Meetings, Gatherings, and Events in Smart Environments

Anton Nijholt
Centre of Telematics and Information Technology (CTIT)
University of Twente, PO Box 217
7500 AE Enschede, the Netherlands
anijholt@cs.utwente.nl

Abstract

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. Our aim is to research representations of what takes place during meetings in order to allow generation, e.g. in virtual reality, of meeting activities. This allows us to look at tools that provide support during a meeting and 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 differences between real, human-controlled, and (semi-) autonomous virtual participants disappear.

1. Introduction

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 are willing to enter video-conferencing and computer-supported collaborative environments. Rather than obliging 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. Meeting knowledge 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 helps us to design advanced and attractive meeting environments.

In this paper the more general point of view is that of ambient intelligence (AmI). 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 AmI 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 AmI. The environment uses knowledge about our emotions, about our moods and about our personality when it provides support. It should be possible to induce social relationships between the AmI environment and its inhabitants.

In the next section we introduce our view on AmI and real and virtual humans. Section 3 is devoted to a discussion on the European projects in which we are involved and that guided our insights in AmI issues and the VR continuum. In section 4 we zoom in on the VR continuum, considering meeting situations. We generalize from observations in the context of smart meeting environments to acts and events in AmI environments.

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 AmI environments.

2.1 Interpretation of Events and Activities

We need to model social and intelligent interactions in an environment between humans, between humans and objects, between humans and autonomous embodied agents (virtual humans) and with the environment (not necessarily addressing an object or human in particular). Input can be obtained from sensors for sound, image, and haptics. The interaction does not only include focused interaction, but also aspects of unfocused interaction. Interpretation requires the fusion of all modalities into 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 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 Agents

Autonomous embodied agents can be part of an AmI 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 related to privacy questions. Who has access to this information and who owns the AmI 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 may have a negative impact on natural behavior of inhabitants and visitors and therefore will have negative consequences for the performance of the environments. Due to eyes and ears, present in objects and walls as electronic dust, we may ask whether being the sole inhabitant of an environment is 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 unfocused interaction behavior.

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 VR environment [Nijholt 2000] or in the context of presence, alienation and privacy [Nijholt et al. 2004]. Involvement in European projects on meeting environments has been fruitful to develop the ideas further. We will discuss these projects first.

3. Meetings: Signal Processing and 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 5th 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.

Events and interactions 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 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 facial expressions, gestures and body posture that support, emphasize or contradict their opinion.

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. Scripted meetings have been organized in which participants act according to prescribed rules that define periods of monologue, discussion, note taking, or presentation. The corpus thus obtained allows study of meeting participants' behavior. Meetings are captured by cameras, lapel microphones and microphone arrays. Recently white board pen capture has been 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

[1] Cf. Michael Coen from MIT about the effects of smart environments on their inhabitants: "The notion of being alone may disappear, .." 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 (2002-2005) is supported by a EU IST programme on Multimodal and Multisensorial Dialogue Modes.

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.

Two approaches 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. Annotation tools and mark-up languages are developed that allow the description of the relevant issues during a meeting, including temporal aspects and low-level fusion of media streams. Higher-level fusion, where 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.

The M4 meeting manager captures the events and interactions in the meeting room. After capturing, the collected information becomes available for participants and non-participants. Clearly, we can look at the project as research on smart environments and on AmI. However, there is no explicit or active communication between user and environment. The user does not explicitly address the environment. Currently, the environment registers and interprets what's going on, but is not actively involved. The environment is attentive, but it is not pro-active. Real-time participation of the environment requires attention and interpretation, but also intelligent feedback and pro-active behavior. It requires also presentation by the environment of multimedia information to the occupants of the environment.

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. It 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 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. We introduce the different research tracks that have been defined. From the point of view of the VR continuum the following tracks are especially relevant:

