Meetings and Meeting Modeling in Smart Surroundings

Anton Nijholt, Rieks op den Akker and Dirk Heylen
Centre of Telematics and Information Technology (CTIT)
University of Twente, PO Box 217
7500 AE Enschede, the Netherlands
{anijholt|infrieks|heylen}@cs.utwente.nl

Abstract

In this paper we survey our research on smart meeting rooms and its relevance for augmented reality meeting support and virtual reality generation of meetings in real-time or off-line. Intelligent real-time and off-line generation requires understanding of what is going on during a meeting. The research reported here takes place in the European 5th and 6th framework programme projects M4 (Multi-Modal Meeting Manager) and AMI (Augmented Multi-party Interaction). Both projects aim at building a smart meeting environment that is able to capture in a multimodal way the activities and discussions in a meeting room, with the aim to use this information as input to tools that allow real-time support, browsing, retrieval and summarization of meetings. In these projects many European research groups participate. Our aim is to research (semantic) representations of what takes place during meetings in order to allow generation, e.g. in virtual reality, of meeting activities (discussions, presentations, voting, etcetera). Being able to do so also allows us to look at tools that provide support during a meeting and at tools that allow those not able to be physically present during a meeting to take part in a virtual way. This may lead to situations where the differences between real meeting participants, human-controlled virtual participants and (semi-) autonomous virtual participants disappear. In this paper we introduce our research aims and ideas and we illustrate them with examples taken from many different projects in related areas.

1 Introduction

When people meet there is interaction. Interaction can be focused and it can be unfocused (Goffman 1963). Meeting means exchange of information. When two people meet this can be information about social status, by looking at clothes or posture. However, they can also start a discussion and exchange information about their family or about themselves. Whether the information exchange, or the interaction, is focused or unfocused, there need to be some common ground in order to make it effective. People meet, people gather, notice each other and communicate with each other, verbally and nonverbally, focused and unfocussed. When two people meet there usually is face-to-face interaction. Attempts to model human face-to- human face interaction have been in order to allow a translation to human-computer interaction. More recently, attempts have been made to model multi-party interaction. If more than two people meet there is another or there are others when you address a particular person. You are aware of the others and they play a role in your communication behavior, verbal and nonverbal.

There are many situations where people meet. In this paper we look at formal meetings, meetings with invited participants and with an agenda that reflects shared goals. Goals may be the willingness to discuss issues, to come to agreement and decision and willingness to accept the outcome of the meeting. Participants of such meetings see each other during different meetings, meetings where a previous meeting is summarized and discussed using its minutes. This is preferably done before starting discussions on new topics or before continuing discussions started in previous meetings. People get to know each other, sometimes know what to expect when someone takes the floor, learn about the body language of other meeting participants, learn how to interpret a participant's verbal utterances, learn about his background, his role during the meeting and learn about his emotions and his humor. In short, meeting participants form a community. They know each other from previous meetings, they share knowledge, culture, ideas and feelings, and generally they share goals. Having shared goals allows self-disclosure during breaks, lunches or informal follow-ups of a meeting (drinks, dinners, outgoing activities, email exchanges, pictures, etc.), smoothen exchanges during next meetings.

How can we support such meeting activities? When meetings take part in smart environments, how can we make use of technology based on models of activity perception, multi-party interaction and event semantics to support meeting participants in their activities (on-line and off-line) and how can we model meeting participants as agents in such a way that remote participation or virtual participation becomes possible?

Our assumption is that people want to meet. They prefer to experience the whole gamut of activities that are associated with physical meetings and only when there are no other possibilities they seem to be willing to enter video-conferencing and computer-supported collaborative work environments. Rather than looking at ways to minimize meetings or to oblige people to use specialized meeting support technology we prefer to consider meetings as a particular case of natural interaction activity between different humans or even between humans and objects or environments. This does not mean that we don't want to distinguish between different kinds of gatherings or meetings. For example, it can be essential to know what a particular meeting is about, what the goals of the meeting or the goals of its participants are and what the reason is to have this particular meeting at this particular moment, in order to be able to understand what is going on during the meeting and, consequently, provide intelligent support to the participants of the meeting. Knowing about meeting goals helps in interpreting the actions (including the spoken utterances of the participants) during the meeting. However, it is also useful to take a more general point of view that will help us to design more advanced and attractive meeting environments.

In this paper the more general point of view is that of ambient intelligence. Ambient Intelligence has been defined as ubiquitous computing + social and intelligent interfaces. Here, 'intelligent' may refer to the original and global AI (Artificial Intelligence) paradigm, its domain-dependent specialization (as in several generations of expert systems), or its translation to agent intelligence with its distinction in believes (knowledge about an application-relevant part of the world), desires (goals of the agent in this particular part of the world) and intentions (short-term goals that bring the agent closer to its goal using a reasoning process). Interfaces between users (visitors, inhabitants) of ambient intelligence environments can be everywhere: in objects that are natural in the environment, in walls or in special devices, including PDA's or tablet PCs. Important are the social aspects of the interfaces in ambient intelligence. The environment should be able to use knowledge about our emotions, about our moods and about our personality when it tries to support us. When useful, it should be possible to induce development of social relationships between the ambient intelligence environment and its inhabitants. Moreover, usefulness of environments should not be understood in terms of efficiency or in terms of efficiency alone. Entertainment issues, feelings of enjoyment, allowing the inhabitant to feel at ease and feel comfortable are important as well.

