well as microphone arrays. Video information comes from multiple cameras. Therefore, a variety of different data streams with a lot of different information has to be processed in the framework of this project. Speech recognition aspects consist of meeting audio transcription, where the aspect of conversational speech is heavily involved, since large parts of meetings consist of conversations between two or more people. One of the interesting aspects is that this project offers the potential to tackle spontaneous speech effects by means of multimodal and multi-stream information processing, e.g. by resolving false segmentation boundaries with support of the visual action or the facial expression that has been simultaneously observed in the video channel. Speech transcription is only one of many different aspects of this project, others include the transcription of visual information, such as e.g. the recognition of gestures, emotions or actions of the meeting participants. Therefore, a variety of novel audiovisual processing techniques were developed within this project and the purpose of this paper is to give the reader an impression of the potential of such methods for the new emerging research area of audiovisual meeting transcription.

SMART MEETING ROOM EQUIPMENT

Smart meeting rooms can be seen as part of a relatively new research field that could be called "smart environments" and rapidly gain more and more attention in the broad area of human-machine communication. Other sub-areas of smart environments are for instance "smart living rooms", "smart offices" or even "smart parking lots". All of these environments have the common feature that they are equipped with a -

From: SPECOM 2003, Academy Press, Moscow

possibly very large number – of sensors that enable the environment to communicate with the humans that are in this environment. This communication ability makes such environments “smart” in that sense, that the user can interact with devices or parts of the room and the environment is able to assist the user in his actions or intentions. In that way, the environment itself becomes interactive and the user typically performs some multimodal interaction in that environment which does not only interact with the user but also assists the user and records the events that take place while the user interacts in such an environment.

A smart meeting room is therefore a meeting room that is equipped with special sensors that are supposed to assist the participants in a meeting and that are able to record the major events while a meeting is being held in such a room. In general, it is possible to distinguish between “active smart meeting rooms” and “passive smart meeting rooms”, where the first variant would enable true interaction with the meeting participants, e.g. by switching automatically the projector on if a participant intends to give a presentation, whereas the passive variant would merely use its sensors to record the meeting events for later evaluation. Although the first variant would be very much desirable, the more realistic version for today´s available technologies is the passive variant, which is also taken up by the M4 project. In this case, the sensors consist of microphones and cameras for event recording. Fig. 1 shows an outline of the smart meeting room as it is currently realized and used in the M4 project by one of M4 project partners. It can be seen that this one is equipped with individual microphones for each speaker and several microphone arrays. There are also several cameras installed in such an order, that each camera can capture three meeting participants. An additional wide angle camera is installed to capture a complete view of the entire meeting scenario.

