Towards Simulating Humans in Augmented Multi-party Interaction

Anton Nijholt

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

Abstract

Human-computer interaction requires modeling of the user. A user profile typically contains preferences, interests, characteristics, and interaction behavior. However, in its multimodal interaction with a smart environment the user displays characteristics that show how the user, not necessarily consciously, verbally and nonverbally provides the smart environment with useful input and feedback. Especially in ambient intelligence environments we encounter situations where the environment supports interaction between the environment, smart objects (e.g., mobile robots, smart furniture) and human participants in the environment. Therefore it is useful for the profile to contain a physical representation of the user obtained by multi-modal capturing techniques. We discuss the modeling and simulation of interacting participants in the European AMI research project.

1. Introduction

Human-computer interaction requires modeling of the user in the interface. User modeling has become a well-respected research area and knowledge about the user makes it possible for a system to adapt its behavior towards the user, e.g. by predicting the user's behavior and preferences and anticipating on this behavior and preferences. There is a tendency to collect as much information of a user as possible. A user profile typically contains preferences, interests, characteristics, and interaction behavior. During the interaction with a system a user displays behavior and makes decisions that can be used to modify a profile. During the interaction it is however more important that the system knows about details of the needs of the user at that particular moment than the global information that is available in a user profile.

During multimodal interaction a system has the possibility, using multiple sensors, to capture real-time the changing characteristics of the user and its way of interacting. This may include facial expressions, gestures, intonation, body posture and biometric information. Fusion and interpretation of that information will make it possible to decide whether a user is satisfied or frustrated about what is going on in the interaction. We have a real-time modeling of the user.

It is certainly not the case that for all human-computer interaction this real-time modeling of the user is required and useful. On the other hand, there are applications for which we need to go several steps further. In smart environments or ambient intelligence environments we encounter situations where the computerized environment has to support interaction between the environment, smart objects (e.g., mobile robots, smart furniture) and human visitors or inhabitants of the environment.

This situation is not really different from a situation where users become part of an augmented reality or virtual reality environment and the environment needs to know about or be able to capture movements and body properties of a user of that environment. Since we are talking about multiple interacting human users or visitors of these interaction supporting environments the question is how to represent these users of such environments. The user profile may contain a physical representation of the user and multi-modal capturing techniques may add in real-time dynamic changes (movements, facial expressions, posture shifts, gestures, etc,). Obviously, the need to present this information to other users in the environment is higher in a situation where users share a virtual environment and one or more of them are not physically present, than in a situation where they share the same physical environment.

In this paper we discuss the modeling and simulation of interacting participants in a smart meeting environment. We report about research that is performed in the context of the European AMI (Augmented Multi-party Interaction) project. Modeling of verbal and nonverbal interactions between meeting participants and having the smart meeting environment understand and support this interaction is the main aim of this project. In our research we do not only look at capturing, interpretation, translation and manipulation of meeting interaction behavior of participants, but also at the possibility to introduce virtual agents that can play particular and useful roles during a meeting, e.g., a virtual chairperson or a virtual meeting participant that has been sent to represent its owner during the meeting.

The organization of this paper is as follows. In section 2 we have some general observations on extensions of, more or less, traditional ways of user modeling. That is, we look at users – or rather visitors, partners, collaborators, colleagues, inhabitants, etc. – acting in smart and virtual environments for which it is useful to include in a profile properties dealing with location preferences and behavior, properties dealing with physical (appearance) and other observable characteristics of verbal and nonverbal behavior. In section 3 we zoom in on meeting behavior and the research that is performed in the European AMI project. Section 4 shows our application of virtual and distributed virtual meeting rooms where meeting participants are represented by virtual humans. Section 5 of this paper contains conclusions and observations on future work.

2. Modeling partners, participants and inhabitants

User profiles allow computer users to be presented with personalized applications. Typically a profile contains preferences, interests, characteristics and behavior. Much more can be added, but in traditional human-computer interaction there is not always a need to process that information. When the system the user interacts with allows multimodality then more information about the user can be extracted in real-time. For example, the system may learn about interaction pattern preferences [1] or detect the user's emotional state and adapt its interaction behavior, its interface and its feedback according to them. The body and what the user is doing with his or her body is becoming important for the system and this is even more the case when the user is allowed to move around and interact from different positions and with various objects, maybe other users and parts of a computer-supported or monitored environment. We not only have users, but also inhabitants, players, partners and passers-by. Not only they need to be characterized, but they need to be characterized in their physical context from information obtained from sensors in the environment and its objects (location sensors, cameras, tracking systems, microphones) including wearables, portable devices and active and passive tags attached to the users. Rather than interaction histories these perceptual technologies allow us to build up and exploit context histories [2].