In the next section of this paper we introduce our view on ambient intelligence and the roles of real and virtual humans in ambient intelligence environments. Section 3 is devoted to a discussion on some of the European projects in which we are involved and that have guided our insights in ambient intelligence research issues related to a virtual reality continuum. That is, we discuss how these projects contribute to the design and implementation of our view on ambient intelligence environments as discussed in section 2. We will also include observations that have become available from the Ambience project, another European project that addresses ambient intelligence issues. Section 4 is about meeting modeling. We survey our research on meeting modeling in the context of the AMI project. We zoom in on models for meeting modeling, addressee detection and the development of annotation tools. In section 5 we introduce our views (in the context of meeting situations) on the virtual reality continuum when considering meeting situations. This whole paper is an attempt to generalize from our observations obtained in the context of meetings supported by a smart environment to a context of whatever kinds of activities in ambient intelligence environments. A short summary of our findings and some notes on future research can be found in the final section (section 5) of this paper.

2 Ambient Intelligence Requirements

As mentioned, ambient intelligence has been defined as ubiquitous computing plus social and intelligent interfaces. As may have become clear from the introduction, we are interested in the interfaces. In the ambient intelligence point of view interfaces don't need to be visible. The environment is the interface. Nevertheless, there may also be many identifiable objects and displays that can be addressed in this environment. And the inhabitant or visitor may have his or her personal assistant, available on a PDA, a tablet PC or migrating from environment to environment that can be addressed. Below are the issues we want to distinguish when looking at ambient intelligence environments.

2.1 Interpretation of Events and Activities in the Environment

This includes social and intelligent interactions in the environment between humans, between humans and objects, between humans and autonomous embodied agents (virtual humans) and interactions with the environment in general (not addressing an object or human in particular). Input can be obtained from sensors for sound, image, and haptics. The interaction that has to be perceived does not only include all aspects of focused interaction, but also aspects of unfocused interaction. Interpretation requires the fusion of all modalities that can be perceived by the environment into various levels of annotation schemes and semantic/pragmatic representations that allow further processing.

2.2 Providing Real-time Support

Based on the interpretation and the resulting representation(s) the environment, its virtual inhabitants and its smart objects need to provide real-time support to the human inhabitants or visitors of the environment. They need to decide how to present this support, through which modalities, and with which content. On the one hand there can be implicit and explicit calls for support by the inhabitant or visitor of the environment, on the other hand the environment can decide that this particular person or group of persons can benefit from its previously obtained knowledge and may suggest or perform, preferably welcome, spontaneous real-time support.

2.3 Multimedia Retrieval and Reporting

Recalling what has been going on in an ambient intelligence environment is another issue. Automatic annotation of information coming from different input sources and fusion of information coming from different input modalities into a representation that allows support to the inhabitant or visitor of an environment also allows indexing and retrieval of events, (hypermedia) browsing of activities, reporting and summarization, and a replay, e.g. in virtual reality, of what has been going on in a particular period of time or before, during and after a particularly interesting event in the environment. For the environment the collecting of such information is useful since it can help in better supporting, in real-time) its inhabitants. These inhabitants may ask such information during a gathering or the environment may supply them with this information when it considered this useful. The interests of off-line users may also guide the attention of the environment in future observations.

2.4 Autonomous and Semi-autonomous Embodied Agents

Autonomous embodied agents can be part of an ambient intelligence environment. However, we can as well have embodied agents in the environment that are real-time controlled by a distant human being or that have been sent to the environment to represent a distant human being, that is, a human not able to be there in person or to take part as a real-time controlled embodied participant of activities going on in the environment. Obviously, a human-controlled virtual being can turn into a (probably less perfect) autonomous embodied agent representing its distant owner when it become less interesting to participate in real-time and a temporary autonomous embodied agent can change into a human-guided agent when activities require attention and real-time guidance by its distant human owner. For these applications we need to be able to present a real-time (a more or less perfect virtual reality) replay of what is happening in the environment in order to allow distant, real-time participation.

2.5 Controlling the Environment and its Inhabitants

Obviously, there can be on-line observation and participation in ambient intelligence or smart meeting environments. Capturing the events into representations that allow retrieval, browsing, summarization and multimedia generation also allows others (owners, providers, visitors) to use this information to influence and control the inhabitants and visitors of these environments. Clearly, this issue is very much related to privacy questions, that is, who has access to this information and who owns the ambient intelligence environment? The inhabitants of an environment are spied on. How does this influence their behavior? Knowing that there are eyes and ears that observe their behavior in unknown ways (details of perception, details of interpretation) may have a negative impact on natural behavior of inhabitants and visitors of ambient intelligence environments and therefore will have negative consequences for the performance of the environments. Due to these eyes and ears, available in natural objects and more or less hidden in the environment, we may even ask whether being the sole inhabitant of such an environment is in fact impossible[1]. Being there assumes to be part of a gathering and also assumes behaving as being in a public environment, including feelings of presence, co-presence, focused and unfocussed interaction behavior (Goffman 1963).

Some of these issues we discussed earlier, for instance in the context of interactive performances where human performers have to interact with objects and virtual performers in a virtual environment (see Nijholt 2000), in the context of social embodied agents (see Nijholt 2003) or in the context of presence, alienation and privacy (see Nijholt et al. 2004;Nijholt 2004). However, in particular our involvement in two European projects on meeting environments (M4 in the 5^{th} framework and AMI in the 6^{th} framework) have been fruitful in developing these ideas further, in particular the issues mentioned in the last two bullets above. For that reason we will present and discuss these projects in the next section.

3 Modeling Meetings: From Signal Processing towards Interpretation

3.1 M4: Multi-Modal Meeting Manager

In this section we first introduce the M4 project. M4 (Multi Modal Meeting Manager) is a large-scale project funded by the European Union in its 5^{th} Framework Programme[2]. M4 is concerned with the construction of a demonstration system to enable structuring, browsing and querying of an archive of automatically analyzed meetings. The archived meetings will have taken place in a room equipped with multimodal sensors.