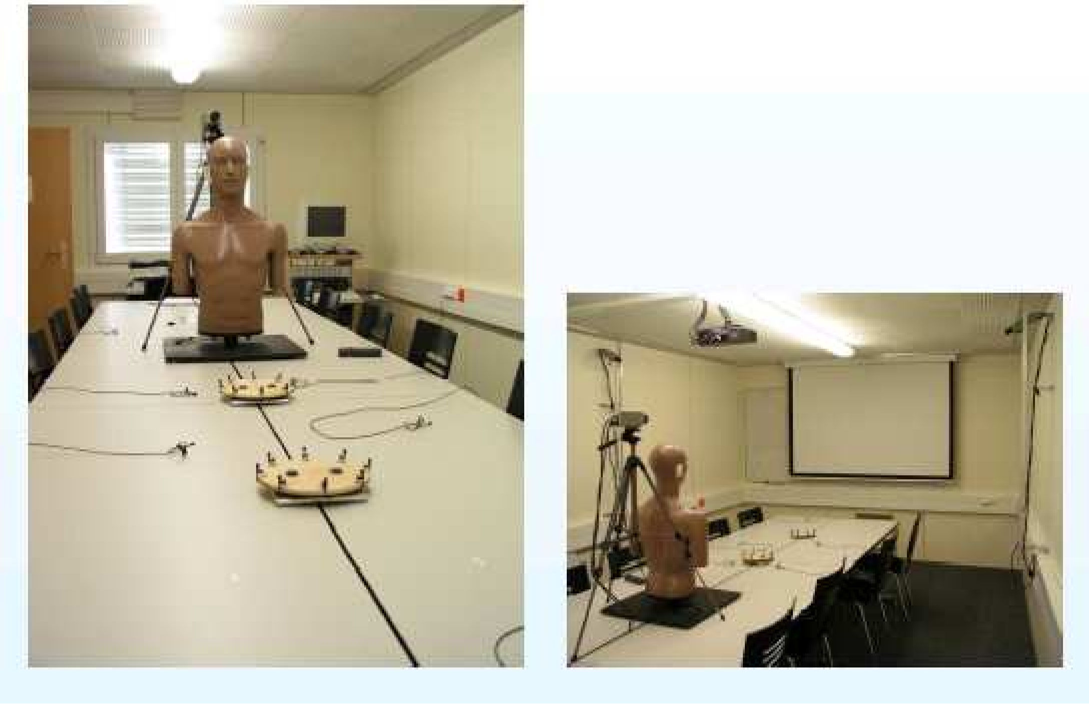


Fig. 1: Example for a smart meeting room

Clearly, the idea of the passive smart meeting room is to record the audiovisual data from the sensors described above and to analyze this data in a later off-line stage in order to transcribe as precisely as possible what happened in that meeting. This concerns of course the transcription of the spoken utterances in the meeting, however is not restricted to audio data, but also takes into account the captured video tracks in order to analyze what actions have been performed by the meeting participants. The final goal is the use these transcriptions in a meeting browser, that basically contains the minutes of the meeting in a multimedia format, containing audio and video. With those transcriptions, it is then possible to query in this browser e.g. for the discussions of the participants or to search who gave a presentation and if some voting took place. Such an information will be extremely valuable e.g. for people who could not participate in the meeting or for participants who want to see the summarization of the meeting.

Obviously, in order to achieve that objective, very demanding multimodal recognition procedures have to be carried out, and the next sections describe partially some of the methods that were developed during this project to perform multimodal recognition of meeting events.

SPEECH AND AUDIO PROCESSING

It would be beyond the scope of this paper to describe all the audio processing techniques that have to be deployed in meeting transcriptions. It is obviously clear that speech transcription in meetings is an extremely demanding speech recognition task, due to the following facts: Speech of meeting participants in such environments is a typical case for spontaneous speech recognition, since in meetings, almost every utterance is spontaneous, with hesitations, repetitions, stuttering and many other effects, except perhaps well-prepared meeting presentations, which could be closer to read speech. However, especially in meetings there are additional burdens, such as e.g. cross-talking, background talking, or background noise.

As in many of today’s speech recognition problems, an appropriate database is also in this project one of the major key factors for success. Since there are almost no databases for meeting recordings available – and especially not publicly available – a special M4 database is currently recorded. Additionally, the consortium has access to the ICSI meeting corpus (see [3,4]), where 75 meetings (72 hours) of fully annotated meeting recordings have been collected, however under different acoustic conditions as foreseen in the M4 project.

Speech recognition experiments for the M4 environment are under way, where in a first stage, training is performed on the ICSI data and the Switchboard database, which is partially suitable due to the fact that it also contains spontaneous speech. A recognition error rate of around 30% (WER) is expected to be obtained with this approach. Besides pure speech recognition, meeting recordings offer many other challenging problems of audio and speech processing. One option visualized in Fig. 2 is the detection of speaker turns:

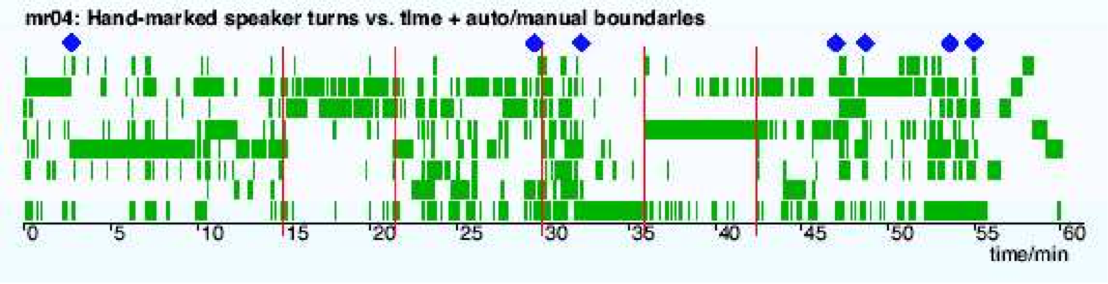


Fig. 2: Speaker turn pattern for a meeting recording

Another option is shown in Fig. 3 and concerns the term "talkativity": Once the speaker turns are found, the average speaking time of each speaker in a meeting can be found out and answers can be given to such questions as e.g. "which participant is mainly talking?" or "how much discussion is there about a certain topic?". In Fig. 3, the histogram on the left side shows the talkativity index of the different meeting participants, estimated from the total meeting time and the talk activitiy pattern for each participant.