- **Understanding Meetings:** Which meeting characteristics play a role in order to understand the multi-party interaction? Turn taking dynamics and multi-party interaction modeling are general areas of research. How do they depend on characteristics as size, status differences, familiarity with each other, the setting, the goal or task (maintaining sociality, sharing information, generating ideas)? Other modes than face-to-face interaction are considered, e.g., asynchronous communication or video-conferencing. The environment needs to understand in order to allow real-time support or off-line access.
- **Uni- and Multi-modal Recognition:** 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. Multimodal syntactic and semantic information need to be extracted in order to recognize and interpret participant behavior, participant interaction and meeting events.
- **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. Example questions that need to be answered are: Who were the participants? Was the agenda covered? How did the discussion progress? What was the atmosphere? Give a summary. Segmentation of a meeting can be done from different viewpoints. We can look at events such as discussion, monologue, presentation, but also at a structuring in terms of decision points, task assignments and topic shifts.
- **Remote meeting assistant:** One of the issues 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 exist other projects on the computational modeling of meetings and the development of tools that help to support meetings or to off-line review and retrieve information. Close to M4 and AMI is the Meeting Room project [Schultz et al. 2001]. Also related is the work done at UCSD on the design of methods for person identification, speaker recognition, face orientation and semantic activity processing. There is both work on smart meeting rooms [Mikic et al. 2000] as on smart environments in general [Trivedi et al. 2000]. The Ambiance project [Aarts et al. 2003] looks at smart home environments, requiring modeling of the home and the objects that can play a role when inhabitants interact with each other and the environment.

4. Meetings in a Virtual Reality Continuum

Developments in AmI 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

[3] AMI started in 2004 and has duration of 3 years. It is supported by the EU 6th FP IST Programme.

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 turn taking process. Again, in AmI 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. This information can be used to generate VR representations of meeting participants in a VR room or an augmented reality supported physical meeting room. Meeting participants can be physically present, they can be represented by an agent that alerts and supports when things become interesting - but otherwise is rather passive - or they can be immersed in the (distributed) VR environment together with other participants, represented as avatars mimicking their owners. In the subsections below we show two examples.

4.1 Multi-party Interaction: BodyChat

In VR environments some research on multi-party interaction can be found. Vilhjálmsson and Cassell [1998] introduced BodyChat, 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 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. Translation of this work to a 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 and add remote participants or add model-based behavior to virtually represented participants. E.g., 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 peculiarities. This view allows a participant to become more lively using more extrovert gestures and facial expressions, it allows to convert a non-native speaker to a native speaker and it allows to change the physical appearance of a participant.

4.2 Multi-party Interaction: MRE

In the Mission Rehearsal Exercise (MRE) environment [Traum and Rickel 2001] we have a VR world inhabited by autonomous agents. The environment allows immersive participation in multi-party interaction. (Semi-) autonomous agents know how to interact with a trainee immersed in the environment. There is direct interaction (the trainee addresses a particular agent in the environment) and indirect interaction (the embodied agents in the environment have their own tasks, not everybody is always involved in every interaction). 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 not yet the case that there is free interaction between the multiple (virtual and human) agents. Currently the layered model underlies a scripted interaction.

4.3 Putting it All Together

We made clear that 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.

5. Conclusions

We discussed different application areas where it is useful to model human interaction behavior. Our main observation is that research in previously separate areas converges and that there is a natural trend towards situations where AmI 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.

References

AARTS, E., COLLIER, R., VAN LOENEN, E., AND DE RUYTER, B. (Eds.). 2003. Ambient Intelligence. Proc. First European Symposium, LNCS, Springer, Berlin.

McCOWAN, I., BENGIO, S. GATICA-PEREZ, D., LATHOUD, G., MONAY, F., MOORE, D., WELLNER, P., AND BOURLARD, H. 2003. Modeling Human Interaction in Meetings. Proc. *ICASSP*, Hong Kong.

MIKIC, I., HUANG, K. AND TRIVEDI, M. 2000. Activity monitoring and summarization for an intelligent meeting room. In: Proc. IEEE Workshop on Human Motion, Austin, Texas.

NIJHOLT, A. 2000. 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., RIST, T., AND TUINENBREIJER, K. 2004. Lost in ambient intelligence? In: Proc. *ACM Conference on Computer Human Interaction (CHI 2004)*, Vienna, Austria.

SCHULTZ, T., WAIBEL, A., BETT, M., METZE, F., PAN, Y., RIES, K., SCHAAF, T., SOLTAU, H., M. WESTPHAL, HUA YU, AND ZECHNER, K. 2001. The ISL Meeting Room System. Workshop on Hands-Free Speech Communication (HSC-2001), Kyoto Japan.

STIEFELHAGEN, R. 2002. Tracking focus of attention in meetings. Proc. IEEE *Conf. on Multimodal Interfaces*, Pittsburgh, PA, 273-280.

TRAUM, D., AND RICKEL, J. 2001. 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., MIKIC, I., AND BHONSLE, S. 2000. Active Camera Networks and Semantic Event Databases for Intelligent Environments. IEEE Workshop on Human Modeling, Analysis and Synthesis (in conjunction with CVPR), Hilton Head, SC.

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