Visual Comput (2006) 22: 965–976
DOI 10.1007/s00371-006-0041-3

ORIGINAL ARTICLE

Anton Nijholt
Rutger Rienks
Job Zwiers
Dennis Reidsma

Online and off-line visualization of meeting information and meeting support

Published online: 5 August 2006

A. Nijholt (✉) · R. Rienks · J. Zwiers · D. Reidsma
Human Media Interaction Lab, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
{anijholt, rienks, zwiers, dennisr}@ewi.utwente.nl

Abstract In current meeting research we see modest attempts to visualize the information that has been obtained by either capturing and, probably more importantly, by interpreting the activities that take place during a meeting. The meetings being considered take place in smart meeting rooms. Cameras, microphones and other sensors capture meeting activities. Captured information can be stored and retrieved. Captured information can also be manipulated and in turn displayed on different media. We survey our research in this area, look at issues that deal with turn-taking and gaze behavior of meeting participants, influence and talkativeness, and virtual embodied representations of meeting participants. We stress that this information is interesting not only for real-time meeting support, but also for remote participants and off-line consultation of meeting information.

Keywords Smart meeting room · Multi-modal interaction · Virtual meeting room · Multi-modal corpus · Corpus annotation · 3D reconstruction

1 Introduction

In this paper, we survey our research on verbal and non-verbal interactions between meeting participants and ways to visualize the information that is obtained from capturing meeting activities in smart meeting environments. This implies a need for theories, models and algorithms that allow us to describe multi-party interaction from which we can design tools and add smartness into these environments. To be able do this we have to deal with the following issues: (i) interpretation of the events and activities in the observed environment; (ii) reactive and pro-active real-time support for activities in the observed environment, for example provision of information about the meeting process or about the meeting topics; (iii) online (during the meeting) and off-line multimedia retrieval, reporting (filtering), browsing and other means of access to meeting information; (iv) online remote observation of activities (monitoring) and remote participation in activities; and, (v) owning and controlling the environment (and its inhabitants).

In more familiar terms, we would like to have technological support during our meeting activities; we want to be able to review and retrieve what occurred during a meeting. We cannot always attend a meeting but want to be able to browse through a record of it, and we want to remotely participate in a meeting or just monitor a meeting when there is no possibility of being physically present. The last issue, owning and controlling the environment is, of course, extremely important from the point of view of controlling access, information, privacy and maintaining a smart environment. There is no strict requirement here to confine ourselves to (smart) meeting rooms. Points of view and technology to be developed have also been applied to smart office environments [27], to educational environments, to home environments [5], and to public spaces. In all these environments we can ask for real-time

support, retrieval and off-line browsing of previous activities and remote participation or monitoring activities.

We pursue our line of research from the viewpoint of what we call the smart meeting room paradigm. In this smart meeting room paradigm, research activities in several fields of research converge. Firstly, research in the area of smart environments and ambient intelligence. That is, research on sensor-equipped environments, where the events and activities detected by the sensors are interpreted by fusing the signals coming from different sources and after interpretation are translated into intelligent support for events and activities taking place in the environment. This support may affect the behavior and characteristics of the inhabitants of the environment. By inhabitants we refer to smart objects, mobile robots, virtual (embodied) agents and also humans.

Secondly, research in the area of multi-modal and multi-party interaction. In traditional human-computer interaction we have mouse, keyboard and a graphical user interface using a single desktop computer. The modest progress in speech recognition allows for interaction through voice. In gaming and some professional environments, we also use other input modalities, for example joysticks, haptic devices, data gloves, input that can be captured using a camera (pointing gestures, facial expressions, and body movements) or even sensors that capture biometrical information from the user. When we change our viewpoint to that of environments where information about users can be captured in different ways, where their activities are not restricted to one particular desktop, and where their activities also include explicit and implicit communication with other entities (a human, a smart device, a virtual agent), then both natural verbal and nonverbal communication modeling and communication between humans and other human inhabitants and available non-human inhabitants become important. That is, we need to model multi-party interaction.