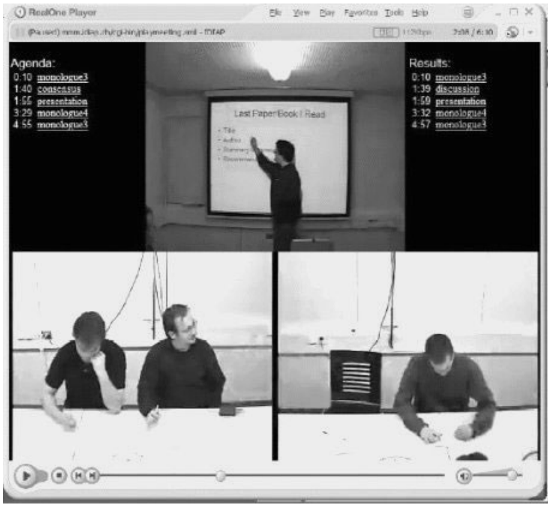
Obviously, events and interactions that take place in a meeting room are of multimodal nature. Apart from the verbal and nonverbal interaction between participants, many events take place that are relevant for the interaction between participants and that therefore have impact on their communication content and form. For example, someone enters the meeting room, someone distributes a paper, the chairman opens or closes the meeting, ends a discussion or asks for a vote, a participants asks or is invited to present ideas on the whiteboard, a data projector presentation is given with the help of laser pointing and later discussed, someone has to leave early and the

[1] Look at remarks made by Michael Coen from MIT Labs about the effects of smart environments on their inhabitants: “The notion of being alone may disappear, or it may be changed drastically.” And, “You may be in a room that’s always alive and aware. And from my experiences here...when the space is ‘off,’ you feel it. You notice that it’s not reacting. There’s a void.”

[2] M4 started on 1 March 2002 and has a duration of three years. It is supported by the EU IST Programme (project IST-2001-34485) and is part of CPA-2: the Cross Programme Action on Multimodal and Multisensorial Dialogue Modes.

order of the agenda is changed, etc. Participants make references in their utterances to what is happening, to presentations that have been shown, to behavior of other participants, etc. They look at each other, to the person they address, to the others, to the chairman, to their notes and to the presentation on the screen, etc. Participants have and use facial expressions, gestures and body posture that support, emphasize or contradict their opinion, etc.

The aim of the M4 project is to design a meeting manager that is able to translate the information that is captured from microphones and cameras into annotated meeting minutes that allow for high-level retrieval questions, and for summarization and browsing. In fact, but this is certainly too ambitious for the current project, it should be possible to generate everything that has been going on during a particular meeting from these annotated meeting minutes, for example, in a virtual meeting room, with virtual representations of the participants.



Agenda:
0:10 monologue3
1:40 consensus
1:55 presentation
3:29 monologue4
4:55 monologue3

Results:
0:10 monologue3
1:39 discussion
1:59 presentation
3:32 monologue4
4:57 monologue3

Figure 1: Three cameras capturing a mock-up meeting

In order to collect multimodal meeting information scripted meetings have been organized in which participants act according to prescribed rules that define periods of monologue, discussion, note taking, or a whiteboard presentation. The corpus thus obtained allows study of meeting participants' behavior. In Figure 1 we show a three-camera view of a meeting between four persons. In addition to the cameras there are lapel microphones and circular microphone arrays available for the meeting manager to capture audio. In the near future it is expected that white board pen capture can be added.

On a more detailed level the objectives of the project are the collection and annotation of a multimodal meetings database, the analysis and processing of the audio and video streams, robust conversational speech recognition, to produce a word-level description, recognition of gestures and actions, multimodal identification of intent and emotion, multimodal person identification and source localization and tracking. Models are needed for the integration of the multimodal streams in order to be able to interpret events and interactions. These models include statistical models to integrate asynchronous multiple streams and semantic representation formalisms that allow reasoning and cross-modal reference resolution. These models form the basis of browsing, retrieval, extraction and summarization methods. Textual "side information" (the agenda, discussion papers, slides) enables the application of useful constraints. It may be used to adapt the language model of the speech recognizer or as query expansion information for retrieval.

A straightforward meeting browser can follow the structure of an agenda. Each agenda item can be associated with different views on that topic. For example, a textual summary, a diagrammatic discussion flow indicating which participants were involved (speaker turn patterns), and audio and video key frames that give the essence of the discussion. Obviously, in order to track the discussion and find the interesting parts features need to be distinguished that can be recognized by the meeting manager.

Presently there are two approaches that are followed. The first one is the recognition of joint behavior, that is, the recognition of group actions during the meeting. Examples of group actions are presentations, discussions, consensus and note taking. Probabilistic methods based on Hidden Markov Models (HMMs) are used for this purpose (McCowan et al. 2003). The second approach is the recognition of the actions of the individuals independently, and to fuse them at a higher level for further recognition and interpretation of the interactions. When looking at the actions of the individuals during a meeting several useful pieces of information can be collected. First of all, there can be person identification using face recognition. Current speaker recognition using multimodal information (e.g., speech and gestures) and speaker tracking (e.g., while the speaker rises from his chair and walks to the whiteboard) are similar issues. Other, more detailed but nevertheless relevant meeting acts can be distinguished. In Zobl et al. 2003 recognition of individual meeting actions by video sequence processing in the context of the M4 project is discussed. Examples of actions that are distinguished are entering, leaving, rising, sitting, shaking head, nodding, voting (raising hand) and pointing (see Figure 2). These are rather simple actions and clearly they need to be given an interpretation in the context of the meeting. Or rather, these actions need to be interpreted as part of other actions and verbal and nonverbal interactions between participants. Presently models, annotation tools and mark-up languages are being developed in the project that allow the description of the relevant issues during a meeting, including temporal aspects and including some low-level fusion of media streams. Higher-level fusion, where also semantic modeling of verbal and nonverbal utterances is taken into account has not been done yet. In some cases it turns out to be more convenient to make shortcuts to a pragmatic level of fusion using knowledge from the application.

