The AMIDA Mobile Meeting Assistant: Remote Meeting Attendance Using a Smart Phone

Lukas Matena
IDIAP Research Institute
Centre du Parc, P.O. Box 592
1920 Martigny, Switzerland
lukas.matena@idiap.ch

Alejandro Jaimes
Telefonica Research
Madrid, Spain
ajaimes@tid.es

Andrei Popescu-Belis
IDIAP Research Institute
Centre du Parc, P.O. Box 592
1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

ABSTRACT

The AMIDA Mobile Meeting Assistant is a system that allows remote participants to attend a meeting through a mobile device. The system improves the engagement into the meeting of the remote participants with respect to voice-only solutions thanks to the use of visual annotations and the capture of slides. The visual focus of attention of meeting participants and other annotations serve to reconstruct a 2D or a 3D representation of the meeting on a mobile device (smart phone). A first version of the system has been implemented, and feedback from a user study and from industrial partners shows that the Mobile Meeting Assistant's functionalities are positively appreciated, and sets priorities on future developments.

Keywords

Remote access, mobile device, user interface, 3D representation, meeting annotation.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Artificial, augmented and virtual realities.

General Terms

Design, Human Factors.

Keywords

Remote access, mobile device, user interface, 3D representation, meeting annotation.

1. INTRODUCTION

Remote participation in meetings has become a necessity as co-workers are less and less often physically collocated. In particular, when participants are away from their offices at the time of a meeting, a mobile device can help them to attend the meeting. However, the remote user's engagement in the meeting might be insufficient if only an audio channel is conveyed from/to the meeting. Therefore, the challenge of conveying and displaying more information from a meeting onto a mobile device must be solved in order to better integrate a remote participant in a discussion.

MobileHCI 2008, September 2-5, 2008, Amsterdam, the Netherlands.
ACM 978-1-59593-952-4/08/09.

In this paper, we demonstrate the *AMIDA Mobile Meeting Assistant (MMA)*, a graphical interface for a smart phone that reconstructs part of the current communication situation in an instrumented meeting room, by using the output of multimodal processing tools such as the detection of the visual focus of attention of the participants and the projected slides. In the next section, we outline our objectives and compare them to available technology. In Section 3, we state the design principles of the MMA, and in Section 4 its implementation. Feedback from a user study and from industrial partners is synthesized in Section 5.

2. OBJECTIVES AND COMPARISON TO OTHER TECHNOLOGIES

The task we investigate is how to enable a remote meeting participant to better understand what is happening in the meeting room. Providing a video stream from the meeting to the remote user and back still suffers from bandwidth constraints (though 3G networks allow for video calls), and does not offer an optimal rendering of the meeting on a small device. Instead, replacing the video by a graphical representation of the meeting room, based on automatic audiovisual annotations, allows the MMA to give the user more relevant information than can actually be conveyed through a conventional video stream. In particular, beyond the words used by speakers, non verbal cues tell a lot about the speakers' role in a conversation [6].

We focus here on conveying automatically-detected aspects of "body language" to the remote user through a graphical interface, and in particular head orientation and visual focus of attention [9, 2]. In addition, real time slide capture from the meeting appears as another important type of visual information [10]. For development and demonstration purposes, we are using the data and annotations of recorded meetings of the AMI Meeting Corpus [3]. Our approach thus copes with essential limitations of mobile phones such as the small size of the screen and limited bandwidth, by taking advantage of increasing computing power and graphic capacities.

Commercial video-conferencing equipment (e.g. Tandberg or HP Halo) uses internet video streaming and special hardware (e.g. TotalView phone). Video calls are available in several P2P communication applications and they are even already available for mobile phones in 3G networks, though bandwidth constraints remain strong. To our best knowledge, there is no widespread working application that would use automatic graphical representation of a remote scene

based on real time events (excluding emoticons), as shown in a recent survey paper [8].

An interesting graphical representation of a meeting flow appears in the MIT Infospaces environment [5], which keeps track of all statements produced by each speaker, focusing on the agree/disagree dimension. This representation is however too complex for the limited screen size of a mobile device, hence potentially confusing to the general user.

3. DESIGN OF THE MMA

The design of the MMA user interface follows the guidelines proposed by L. Chitarro [4], especially the first three steps.