Thirdly, we need to mention developments in the area of teleconferencing and collaborative virtual environments. Clearly, teleconferencing has a long tradition [24]. Current teleconferencing systems are biased towards transmitting video and hardly consider other ways of transmitting participantsŠ contributions. There are exceptions: for example, there were early attempts to use virtual reality environments where avatars represent meeting participants, where shared blackboards can be accessed and where other meeting equipment is visualized [10]. Neither image processing, artificial intelligence, animation, virtual reality nor information visualization have played important roles in teleconferencing, although there have been attempts to include tools to convey gaze information to the participants [26]. The situation is slightly different when we look at collaborative environments that developed from computer supported collaborative work (CSCW) systems. Here, from the beginning, research issues were more advanced since workers are assumed to collaborate in non-verbal ways (sharing notes, sharing objects), and therefore it is an advantage to have their actions made visible to each other and to have a virtual environment designed to support these activities [8, 9]. One of the aims of our research is to allow real-time transformation of detected events from one world into another (virtual) world. To realize this aim we regenerate observations from one meeting room into another (remote) meeting room. For this purpose, we created a virtual version of a smart meeting room, where apart from visualizing captured meeting activities we are also able to look at visualizing meta-information about the meeting, as well as augmenting it with all sorts of supportive meeting assistants. Real-time support in smart meeting rooms has been discussed before, e.g. in [2, 13].

The remainder of this paper is organized as follows. Since our research takes place in the context of the European Augmented Multi-party Interaction (AMI) project, the project will be discussed in Sect. 2. The emphasis in the project is on the design of a corpus of meeting interactions and tools to annotate this corpus in order to analyze the data and to achieve models that underlie the data. Section 3 is devoted to our attempts to introduce off-line and online support to meeting activities. In order to provide support, the activities in the smart meeting environment need to be interpreted, and pro-active and reactive support will be based on access to this interpretation. A virtual reality representation of meeting activities enriched with useful meta-information is a very advanced form of interpretation. In Sect. 4, we discuss how we can obtain such a representation from captured data (using cameras and microphones) and we also discuss how it can play a useful role: for research purposes, for browsing meeting information and for remote participation, that is, a desktop, augmented reality or an immersive virtual reality environment where participants meet and participate and where meeting information can be displayed to participants and visitors. As a preliminary example of a virtual (embodied) meeting assistant, we discuss some work in progress on a virtual presenter. In Sect. 5, we reveal some preliminary observations on an initial distributed version of our virtual meeting environment. Conclusions and some directions for future work are mentioned in Sect. 6.

2 The augmented multi-party interaction (AMI) project

Well-known projects on multi-modal interaction have been created to bridge research on smart environment and ambient intelligence research on the one hand and meeting or teleconferencing research on the other hand. As explained, our work takes place within one of these projects: the AMI [14, 17] project. AMI is a research project in the European 6th Framework program and concerns research on multi-modal interaction, and, as the name suggests, multi-modal interaction in a multi-party context; a context

where we have two or more persons interacting with each other and/or with smart entities (objects, virtual humans, robots, etc.) present in the environment. The AMI project concentrates on multi-party interaction during meetings and its main aims, therefore, are to develop technologies for the disclosure of meeting content and the provision of live meeting support of (possibly remote) meetings. This kind of research fits into paradigms of interactions in smart environments, ubiquitous computing, disappearing computers, and ambient intelligence.

Obviously, there is multiparty interaction in educational settings, in class rooms, in offices, in workspaces, in home environments and in public spaces. It is assumed that models and technology developed in the meeting context will be useful in these other situations as well. Other related projects dealing with small group meetings are the late CHIL (computers in the human interaction loop) project [28] and the CALO project (cognitive agent that learns and organizes) [11].