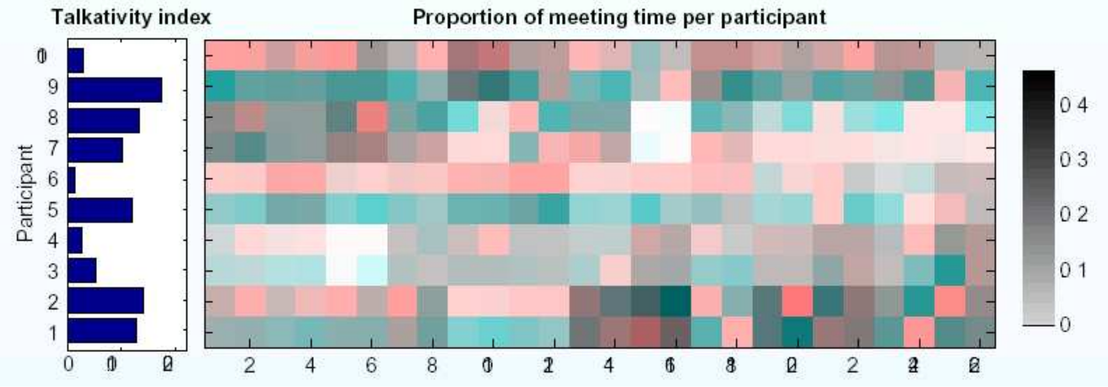


Fig. 3: Computation of talkativity for a meeting with 10 participants

MULTIMODAL PROCESSING OF AUDIOVISUAL DATA

As already mentioned, in meetings, other communication channels than audio alone are valuable sources of information. This is especially true for the visual information channel, that can be evaluated in order to analyze the events and actions that were performed during a meeting. Very often, these visual cues are in strong correlation with optical cues, e.g. when a meeting participant gives a presentation in front of the other participants, draws a sketch on the board and explains his presentation by spoken comments.

If such information is intended to be evaluated and interpreted, the key for this is to find the position of the meeting participants. If this is feasible, then it is possible to find out who did something, at what time it was done and what has been done by this person. In meetings, the most typical position of people is when they are sitting around the meeting table and with sufficiently high resolution of the camera, the most efficient way to locate the position of a person is to localize his face. Fig. 4 shows an example for multiple hypotheses of a face in a meeting situation. Neural network techniques in conjunction with colour analysis can be successfully employed for this task.

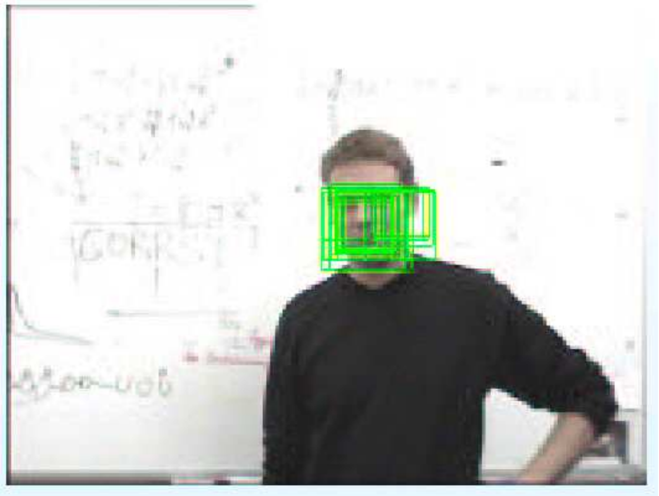


Fig. 4: Face detection in a meeting scenario

Action recognition is another important issue for meeting transcription. In this case, a few pre-defined actions are described, such as e.g. entering, leaving, rising, nodding or voting. Fig. 5 shows a few examples for such actions in meetings.