Mapping. A meeting processing application outputs annotations in real time, and sends them to the Hub – a client-server architecture for data exchange [1] – from where the remote user's device can retrieve them in real time using a subscription mechanism. Each of the incoming annotations must then be mapped into a multimedia event within a virtual space. This space uses simple avatars for participants, photographic representation of slides, red color flashing highlighting for speech events, and green color arrows (3D) or contours (2D) for the object of the visual focus of attention. We reserve sound and/or vibration events for further use, e.g. keyword spotting, or a "you are expected to speak" notication.

Selection. The meeting processing tools provide a range of abstracted information from a meeting, for both short and long time events (see available annotations of the AMI Corpus [3]). Moreover, multiple audio and video streams are available. It is, of course, virtually impossible and highly unpractical to fit all this information into the tiny screen of a mobile phone (usually 320x240 pixels). Therefore, given the goals of the MMA mentioned above, the following annotations were selected:

- head orientation and visual focus of attention of selected participating (i.e. "who is paying attention to whom"), giving an idea of the dynamics of the meeting;
- automatic speech recognition, used to display mouth movement;
- speaker segmentation (i.e. "who speaks when");
- clothes color, represented on the avatars (manually input at present);
- slide content and real time slide change.

In addition, the speech signal from the meeting room is conveyed to the remote participant. To further reduce the amount of information displayed at the same time, we decided to show the focus of attention of only one meeting participant selected by the user. We also experimented with the automatic selection of the participant whose focus of attention to show, such as the speaker – but as the speaker changes sometimes very fast, this was difficult to follow. In the future, rendering all foci of attention, or, better, a joint focus of attention, will likely improve the overall experience of the interface. This could be done using an advanced mechanism for annotation processing on the mobile client side.

Annotation producers that are represented graphically sometimes generate too many events to display them all, as that would produce a "jitter" on the screen that would make the graphic interface difficult to use. Therefore the amount of data sent to the Hub was reduced by filtering over the timestamps: for example, the standard output of the head orientation producer is 25 frames per second (fps), but we use only a 1 fps rate computed as the average of the 25 preceding frames received via the Hub. As for the visual focus of attention, as the change rate is slower, we sample 1 frame out of 25.

Presentation. The next design task is to place all selected visual items on the device's screen. We propose two alternative displays representing the meeting room: the 2D and the 3D view, shown respectively left and right in Figure 2, and a full-screen slide preview. The 2D view is a top-view representation, where each participant is represented by a simple avatar consisting of three basic shapes to represent the body and arms. The person or place in the focus of attention of the participant selected by the user for display is marked by a light green circle. The head orientation is represented by a sector starting from the avatar, which becomes red when the person is speaking. This color scheme has been chosen to improve the intelligibility of the interface: most important features – focus of attention and speaker state – are represented by signal colors that remain the same for both user interfaces.

The 3D view gives the user the possibility to change the view of the virtual representation of the meeting room, for instance by turning around the meeting table or by zooming on the slide screen or on any other region of the room. The focus of attention is represented by a light green arrow, while the speaking status is shown by red moving lips.

4. IMPLEMENTATION OF THE MMA

The architecture of the implemented Mobile Meeting Assistant, showing its main modules and the data flow between them, is represented on Figure 1. The system assumes that the meeting takes place in a smart meeting room [7], where audio and video streams for each participant as well as for the overall scene are captured and recorded to a media/document server. Automatic feature extraction software is used to produce annotations such as automatic speech recognition (ASR), speaker segmentation, head orientation, and visual focus of attention [9, 2]. The architecture supposes that these algorithms are running in real time and that resulting annotations circulate through the real time data distribution entity, the Hub [1]. However, as some of these algorithms do not run in real time at present, their output is simulated on past meetings from produced offline annotations.

A Java 2 Mobile Edition application enables the mobile phone's communication with the Hub using an Internet connection. The 3D interface is using Java Mobile 3D API. The application was developed on an emulator provided by Sun Microsystems and is tested on a Nokia N95 smart phone. The Java application subscribes via the Hub to the annotation flow from a particular meeting and listens for updates; upon reception of an element, it renders the corresponding graphical element directly to the GUI. If the annotation consists of an URI of a document (a captured slide for example), the application retrieves the document from the document server and displays it in the user interface.