Fig. 1. A global meeting view

2.1 Corpus design and collection

In the AMI project, meeting data is captured in smart meeting rooms at IDIAP (Martigny), at the University of Edinburgh, and at TNO Human Factors (Soesterberg). The meeting data in these meeting rooms is captured using various kinds of sensors: videos, microphones, white-board, PowerPoint presentations and smart pens. The emphasis in the project is on speech and image processing: how to fuse information coming from these two media sources and how to make this information accessible using multimedia presentation tools. The main corpus consists of one hundred hours of video/audio registration of meeting activities, captured from some global video cameras, global microphone array arrangements, individual cameras and (lapel) microphones. In order to collect data and to study meeting activities, meetings have been arranged and recorded.

Scenarios were used to achieve more natural interaction between the participants of these meetings, rather than predefined scripts where participants are explicitly told what to do and how to behave. See Fig. 1 for a global camera view of such a meeting. The scenarios that were used described design meetings of a project team that had to develop a design for a remote control. The participants were assigned various roles: a project manager, a marketing expert, a user interface designer and an industrial designer. The question arises: will these teams ever agree on a design?

2.2 Corpus annotation and corpus annotation tools

How does a development team agree on the design of a remote control? To answer such a question one needs to study the information available in the corpus of collected meetings. To study the corpus, annotation schemes have to be designed, which is an activity that requires interaction between data observed, as well as social psychological models found in the literature. Annotations apply descriptions to data. These descriptions can apply to phenomena at an individual level or at a group level. Moreover, these descriptions should preferably be based on theoretical models (of interaction) or they have been chosen because they are useful for the particular domain of application. Obviously, annotations as well as the theoretical models can describe meeting activity at different levels, for example, names of meeting participants, parts of speech, dialogue acts, gestures that are made, speaker, head and gaze orientation, who is addressed, focus of attention of a particular person or the group, displayed emotions, and the current topic.

Tools based on theoretical models are available to obtain some of these annotations automatically, although not with a hundred percent accuracy. This is especially true for annotations based on models from computational linguistics, dialogue modelling, low-level image processing and multi-modal tracking. Tools for real-time support or off-line information access can be based on this automatically obtained information from a meeting and their sophistication and usefulness directly depend on the state-of-the-art of automatically collecting this information and the intelligence needed to interpret it. In order to advance this state-of-the-art it is an unavoidable task to collect and annotate the data manually. These annotations are useful for data analysis and for the design and evaluation of more advanced theoretical models that describe the issues of interest. Although the annotations have to be performed manually, one can develop tools that allow efficient creation of annotations and tools in which knowledge about the phenomena to be annotated is embedded. This embed-

ding allows tools to suggest annotations, to limit choices, or to pre-fill values of attributes. Among the annotation tools we have developed are a dialogue act coder (DA) and a continuous video labelling tool (CVL, [20]). These tools are based on the NITE XML Toolkit (NXT, [4]), which allows display and analysis of cross annotated multi-modal corpora. In addition, NXT has a query language for the exploitation of annotated data. The CVL tool that we developed supports real-time and off-line annotation of observations and interpretations from video. Examples are gaze direction, head orientation, postures, and target of pointing gestures.

Generally, when looking at annotation creation, it is useful to distinguish between annotations that are based on observations (e.g., head nods) and annotations based on interpretation (e.g., dialogue acts and emotions). It is also useful to distinguish different layers of annotations, where a layer describes one particular aspect of annotating (e.g., gestures). Moreover, a layer can contain input for other layers, for example, a dialogue act layer can use input from a layer describing facial expressions or gaze. Labels for annotating can also be taken from an ontology describing the important concepts of the application. Segmentation of the input into fragments that can be referred to is another important issue. Clearly, all these aspects have to be addressed when our research aims at making a change from manual annotation towards semi-automatic and fully automatic annotation and interpretation of data in order to provide real-time support to the inhabitants of a smart environment. Apart from designing models and rules from the observed and annotated events, there is also the possibility of training and teaching the computer to build models that are able to automatically derive these annotations (that is, automatically find interesting and useful features for interpreting events).