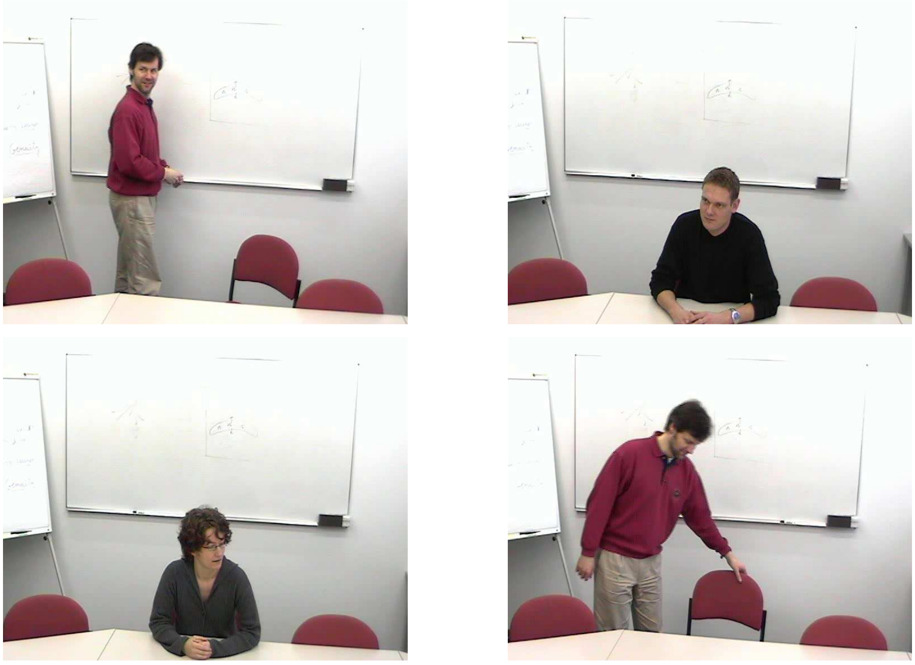


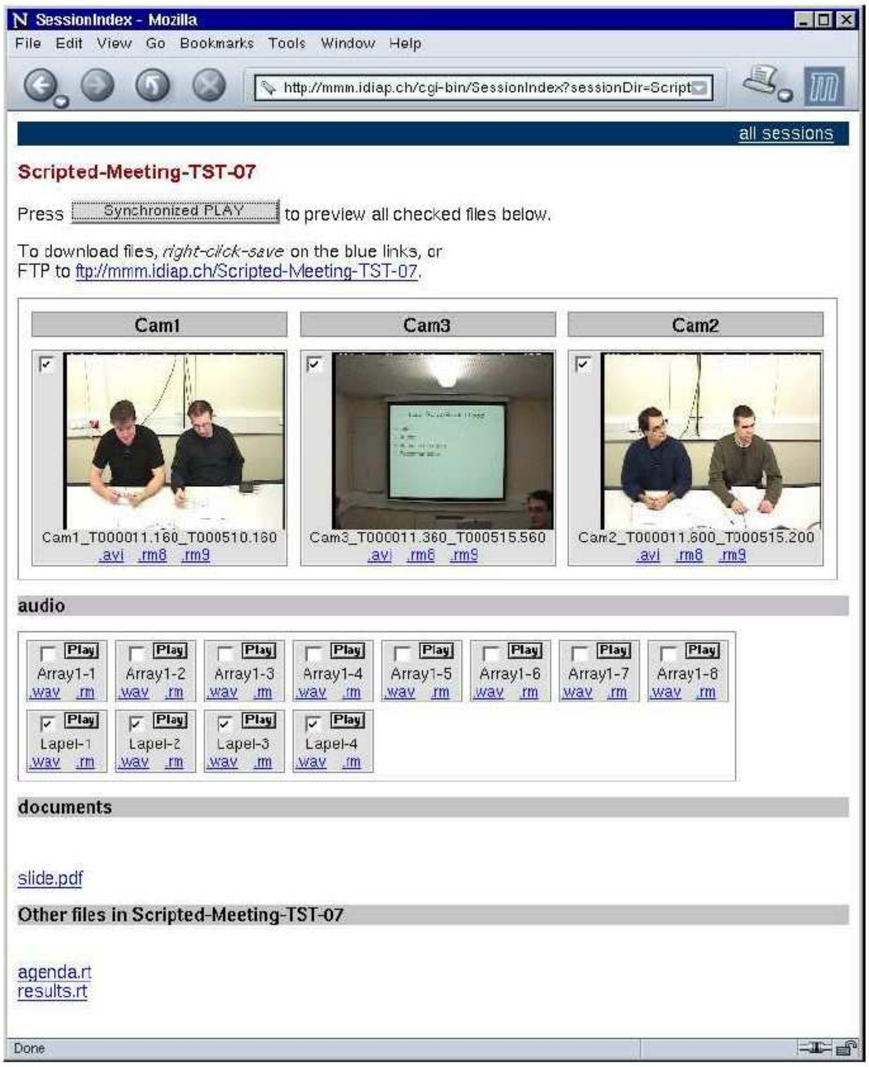
Fig. 5: Examples for typical meeting actions

Special measures are taken and described in more detail in [2] to extract global motion features from the recordings of such actions and to classify them with Hidden-Markov-Models. In this way, a respectable action recognition rate of up to approx.