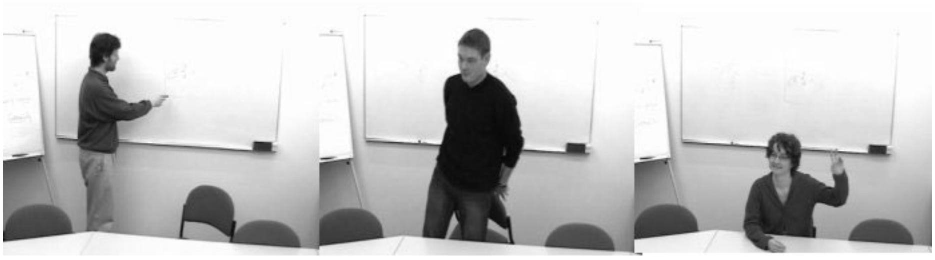


Figure 2: Pointing, rising and voting

The M4 meeting manager captures the events and interactions in the meeting room. After capturing the gathered information becomes off-line available for both participants and non-participants. Clearly, we can look at the project as research on smart environments and on ambient intelligence. However, there is no explicit or active communication between user and environment. The user does not explicitly address the environment, although it would be possible, but not done in this project, that a meeting participant explicitly addresses the meeting manager the way she would address a human note taker during a meeting. Currently, the environment registers and interprets what's going on, but is not actively involved. The environment is attentive, but does not give feedback or is pro-active with respect of the users of the environment. Real-time participation of the environment requires not only attention and interpretation, but also intelligent feedback and pro-active behavior of the environment. It requires also presentation by the environment of multimedia information to the occupants of the environment.

Our involvement in the project is modest and it should be understood that most of what we explained above is based on work done by our partners. In our work, see e.g. Jovanovic (2003), we try to explore different aspects of the interpretation point of view. In addition we hope to integrate recent research in the area of more traditional multimodal dialogue modeling (Hofs et al. 2003). These issues will become more important in the recently started AMI project, an overlapping successor project of M4.

3.2 AMI: Augmented Multi-party Interaction

The AMI (Augmented Multi-party Interaction)[3] project is concerned with new multimodal technologies to support human interaction, in the context of smart meeting rooms and remote meeting assistants. The project aims to enhance the value of multimodal meeting recordings and to make human interaction more effective in real time. These goals are being achieved by developing new

[3] AMI started on 1 January 2004 and has a duration of three years. It is supported by the EU 6th FP IST Programme.

tools for computer supported cooperative work and by designing new ways to search and browse meetings as part of an integrated multimodal group communication, captured from a wide range of devices. The project also makes recorded and annotated multimodal meeting data widely available for the European research community, thereby contributing to the research infrastructure in the field.

In the next paragraphs we introduce the AMI project. Clearly, since the project has to start yet, we have to confine ourselves to the project proposal and the different research tracks that have been defined there. From the point of view of the virtual reality continuum (see the next section) the following tracks are especially relevant:

Understanding Meetings: Which meeting characteristics play a role in order to understand the group's communication? Multimodal turntaking dynamics and multi-party interaction modeling are general areas of research. How do turntaking and dialogue structure depend on these meeting characteristics? Examples of characteristics are size, status differences, familiarity with each other, the setting, the goal or task (maintaining sociality, sharing information, generating ideas), etc. Although presently M4 is about face-to-face discussions, other meeting modes, supported by communication technology, can be considered, for example allowing asynchronous communication or video-conferencing. That is, in AMI not only face-to-face but also remote meeting dynamics has to be studied. Clearly, a wealth of research has been done in these areas and can be made use of, but in addition to that meeting support research, here we need also the environment to understand the meeting in order to allow later access for retrieval, replay and explanation.

Uni- and Multi-modal Recognition: There are many challenges for audio and video processing in smart environments. There are multiple sound sources, speech is conversational and there may be non-native speakers, to mention a few problems for speech recognition. For video processing we have to deal with unrestricted behavior of participants with variations of appearance and pose, different room conditions, occlusion, etc. Speaker turn detection, speaker localization and speaker tracking can be done using speech recognition and identification; visual processing is needed for visual tracking, face detection and recognition, facial expression recognition, gesture and action recognition. However, multi-channel processing, i.e., combination of audio and video streams allow better and more complete person identification and tracking and understanding of human-human interaction in a smart meeting environment. Multimodal syntactic and semantic information need to be extracted in order to recognize and interpret participant behavior, participant interaction and meeting events.

Multimodal Content Abstraction and Multimedia Presentation: Retrieval from meetings and browsing of meetings requires a natural structuring of meeting content. This structuring is obtained from recognition and interpretation of sequences of meeting acts and indexing the multimodal recordings. Some example questions that the AMI demonstration system should be able to answer are: Who were the participants? Was the agenda covered? How did the discussion progress? What was the atmosphere? Can I have a summary of the meeting? Segmentation of a meeting can be done from different viewpoints. We can look at events such as discussion, monologue, note taking, presentation (as is already done in the M4 project), but also at a structuring in terms of decision points, task assignments and topic shifts. An intelligent meeting browser can be designed that uses a hypertext view of the meeting in which the different structuring viewpoints are embedded.