2.3 Meeting modelling

A meeting model captures the knowledge that explains the verbal and nonverbal multi-party interaction in the meeting context. There are individual and group activities that follow from the goals of a meeting and from properties of a group where individuals get together, have different roles, and may have different aims. These higher level phenomena show in relationships between the different features that are annotated and knowledge about them may determine the set of features to be annotated. Knowledge on the level of meeting models helps in predicting what is likely to happen next in a meeting on a more global level. Obviously, attempts can be made to structure meetings based on low-level phenomena. For example, in the AMI project, stochastic models (variants of hidden Markov models) are used to describe meetings as sequences of meeting actions [29]. In such models, many important aspects of a meeting are not addressed. What does a meeting participant communicate or intend to communicate, when are there topic shifts and how can these be recognized, both from verbal and nonverbal cues and who is being addressed by a speaker? Group behavior and how individuals behave in a group setting are important issues. They require input from behavioral and social sciences [1, 15]. Theory and models of social interaction, for example, Bales' theory of social interaction systems need to be translated into interaction issues in order to help us to understand what is going on during a meeting and theories of verbal and nonverbal communication provide input for this understanding. Yet another important issue is the embedding of organizational information models that focus on the decision making process. Our own Twente argument schema [22] is an attempt to crystalize the decision making process through the creation of argument diagrams. Other high-level aspects that need to be taken into consideration are the organizational context of a meeting, the particular context (e.g., a project) of a meeting and the individual context for its participants.

2.4 Real-time, off-line and remote meeting assistance: technology

In order to guide the research in the AMI project, several use cases have been developed. The general assumptions behind these use cases are that there is a need to access available meeting information off-line. The information may be made available through a browsing facility and by asking questions about the meeting (a specific item of information, a summary, some global information, etc.). There is also a need to have real-time support during a meeting, for example, to be able to access information from previous meetings on the same topic or from a previous part of the current meeting. A third aim that needs to be mentioned is the possibility of taking part in a meeting as a remote participant, where the meeting tools allow for the lack of available information (not knowing who plans to take the floor, not knowing who is currently speaking, not really being aware of something interesting that is going to happen, etc.). The context for the use cases are the previously mentioned design team meetings. The use cases that have been introduced are about looking up information on previous meetings, auditing unattended meetings, reminders during meetings on contents of prior meetings and catching up on a meeting you are late for. Similarly, use cases have been developed for remote meeting assistance and for live meeting assistance. A meeting assistant can, for example, alert a participant that he still needs to give an opinion or that the matter under discussion contains some elements that are important for him. A lot of useful information that can be collected during a meeting can be presented to the chairman in order to guide his decisions. An example of a possible meeting assistant is described in work by Joan Dimicco [6]. She

developed a system called second messenger that shows real-time text summaries of participantsŠ contributions. As to whether the system really influenced the meeting, it appeared that after increasing the visibility of the less frequently speaking group members, they started to speak more frequently than before, whereas the more dominant people started to speak up to 15% less. This example shows that meeting assistants are now able, and in the near future will be even more so, to influence and regulate the meeting process. We will return to this meeting assistance in the next sections.

In order to collect and interpret data from a meeting in progress, a lot of multi-modal recognition and interaction technology, based on models of verbal and nonverbal communication, multi-party interaction and group behavior in a meeting context has to be developed. In fact, this technology development takes up the main part of the current AMI project. Since this paper is not about the development of that particular technology, but rather about how to use it in augmented reality meeting environments, we confine ourselves here to mentioning the main research topics and, where necessary for our purposes, to discussing more about the theory and technology in later sections that zoom in on augmented reality meetings. Hence, a short list of topics is presented here:

- Multi-modal source localization, tracking, participant and speaker identification.
- Recognition of speech, gestures, postures, facial expressions, and emotions.
- Fusion (integrating information obtained from different media sources) and fission (selecting information and multi-media for information presentation).
- Automatic identification and modelling of conversational roles (speaker, listener, addressee, audience, etc.). Recognition of individual behavior using verbal and nonverbal cues.
- Detecting and modelling dialogue acts, turn-taking behavior and focus of attention. Detection of argumentative structures in meetings, detection of topics and topic shifts, detection of decision points.
- Segmentation of multi-modal streams, structuring by meeting events, identification of group activity.
- Technology for access to meeting information (retrieval, filtering, browsing, multi-modal summarization, visualization, replay).
- Design of (real-time) meeting support, design of intelligent and pro-active meeting assistants allowing remote and virtual presence, mixed reality and virtual reality tools and meeting environments.

A few additional remarks are in place. First of all, to make this research possible, data has to be collected and environments need to be created (smart meeting rooms) that allow the collection of meeting data. This has been mentioned in Sect. 2.1. For the data that is we designed and developed theory, algorithms and tools that allow for automatic, semi-automatic and manual annotation. This has been discussed in Sect. 2.2. For the development of theories, algorithms and tools it is useful to have a knowledge about meeting processes and group interaction and behavior (Sect. 2.3). Again, as mentioned above, research in these areas is guided by a collection of use cases. Obviously, when translating research results into meeting support technology, we can distinguish different levels of functionality in the technology (e.g. the intelligence of meeting assistants) and, therefore, different levels of intelligence to be obtained from the research that aims at automatic annotation and interpretation. Current state-of-the-art research covers the areas mentioned and also partial results from different areas can be integrated into representations and interpretations on which the development of useful tools for online and off-line meeting support can be based.

3 On and off-line support for interpreting activities

Theory, models and algorithms that describe multi-party interaction allow for the design of tools and environments that support such interaction. We distinguish:

- Reactive and pro-active real-time support to recognized activities in the observed environment, for example by providing information about the meeting process or about the meeting topics.
- Online (during the meeting) and off-line multimedia retrieval, reporting (filtering), browsing and other ways of getting access to meeting information.
- Online remote observation of activities (monitoring) and remote participation in activities.
- Ownership and control of the environment (and its inhabitants)

As mentioned, there is no need to confine ourselves to (smart) meeting rooms. Points of view and technology to be obtained can be applied to smart office environments, to educational environments, to home environments, and to public spaces. Depending on the point of view and the environment, more or less attention can be paid to issues of efficiency, privacy, control, ownership of access and information, trust, presence, well-being, family-feeling, social relationships, entertainment, and education. One other issue that should be mentioned is the role of autonomous and semi-autonomous (embodied) agents. This role will be discussed further in the forthcoming sections. It is not the aim of this paper to discuss all the issues mentioned above. We will focus on providing real-time support to activities in the observed environment and we will look at issues related to online observation and participation in activities.

Participants in activities, whether they are in a joint physical space or participate remotely, can be supported by the environment in their activities and this support can be realized by introducing virtual agents into the environment to act as assistants. For example, meeting assistants may have knowledge about previous meetings or about a current meeting. They can be available for all participants, they can act as a personal assistant to one particular meeting participant, or they can take part, or even all of the responsibility for the meeting organization and outcome. Obviously, the latter should be done in close cooperation with a human chairperson who is responsible for the organization and outcome of the meeting.

To provide this support, we can introduce software agents acting in the meeting environment, collecting and interpreting information that has to be derived from verbal and nonverbal interactions between meeting participants (including interactions with smart objects, the environment, virtual and embodied agents, etc.). Consider, for example, a global agent that acts as a personal assistant to the human chairman; it collects statistics about the talkativeness of the participants, their mutual influence, or other signs of participant involvement and reports these to the chairman. Meeting assistants can also be content assistants that know about finding and presenting related information from previous meetings and other information sources, organizational assistants that take responsibility for planning the meeting (negotiating time and place), meeting preparation (room reservation, preparing data projector and set-up of presentation, advise on time constraints during meeting), and remote control assistants (that take care of automatic slide changes during a presentation, dim the lighting when the presentation starts, etc.). See [2] and [23] for a more elaborate overview.