Two types of interfaces were implemented (Figure 2), both reconstructing a view of the meeting based on the layout of

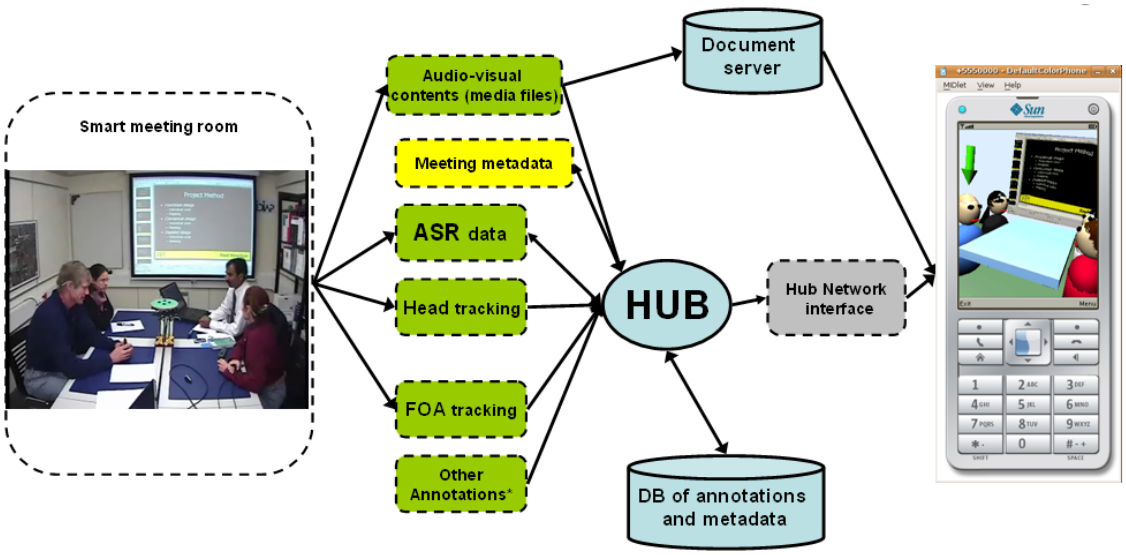

Figure 1: Architecture of the Mobile Meeting Assistant, with a view of the smart meeting room (left) and of the mobile phone emulator with the 3D interface (right).

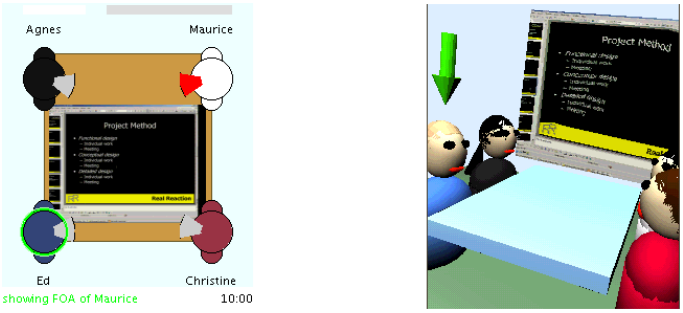

Figure 2: Snapshots of 2D and 3D interfaces.

the actual meeting room (information about seating is entered manually for the moment). As explained in the previous section, the interfaces display the following informations: current speaker, head orientation, visual focus of attention of selected participant and slides.

5. RESULTS, EVALUATION AND DISCUSSION

5.1 User Study

A small-scale user study was performed with 13 subjects, all of whom use information technology every day and have a university degree in computer science. The subjects were given a demo of the MMA application running on an emulator in real time, with a video recording of the meeting playing on second computer, for a duration of 5 minutes (meeting IS1008a from the AMI Corpus). They had then the possibility to interact with the application, for instance to change the interface, the viewing angle or the selected participant, for a maximum duration of 5 minutes. The subjects answered a questionnaire shortly after the demonstration, with the possibility to make comments in writing. Some questions required them to rate implemented or potential functionalities of the MMA, while others inquired about their own needs for a remote meeting assistant. Numeric ratings of their answers are coded in what follows from 1 to 5, 1 being *best* and 5 *worst*.