Remote meeting assistant: One of the issues that will be explored in the AMI project is the design of a real-time, on-line remote meeting assistant. The system will allow a remote participant to a meeting to browse recent events in the meeting or to be automatically alerted at points of interest. Obviously, this empowerment of a remote participant can be useful for others present at the meeting too.

3.3 Related Research Projects

There have been several other research projects concerned with the computational modeling of meetings or, more modestly, the development of tools that help to support meetings or to off-line review and retrieve information available in recordings of meetings. For example, the ICSI project is also concerned with the development of a system for recording and browsing meetings, however, it is only based only on audio data (Morgan et al. 2001). A project very much related to M4 is the Meeting Room project at Carnegie Mellon University (Schultz et al. 2001). It is concerned with the recording and browsing of meetings using audio and video data. Closely related to AMI is for example the work done at the University of California, San Diego, which includes the development of methods for person identification, current speaker recognition, models for face orientation, semantic activity processing and graphical summarization of events. There is both work on intelligent meeting rooms (Mikic et al. 2000) as on smart environments in general (AVIARY: Audio-Video Interactive Appliances, Rooms and sYstems see Trivedi et al. 2000). Neem (Ellis and Barthelmess 2003) is a project of the University of Colorado that aims at introducing different intelligent agents in a distributed business meeting environment. These agents have to assist the meeting participants. Three agents are considered: an informing agent (assisting in obtaining necessary information, e.g. through a web search), a social agent (helps to build common ground) and an organizational agent (keeping track of time, etc.). Underlying their behavior is Bales' Social interaction Systems (Bales 2001) theory and organizational theories of problem solving. The Ambiance project, done in the context of a European project, is also more general than 'just' an attempt to model meeting situations. Rather it looks at smart home environments (Aarts et al. 2003), requiring much more modeling of the environment, including the many objects that can play a role in activities among inhabitants or between inhabitants and the global environment.

4 Meeting Modeling

In this section we have a few preliminary observations on meeting modeling. The various behaviors of peoples in a meeting can be analysed and studied from different perspectives. Meetings are social events: familiarity, social roles, personalities influence the behavior of participants. In many meetings a group meets to work on a project, conversations take place that have the form of a discussion. The task of the group implies taking decisions what to do to reach the goals of the project, and often to become clear about the goals of the project. Thus, an important part of a meeting model, a model that describes the joint meeting activities, is a discussion model. We could look at the meeting as just a series of conversational, verbal or non-verbal behaviors, observe for instance turn taking and turn giving behavior, or see how topic change is realized, or how participants address other participants, but we feel that without taking into account the goals that the participants want to realize by meeting we can not fully understand their behaviors and the joint activities that take place. It is the goal of the group and the -possibly conflicting-interests of the participants that finally motivate what is being said and how people react on each other.

To give a concrete example, consider the following situation. After a student has given his final presentation of his master thesis, a small group of people, involved in the student's project, has to judge the student's work. The judgment has to be expressed in the form of a mark on a scale between 5 and 10. The four people meet and they have about a quarter of an hour to come to a decision; the student is waiting outside the meeting room for the outcome of the decision.

Although this is a rather simple situation: the topic of discussion is clear, the possible outcomes of the process are clear and fixed in advance, and there may even be prescriptions what aspects have to be taken into account for such a decision, many of the ingredients of discussions in which a group has to make a decision can be observed in this situation. The following question may be of interest for such a proces.

- How was the decision made?
- Did all members agree on the outcome?

- How long did it take before a decision was made?
- Was the discussion well organized and structured or were there many topical shifts?
- Did everyone have the chance to give his opinion?
- Was there interaction between participants having different opinions?
- Were there argument given for or against statements?
- Was there a discussion about the criteria that had to be taken into account?
- Was there a discussion about the weights of the different factors that were of influence on the outcome?
- Was there a group member who was convinced by other members and changed his opinion?
- Was every member evenly involved in the discussion or were there clearly distinguished parts in which some members showed more involvement than others?

Notice that we don't ask whether the outcome of the decision making process was a *rational* one.

Of interest is the way the group comes to an agreement, not whether the conclusion is a reasonable or logical conclusion. The model is a *descriptive* model not a model that prescribes how the participants *should* behave or discuss, or how they should come to a conclusion.

If we observe all relevant information, and aspects of conversational behavior, of a large number of similar groups making the same task, we can compare the results and see what factors influence the outcome of the decision and the time it took the group to come to a decision.

A meeting model should be general enough so that it models not only one type of meeting in which a group discusses one specific topic, but whatever topics and issues that are discussed. The meeting model needs a model of discussions in general. What are the basic elements of a discussion and how are they structured?

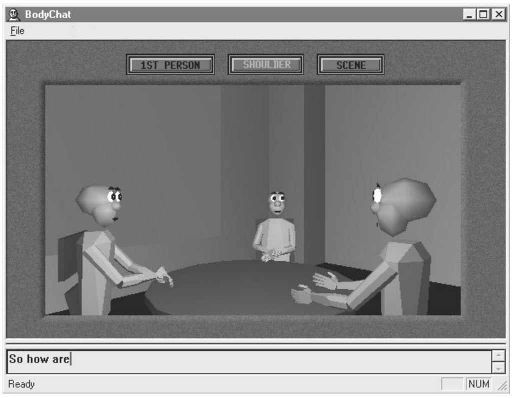
A discussion has a topic: the issue the discussion is about. The topical structure of a conversation show where a subtopic or a new topic is introduced and by whom. The discussion starts when someone gives his opinion, explains his position and gives the floor to other participants to give their opinion about it. We can distinguish a number of types of contributions to the discussion: give a new statement or opinion, react on a previous given statement, either by agreeing or disagreeing, or by partial agreeing with the statement. One can ask someone for his opinion, or ask for clarification. Finally, one can ask whether every one agrees on a particular conclusion.