A distinction between agents that are embodied and agents that show themselves using other representations (prompts, menuŠs, question and answer forms, etc.) will be discussed in the next section. We will return to meeting assistants in section 4.4 where we will discuss them in the context of our research.

4 Meetings in the virtuality continuum

Localization and tracking technology and technologies for recognition and interpretation of verbal and nonverbal meeting activities allow for real-time support of human activities taking place in a meeting environment. Since the technologies do not cover the recognition and interpretation of all aspects of activities we should also understand that support that can be derived from this recognition and interpretation can only be available on a similarly limited level of functionality. This level will increase when theory, models and technology improve.

Real-time support allows transformations from the environmentŠs multi-modal low-level input to representations that can be mapped on multimedia high-level output that is useful for the meeting participants, whether they are present in the smart environment or whether they are participating remotely. During this transformation the information that is received as input can be enriched or even manipulated. Enrichment in this sense could be the employment of (i) historical data containing information about previous activities and interactions, (ii) knowledge obtained from formal models such as ontologies underlying the domains of discussion, (iii) common sense knowledge, and, finally, (iv) knowledge about the participants and their (current) preferences.

One possible representation that can be obtained this way is a virtual reality representation of the meeting and the meeting activities. That is, there is a (preferably real-time) transformation from events and interactions in the physical meeting room to a virtual representation of events and interactions in a virtual meeting room. How to do this and the usefulness of doing so is discussed below. We mention that this idea fits into our earlier observations on sharing physical and virtual spaces and it also fits in with general observations on augmented and mixed reality [16]. When we represent humans in a virtual meeting setting we need to discuss their virtual representatives or replicas (avatars, virtual humans, embodied conversational agents). They can be represented as a copy of themselves, but equally well in any other appearance imaginable. Moreover, just as all their actions and movements can be replicated on their representation, also just a filtered subset or summary of the actions and movements can be replicated, giving the representatives a sort of autonomous behavior. Even more, it will be, and even already is, possible to design embodied agents that have specific tasks in the environment and that communicate with actual persons or replicas of actual persons and act completely autonomously. Before discussing possible roles for virtual humans in virtual meeting environments, we will consider two questions: *Why do we want a virtual representation?* and *How can we obtain a virtual representation?* Once we have discussed these issues, we will turn to roles that can be played by the virtual humans in a virtual meeting environment.

4.1 Why do we want a virtual representation?

There are several reasons for interest in realizing a transformation from meeting activities to their virtual reality representations and in realizing a virtual meeting room (VMR) (see also [19]).

– First of all, this transformation allows a 3D presentation and replay of multimedia information obtained from a captured meeting. Depending on the state-of-the-art of speech and image processing (recognition and interpretation), one may think of manual annotation replay, replay based on both manually and auto-

matically obtained annotations and interpretations and replay based purely on fully automatically obtained interpretations. Obviously, when the meeting environment has the intelligence to interpret the events therein, it can transform events and present them in other useful ways (summaries, answers to queries, replays offering extra information, visualization of meta-information, etc.).

- Secondly, transformed annotations, whether obtained manually or automatically, can be used for the evaluation of annotations and annotation schemes and of the results obtained by, for example, machine learning methods. Applicable models of verbal and nonverbal interaction, multi-party interaction, social interaction, group interaction and, in particular considering our domain of meeting activities, models of meeting behavior on an individual or on a group level, are not available or only available for describing the rather superficial phenomena of group interaction. Our virtual room offers a test-bed for eliciting and validating models of social interaction, since in this representation all data is digital, we are able to control it. This enables the study of observation interdependencies (voice, gaze, distance, gestures, facial expressions). Therefore, VMR can be used to study how these individual observations influence features of social interaction and social behavior.
- Thirdly, a virtual reality environment can be used to allow real-time and natural remote meeting participation. In order to facilitate this, we need to know which elements of multi-party interaction during a meeting need to be presented in a virtual meeting in order to obtain as much naturalness as possible. The test-bed function of a virtual meeting room, as mentioned above, can help one to find out which (nonverbal) signals need to be mediated in one way or another.