80% can be reached. The further reaching goal of these activities is the recognition of complex multimodal events, such as e.g. mentioned at the beginning of this section for the presentation of a participant, where the multimodal recognition of audio-visual data streams should lead to the recognition of events such as presentations, group discussions, or even meeting interruptions, such as e.g. coffee breaks.

MEETING BROWSER

Finally, the expected result of the meeting transcription process is the display of the transcription results in a so-called meeting browser. Fig. 6 shows an example for such a system. The basic idea is that – similarly to browsing through websites – it should be possible to load the transcription and index files resulting from the multimodal meeting analysis and indexing into the browser and to navigate through these results and query the system e.g. about the content of the speech transcription, the tracking of various speakers, the analysis of the previously discussed talkativity of speakers, or the search for complex events, such as votings or presentations of participants.


SessionIndex - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://mmm.idiap.ch/cgi-bin/SessionIndex?sessionDir=Script
all sessions
Synchronized PLAY
Play
Done

Scripted-Meeting-TST-07

Press Synchronized PLAY to preview all checked files below.

To download files, *right-click-save* on the blue links, or
FTP to ftp://mmm.idiap.ch/Scripted-Meeting-TST-07.

Cam1	Cam3	Cam2
☑ Cam1_T000011.160_T000510.160 .avi .rm8 .rm9	☑ Cam3_T000011.360_T000515.560 .avi .rm6 .rm9	☑ Cam2_T000011.600_T000515.200 .avi .rm8 .rm9

audio

□ Play Array1-1 .wav .rm	□ Play Array1-2 .wav .rm	□ Play Array1-3 .wav .rm	□ Play Array1-4 .wav .rm	□ Play Array1-5 .wav .rm	□ Play Array1-6 .wav .rm	□ Play Array1-7 .wav .rm	□ Play Array1-8 .wav .rm
☑ Play Lapel-1 .wav .rm	☑ Play Lapel-2 .wav .rm	☑ Play Lapel-3 .wav .rm	☑ Play Lapel-4 .wav .rm				

documents

slide.pdf

Other files in Scripted-Meeting-TST-07

agenda.rt
results.rt

Fig. 6.: Example for a multimodal meeting browser

SUMMARY AND CONCLUSION

Some preliminary results have been presented on how multimodal recognition of events in smart meeting rooms can be achieved. It has been shown that complex algorithms from the area of speech recognition, computer vision and multimodal pattern recognition are required for such a task. The research in this area is just at its beginning stage and the author would like to point out here that this is a joint effort of the consortium representing the EU project “MultiModal Meeting Manager” (M4). It is expected that the area of meeting analysis and smart meeting rooms will have an enormous development in the next years to come and much further sophisticated methods will be required and developed in order to make the vision of automatic meeting transcription and browsing a reality.

REFERENCES

[1] http://www.dcs.shef.ac.uk/spandh/projects/m4/

[2] Zobl, M, Wallhoff, F and Rigoll, G (2003). Action Recognition in Meeting Scenarios Using Global Motion Features. *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Graz, Austria, March 2003*.

[3] Janin, A, Baron, D, Edwards, J, Ellis, D, Gelbart, D, Morgan, N, Peskin, B, Pfau, T, Shriberg, E, Stolcke, A and Wooters, C. The ICSI Meeting Corpus. *Proc. ICASSP-03, Hong Kong, April 2003*.

[4] Morgan, N, Baron, D, Bhagat, S, Carvey, H, Dhillon, R, Edwards, J, Gelbart, D, Janin, A, Krupski, A, Peskin, B, Pfau, T, Shriberg, E, Stolcke, A and Wooters, C. Meetings about meetings: research at ICSI on speech in multiparty conversations. *ICASSP-2003, Hong Kong, April 2003*.