For all these types of actions people use verbal and non-verbal expressions to communicate them.

The higher level information stored in the model, that is the information on the level of the discussion and the decision making process, is 'backed up' by information about events on a lower level of general speech acts and conversational behavior: the transcripted speech, the voice and prosody of the speech, the information about nonverbal conversational behavior, like head nodding, pointing gestures. From observations obtained form data received over the video and audio channels we may conclude for instance that the speaker is person A and that he strongly disagrees with the current statement of the discussion.

We may view the meeting browser as an interface of an meeting expert system that can be asked not only to give information about an event but also to show the audio and video data that together form the evidence for its conclusions about what happened in terms of semantic actions.

The *back up* relation between the information on the higher semantic level of group processes and the information on the lower level of individual behaviors that take place is one of the types of relations that exists between the various actions that we distinguish in the meeting model. The *constitutive* relation between two types of actions x and y is that relation that we express by saying that a person is doing x by doing y. An example is: switch on the light by pressing a button. Another is to vote by raising the arm. The constitutive relation can either be conventional (ritual)



So how are

Figure 3: BodyChat: Conversational gestures

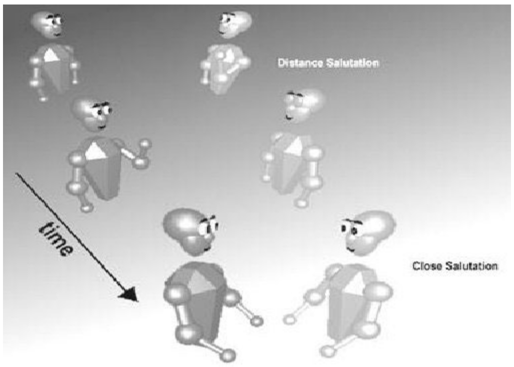


Figure 4: People meet: Salutations

or natural and based on causal relations between the two events. Other types of relations between actions are the *sequential* relation; action: x is followed by action y, and the *joint-relation*: x and y are simultaneous actions on the same level that together constitute one joint-action: shaking hands is a typical example.

5 Meetings in a Virtual Reality Continuum

As may have become clear in the first sections of this paper, developments in the area of ambient intelligence or in more restricted environments such as smart meeting rooms and future workspaces have drawn attention to the modeling of multiparty interaction, where the members of the party may be human only or, when smart objects and other support technology become available, both humans and objects. There is an obvious trend in meeting support technology to allow remote participants or to only have geographically distributed meeting participants. This has been the start of research on video conferencing and collaborative environments where attempts were made to provide information about gaze in order to facilitate the turntaking process (see e.g. Vertegaal 1998). Again, in ambient intelligence environments and certainly in smart meeting rooms similar research issues emerge with the aim to understand behavior, interactions and events, while making use of audio, video and biometric sources. As mentioned before, this information may as well be used to generate virtual reality representations of meeting participants in a virtual meeting room or an augmented reality supported physical meeting room. Meeting participants can be physically present, they can be represented by an (embodied) agent that alerts and supports when things become interesting (just as the remote meeting assistant) - but otherwise is rather passive - or they can be immersed in the (distributed) virtual environment together with the other participants, all represented as avatars mimicking their owners.

In the subsections below we show some examples from the literature and some of our own research.

5.1 Multi-party Interaction: BodyChat and Situated Chat

hesubfigure In virtual reality environments examples of research on multi-party interaction can be found. For example, Vilhjálmsson (1998) has worked on BodyChat (Figure 3, Figure 4), a chat environment system that allows users to communicate via keyboard input, "while their avatars automatically animate attention, salutations, turn taking, back-channel feedback and facial expression, as well as simple body functions as the blinking of the eyes." Hence, human-like conversational behavior for virtual humans that represent real users is simulated. In this system, apart from what is derived from the situation and the utterances, there is not necessarily a relationship between what a particular chat participant is doing in real-life (posture, gestures, facial expressions) and its nonverbal communication characteristics in the virtual world. It is the avatar that knows how to use his body during communication. This work has been continued in a project called Situated Chat (Vilhjálmsson). In addition to the social conversational rules Situated Chat also used a discourse context model to automatically generate referring gestures in the shared visual environment of the animated avatars.

Translation of this work to a smart meeting environment is straightforward. Once we can capture the events in a physical meeting room we can translate them to events in a virtual meeting room (see e.g. Figure 5) and add remote participants or add model-based behavior to virtually represented participants. For example, focus tracking (Stiefelhagen 2002) can be enhanced and converted into gaze behavior of virtual meeting participants. Assigning desirable properties to avatars that represent human participants during a meeting may much more smoothen the progress of a meeting than when the real participants are represented with all their particularities. This view allows a particular participant to become more lively through more extrovert gestures and facial expressions, it allows to convert a non-native speaker to a native speaker and it even allows to change the physical appearance of a particular participant.

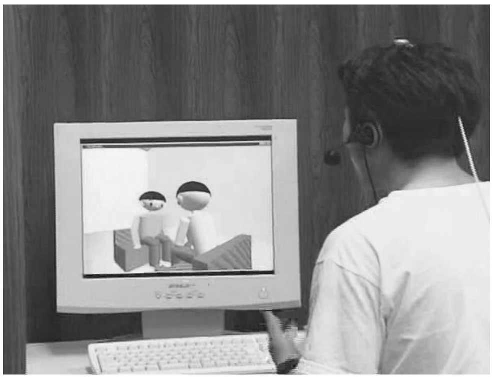


Figure 5: Real-time transformation of conversational gestures

6 Multi-party Interaction: Mission Rehearsal Exercise

Another example, where the starting point is the virtual world inhabited by autonomous agents, is the Mission Rehearsal Example (MRE) environment (Traum and Rickel 2001) developed at the Institute for Creative Technologies.