In ambient intelligence research the aim is to model verbal and nonverbal communication and other human behavior in such a way that the environment in which this communication and other behavior takes place is able to support these human activities in a natural way. Obviously, the purposes of the environments and the aims of the inhabitants of a particular environment can very much constrain and guide the interpretation of the activities and the support given by the environment. Entertainment, education, profession, home, family, friends, etc., all provide different viewpoints on activities,

communications, and desirable real-time support and sometimes also on off-line support allowing intelligent access to archived activities and multi-media presentation of such information.

3. Modeling meeting environments, activities and interactions

The meeting support application researched in the AMI project [3] requires the development of tools and methods that take into account the particular meeting context. Rather than zooming in on constraining general methods of detecting and interpreting events in physical environments, we have a bottom-up approach starting with observed events in meeting environments and attempting to model and explain them using more general observations on theories of verbal and nonverbal communication display. The research issues in the AMI project are:

- **Understanding Meetings:** Which meeting characteristics play a role in order to understand the group's communication? Multimodal turn-taking dynamics and multi-party interaction modeling are general areas of research. How do turn-taking 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. Meeting support research is also about the environment understanding the meeting in order to allow intelligent off-line access for retrieval, summarization, 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.
- **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. Segmentation of a meeting can be done from different viewpoints (discussion, monologue, note taking, presentation, decision points, task assignments and topic shifts). A meeting browser can be designed that uses a hypertext view of the meeting in which the different viewpoints are embedded.
- **Remote meeting assistant:** One of the issues that is 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 browse recent events in the meeting or to be automatically alerted at points of interest.

4. Meeting participants in (distributed) virtual meeting rooms

In our research we have looked at capturing meeting activities from an image processing point of view and at capturing meeting activities from a higher-level point of view, that is, a point of view that allows,

among others, observations about dominance, focus of attention, addressee identification, and emotion display. We studied posture and gesture activity, using our vision software package. Our flock-of-birds software package was used to track head orientation of some of our 4-party meetings. It allowed us to display animated representations of meeting participants in a (3D) virtual reality environment [4]. In this environment visualized events can be augmented with meta-observations provided by support agents and displayed in the virtual environment. This is illustrated in Figure 1.

Even more attractive is it to have meetings represented in a virtual meeting room (VMR), where the participants do not all share the same physical space. We introduced a prototype version of a distributed meeting room set-up

This set-up [5], illustrated in Figure 2, allows the connection of several inhabited meeting rooms equipped with standard AMI sensors (cameras, microphones) and the representation of the meeting participants and their activities in a shared virtual meeting room that is made accessible for all participants (and possible observers) in real-time. The set-up allows the participants to take part in the meeting, perceiving the verbal and nonverbal communication by other participants

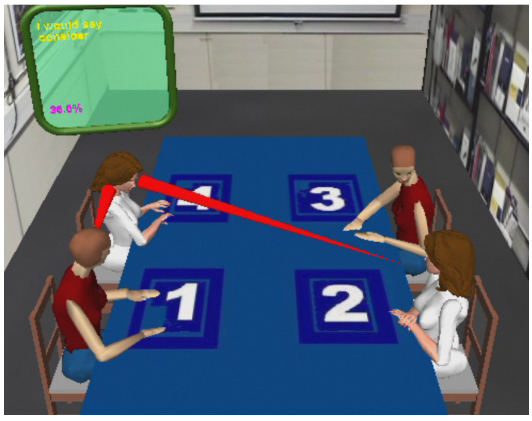

Figure 1. The virtual meeting room showing gestures, head movements, speech transcript, addressee(s) and the percentage a person has spoken until that moment

through their avatars, from their assigned position around the meeting table. As shown in Figure 2, also in this distributed version we can add meta-information about the meeting and its progress to the visualization of the virtual meeting room.

The technology used within the DVMR experiment differs substantially from normal video conferencing technology. Rather than sending video data as such, this data is transformed in a format that enables analysis and transformation. For the DVMR experiment the focus was on representing poses and gestures, rather than, for example, facial expressions. Poses of the human body are easily represented in the form of skeleton poses [6], essentially in the same format as being used for applications in the field of virtual reality and computer games. Such skeleton poses are also more appropriate as input data for classification algorithms for gestures.