The subjects judged the MMA very positively: they liked the concept (1.5/5) and the present approach (1.9/5). They would use such an application "sometimes" (9 out of 13), mainly for design/technical meetings (11 out of 13) or business meetings (10 out of 13), but less for personal meetings (5 out of 13). They would mainly use the application while waiting at the train station or at the airport (10 out of 13), in the office or on a train/airplane (9 out of 13 both). The main limitations for use in such conditions is the available attention if the user must do something else (e.g. go to a gate or catch a train, 10 out of 13), the small size of the screen, and noise from the environment (7 out of 13 both).

In terms of user experience, users seem equally satisfied with the 2D and the 3D interfaces (2.3/5 and 2.2/5). The interface and color schemes are at the appropriate level of complexity (11 out of 13). The most appreciated information is "who is speaking when/to whom" (1.7/5), followed by the full-screen slide preview (2.0/5), the focus of attention (2.1/5) and head orientation (2.5/5). Possible features to be added in the future have been rated similarly: "you are expected to speak" alert seems the most desired one (2.0/5), followed by the "enter/leave room" alert (2.2/5), use of personalized avatars (2.4/5), and display of speech transcript (2.5/5).

Finally, most of the subjects would also use a desktop version of the MMA (11 out of 13), and many would even prefer it (8 out of 13), a fact that meets some of the explicit suggestions received from industrial partners.

5.2 Verbal Feedback: Evaluation and Future Work

The suggestions expressed verbally in the user study and at a workshop involving potential industrial partners point to the following possible improvements, which are part of

the future plans for the MMA. The MMA appears quite intuitive to use, and seems especially useful to someone who is not fully acquitained the meeting participants; however, 2D graphical conventions could be made clearer, and the 3D representation seems unnecessarily complex to some users. In both cases, slides should be made more visible, as well as the identity of the speakers. Ideally, more subtle body language, such as signs of disagreement and agreement, should be conveyed.

More realism was also required by the subjects, who suggested to include actual photos of users, or at least let them choose their own avatars. Accessing information about the other participants as well as consulting meeting documents was another suggestion, while the slide capture itself appeared to be already a good candidate for a commercial product. Of course, a system running on a mobile phone is a primary objective, bringing the audio stream to users via the data stream (using VoIP), and synchronizing it with annotations and with the graphical representation of the meeting. This implementation is ongoing at the time of writing.

An important development will be the converse representation of the remote participant into the meeting room, because people in the meeting also need to improve their understanding of his/her presence beyond pure speech. At this point, the system could also be extended to support purely remote meetings, so that it creates the feeling of a virtual meeting room on a set of mobile devices.

Acknowledgments

This work was supported by the European IST Programme through the AMIDA Integrated Project FP6-0033812.

6. REFERENCES

[1] AMIDA. Commercial component definition. Deliverable 7.2, AMIDA Integrated Project IST-033812 (Augmented Multi-party Interaction with Distance Access), November 2007.

[2] S. Ba and J.-M. Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In S. Renals, S. Bengio, and J. G. Fiscus, editors, *Machine Learning for Multimodal Interaction III*, LNCS 4299, pages 75–87. Springer-Verlag, Berlin/Heidelberg, 2006.

[3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI Meeting Corpus: a pre-announcement. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction II*, LNCS 3869, pages 28–39. Springer-Verlag, Berlin/Heidelberg, 2006.

[4] L. Chittaro. Visualizing information on mobile devices. *Computer*, 39(3):40–45, 2006.

[5] H. Drew and J. Donath. Information spaces: Building meeting rooms in virtual environments. In *CHI 2008 (25th SIGCHI Conference on Human Factors in Computing Systems)*, Florence, Italy, April 5–10 2008.

[6] A. Kendon. *Nonverbal communication, interaction, and gesture: selections from Semiotica*. Mouton, The Hague, 1981.

[7] D. J. Moore. The IDIAP Smart Meeting Room. Communication 02-07, IDIAP Research Institute, July 2002.

[8] A. Nijholt, R. Rienks, J. Zwiers, and D. Reidsma. Online and off-line visualization of meeting information and meeting support. *The Visual Computer*, 22(12):965–976, 2006.

[9] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.

[10] A. Vinciarelli and J.-M. Odobez. Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006.