This training environment allows immersive participation in multi-party interaction. In this system there are autonomous agents in a virtual world that are able to interact with a human visitor (in this case, a trainee that has to perform a certain task) that is immersed in the environment. There is direct interaction (the trainee addresses a particular agent he sees in the environment) and indirect interaction (the embodied agents in the environment have their own tasks, not everybody is always involved in every interaction). See Figure 6. Hence, we have multimodal interaction between multiple (human and virtual) agents in the environment. Important are the locations of the conversants and the objects they are discussing. Agents are aware that others are listening. An important aspect of this system is the underlying dialogue model. It consists of several layers: a contact layer (whether and how individuals are accessible for communication), an attention layer (the objects or process that agents attend to), the conversation layer (where separate dialogue episodes are modeled), a layer of social commitments and a layer of negotiation (how agents come to agree on commitments). Although the models are there it is certainly not yet the case that in this environment there is free interaction between the multiple (virtual and human) agents. Currently the layered model underlies a scripted interaction.

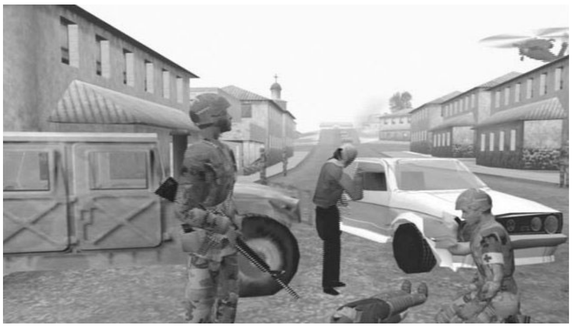


Figure 6: Multi-party interaction in the Mission Rehearsal Exercise

A similar environment for learning Lebanese Arabic language and culture is being developed

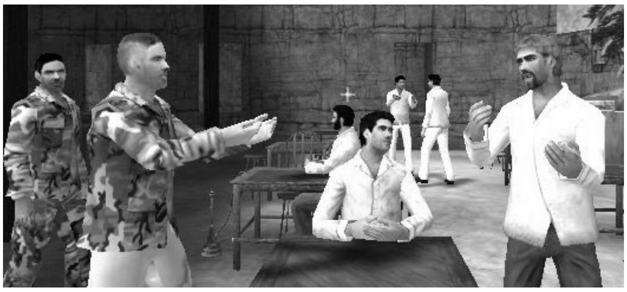


Figure 7: Tactical Language Training project

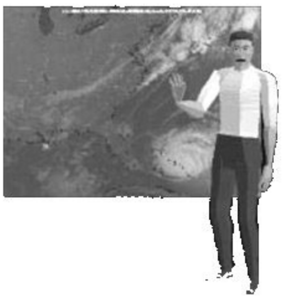


Figure 8: Virtual presenter

at the CARTE institute. The environment is inhabited by animated agents representing local people with who a learner has to communicate (see Figure 7). The learner is also represented in the environment where his avatar displays the chosen gestures.

6.1 The HMI-Parlevink Virtual Meeting Environment

The AMI project just started. On the other hand, our research group has a background in modeling embodied agents in their 'natural' environments. Some preliminary research on modeling meeting behavior and displaying it in a virtual meeting room is under way. One of the topics we look at this moment is the role of a virtual presenter in a virtual meeting room. Previous work in this area has been done by Nomay et al. 2000 (see Figure 8). One thing we would like to model is to have a remote participant showing a presentation as an embodied agent. It certainly should allow interaction with this embodied representation and probably also with the remote participant who is (semi-)controlling the presentation. However, in our situation we also want to allow fully synthetic presenters that know about the presentation and that are able to interact with meeting participants (maybe present in a physical meeting room, maybe a remote human meeting participant and maybe a fully synthetic virtual assistant). Can we interrupt this synthetic presenter while he or she is showing a PowerPoint presentation? The presenter knows about all sheets in the presentation and should at least be able to tell that the answer to a particular question will be on a next sheet. Or that he or she has already handled that, but is willing to go back to a particular sheet in order to explain it again or in more detail.

Apart from increasing the notion of (real-time) presence, when we combine virtual, real, mixed and augmented meeting settings, there is also the notion of validation of theories of meeting interactions by looking at possibilities to generate such interaction behavior from models of interaction or from (semi-)automatically obtained annotations from meeting interactions.

6.2 Putting it All Together

In this section we made clear that some modest research attempts are underway to achieve models that cover verbal and nonverbal communication aspects of human behavior in different situations. These models are necessary to allow for a smooth transition from real to virtual worlds and to a merging from real and virtual worlds. Due to our participation in European projects on meeting modeling, meeting situations and meeting interactions our main efforts are in the area of meetings. However, there are so many different kinds of meetings, meeting situations, meeting interactions and meeting participants that we don't think this domain very much restricts our interest in modeling human interaction in all possible kinds of situations.

7 Conclusions

We discussed different application areas where it has become useful to model multi-party human interaction behavior. Our main observation in this paper is that we see research in previously separate areas converge and that there is a natural trend towards situations where ambient intelligence environments (exemplified in this paper with smart meeting rooms) and virtual reality environments merge in order to obtain shared environments where people live, work and meet. In this paper we surveyed our research and research ideas in the framework of the European AMI (Augmented Multi-party Interaction project). We hardly touched upon our technical work in this project.