Another advantage for remote meetings, especially when relying on small handheld devices, using wireless connections, is that communicating skeleton data requires substantially less bandwidth than video data. A more abstract representation of human body data is also vital for combining different input channels, possibly using different input modalities. Here we rely on two different input modalities: one for body posture estimation based upon a video camera, and a second input channel using a head tracker device. Although the image recognition data for body postures also makes some estimation of the head position, it turned out that using a separate head tracker was much more reliable in this case.

The general conclusion is, not so much that everyone should use a head tracker device, but rather that the setup as a whole should be capable of fusing a wide variety of input modalities. This will allow one to adapt to a lot of different and often difficult situations. In the long run, we expect to see two types of environment for remote meetings: specialized meeting rooms, fully

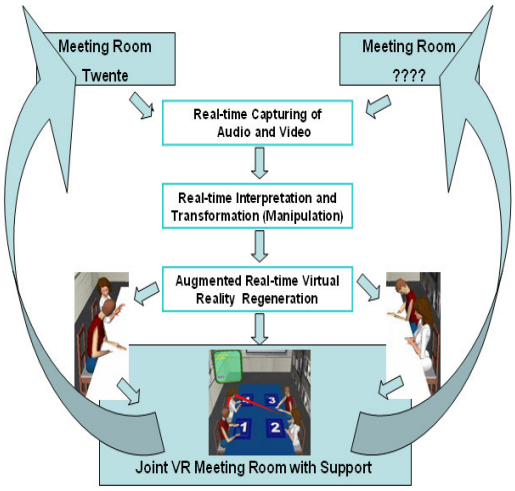

Figure 2. Capturing, manipulation and re-generation of activities in remote locations

equipped with whatever hardware is needed and available for meetings on the one hand side, and far more basic single user environments based upon equipment that happens to be available. The capability to exploit whatever equipment is available might be an important factor for the acceptance of the technology. In this respect, we expect a lot from improved speech recognition and especially from natural language analysis. The current version of the virtual meeting room requires manual control, using classical input devices like keyboard or mouse, in order to look around, interact with objects etcetera. It seems unlikely that in a more realistic setting people that are participating in a real meeting would like to do that. Simpler interaction, based upon gaze detection but also on speech recognition should replace this situation.

5. Conclusions and future work

We surveyed our research towards representing meeting participants as virtual humans in (distributed) virtual meeting environments. The characteristics assigned to these virtual humans hardly include properties that can be displayed in individual and (semi-)autonomous behavior of the avatars. This will be researched in the future. That is, how can we use the information that is available or has been collected during interaction and interaction histories to provide an avatar representing a human meeting participant with realistic behavior and, when useful, realistic autonomous behavior?

Acknowledgements

Members of our group involved in the research reported here are, among others, Job Zwiers, Rutger Rienks, Dennis Reidsma, Ronald Poppe and Hendri Hondorp. We also should acknowledge Jan Peciva from the Technical University of Brno who collaborated with us in order to realize the distributed virtual meeting room.

This work was partly supported by the EU 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-132).

References

[1] Oviatt, S., Lunsford, R. & Coulston, R. Individual differences in multimodal integration patterns: what are they and why do they exist? *CHI '05*: Proc. of the 2005 Conf. on Human Factors in Computing Systems, 2005, 241-249.

[2] Prante et al. ECHISE 2005, 1st International Workshop on *Exploiting Context Histories in Smart Environments*. Workshop at PERVASIVE 2005, Munich, Germany, 2005.

[3] McCowan, I., Gatica-Perez, D., Bengio, S., Moore, D., Bourlard, H. Towards Computer Understanding of Human Interactions. In: *Ambient Intelligence*, E. Aarts et al. (Eds.), LNCS, Springer-Verlag Heidelberg, 235 - 251.

[4] Nijholt, A. Meetings in the Virtuality Continuum: Send Your Avatar. In *Proceedings Cyberworlds 2005*, A. Sourin (Ed.), November 2005, Singapore, to appear.

[5] Nijholt, J. Zwiers & J. Peciva. The Distributed Virtual Meeting Room Exercise. In: Proceedings *ICMI 2005 Workshop on Multimodal multiparty meeting processing*, Trento, Italy, October 2005, to appear.

[6] Poppe, R., Heylen, D., Nijholt, A., Poel, M. Towards real-time body pose estimation for presenters in meeting environments. *Proc. 13th Intern. Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*. V. Skala (Ed.), Plzen, Czech Republic, 2005, 41-44.