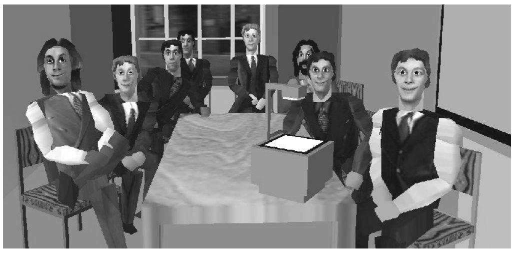


Figure 9: Meeting audience showing appreciation

Apart from meeting modeling (see section 4) we are in particular concerned with the design of annotation tools, image processing (posture, gesture and facial expressions), modeling of turntaking and addressee detection and emotion modeling, all in the context of meetings in smart environments. There is a lot of research that is extremely important, but is not discussed here and not part of the project. We would like to mention privacy issues, presence issues and issues related to the fact that people know that there actions are recorded and interpreted (cf. Nijholt 2004). Presence issues in a meeting environment have been researched by Slater (Pertaub et al). Slater studied the illusion of sentience in a virtual meeting environment with the objective to present evidence that people react to virtual characters as if they were real. See Figure 9 where someone is presenting for an interested (virtual) audience. Obviously, these observations are interesting when we allow mixtures of virtual and (representations of) real people in the same meeting environment.

References

Aarts, E., R. Collier, E. van Loenen & B. de Ruyter (Eds.). Ambient Intelligence. Proceedings First European Symposium, EUSAI 2003, Lecture Notes in Computer Science, Springer, Berlin, 2003.

Bales, R.F. *Social Interaction Systems. Theory and Measurement.* Tranaction Publishers, New Brunswick, 2001.

Ellis, C. and Barthelmess. The Neem dream. Proceedings *Tapia '03*, October 2003, Atlanta, Georgia, USA, 23-29.

Goffman, E. *Behavior in Public Spaces.* Notes on the Social Organization of Gatherings. The Free Press, New York, 1963.

Hofs, D., R. op den Akker & A. Nijholt. A generic architecture and dialogue model for multimodal interaction. Proc. 1st Nordic Symposium on Multimodal Communication, P. Paggio, K. Jokinen & A. Jönsson (Eds.), CST Publication, Center for Sprokteknologi, Copenhagen, 2003, 79-92.

Jovanovic, N.. Recognition of meeting actions using information obtained from different modalities: a semantic approach. TR-CTIT-03-48, October 2003, 44 pp.

McCowan. I, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner and H. Bourlard. Modeling Human Interaction in Meetings. Proc. *IEEE ICASSP 2003*, Hong Kong.

Mikic, I., K. Huang & M. Trivedi. Activity monitoring and summarization for an intelligent meeting room. In: Proceedings IEEE Workshop on Human Motion, Austin, Texas, December 2000.

Morgan, N., D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg & A. Stolcke. The Meeting Project at ICSI. Human Language Technologies Conference, San Diego, March 2001.

Nijholt, A. Towards virtual communities on the Web: Actors and audience. Proc. *Intelligent Systems & Applications (ISA'2000)*, Vol. II, F. Naghdy et al. (Eds.), ICSC Academic Press, Canada, 2000, 725-731.

Nijholt, A. Multimodality and Ambient Intelligence. In: *Algorithms in Ambient Intelligence.* W.F.J. Verhaegh, E.H.L. Aarts & J. Korst (Eds.), Kluwer Academic Publishers, Boston/-Dordrecht/London, 2003, 21-53.

Nijholt, A., T. Rist & K. Tuinenbreijer. Lost in ambient intelligence? In: Proceedings *ACM Conference on Computer Human Interaction (CHI 2004)*, April 2004, Vienna, Austria, ACM New York, 1725-1726.

Nijholt, A. Where computers disappear, virtual humans appear. *Computers and Graphics*, Vol. 28, No. 4, Elsevier, ISSN 0097-8493, 2004, to appear.

Nomay, Ts., L. Zhaoz & N.I. Badler. Design of a Virtual Human Presenter. Internal report, University of Pennsylvania, 2000.

Pertaub D.-P, M. Slater & C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments* 11 (1), 68-78.

Schultz, T., A. Waibel, M. Bett, F. Metze, Y. Pan, K. Ries, T. Schaaf, H. Soltau, M. Westphal, Hua Yu & K. Zechner. The ISL Meeting Room System. Proceedings of the Workshop on Hands-Free Speech Communication (HSC-2001), Kyoto Japan, April 2001.

Stiefelhagen, R. Tracking focus of attention in meetings. Proc. IEEE *International Conference on Multimodal Interfaces*, Pittsburgh, PA, USA, 2002, 273-280.

Traum, D. and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Agents 2001 Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents.*

Trivedi, M., I. Mikic, S. Bhonsle. Active Camera Networks and Semantic Event Databases for Intelligent Environments. IEEE Workshop on Human Modeling, Analysis and Synthesis (in conjunction with CVPR), Hilton Head, South Carolina, June 2000.

Vertegaal, R. Look who's talking to whom. Mediating joint attention in multiparty communication & collaboration. Ph.D. Thesis, University of Twente, 1998.

Vilhjálmsson H., and J. Cassell. BodyChat: Autonomous Communicative Behaviors in Avatars. In: Proc. *2nd Annual ACM International Conference on Autonomous Agents*, Minneapolis, 1998.

Vilhjálmsson, H. Avatar Augmented Online Conversation, Ph.D. dissertation, Program in Media Arts and Sciences, Massachusetts Institute of Technology, Cambridge, MA.

Zobl, M., F. Wallhoff & G. Rigoll. Action recognition in meeting scenarios using global motion features. Proc. IEEE International Workshop on *Performance Evaluation of Tracking and Surveillanc*e, 2003.