Although these items suggest that the main uses of VMR are for replay, test bed and remote participation, we can, of course, also offer VMR to the meeting participants inhabiting the physical meeting room while they are interacting. While meeting they can get all kinds of information about the meeting presented in this virtual environment and they can use it as a domain-dependent browser to ask questions such as: Who is this person, what did he/she say about this topic in a previous meeting, why is this person getting upset when we talk about this topic, etc.? Hence, due to this visualization, meeting participants may feel stimulated to ask questions related to behavior of meeting participants, meta information displayed in the environment and events taking place (without disturbing the meeting). Clearly, when looking at VMR from this point of view it serves the role of providing live meeting assistance to the meeting participants present in the real meeting room. The visualization provides the context for the user to interact with the system and it provides the context for the system to interpret and assist the user. Remote or on-line access of VMR is, of course, no problem. That is, nonmeeting participants can view in from a predefined position, or even experience the meeting through the eyes of one of the representations.

This audience is not necessarily visible to the meeting participants. A slight extension should allow visualization of the audience, for example as avatars, and make the meeting participants, still assuming that they use the VMR as a live meeting assistant, aware of who is in the audience. We have not done this yet, but it fits in with the tradition of multi-user virtual environments, where in this case the multi-user environment can be constrained to a public gallery, not disturbing the meeting. Obviously, many other ideas common in multi-user virtual reality environments and distributed virtual environments, including the various ways of distributing data and processes, can be introduced here. As mentioned in the first item, VMR allows us to reconstruct a meeting, but not necessarily in a way that tries to stay close to the real activities. Gestures can be exaggerated, pointing can be done such that it is more easily recognizable, speech can be improved, and we can even have different combinations of modality display than were present in the real meeting.

4.2 How can we obtain a virtual representation?

Information needed to build a virtual representation of meeting activities can be obtained in real-time from recordings of behaviors in real meetings (e.g. tracking of head or body movements, voice), from manual annotations or from machine learning models that induce higher level features from lower level signals, recordings or annotations. Obviously, when the main part of the annotations is obtained manually by annotators that annotate the meeting off-line, generation or presentation of this meeting information can become close to being perfect, although far from real-time. This assumes that the annotation schemes that are used by human annotators are sufficiently detailed and to allow for (re-)generation of verbal and nonverbal behavior by virtual meeting participants. The more complete the automatic annotation, the more complete a real-time regeneration in virtual reality.

Our computer vision software processes low resolution, monocular image sequences from a single camera. A silhouette is extracted, shadows are removed and skin color is extracted from the silhouette in order to locate hands and heads. Silhouette matching is used to match a projection of a human body model to the extracted silhouette (see Fig. 2). This allows us to display animated representations of meeting participants in a (3D) virtual reality environment (see Fig. 3).

The 3D positions of head, elbows and hands can be calculated reasonably well. 3D technology based upon portable standards, like VRML/X3D and H-Anim avatars is used. For some recordings, electromagnetic sensors

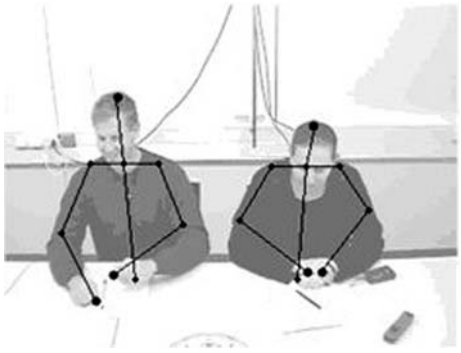

Fig. 2. Human pose estimation using computer vision

Fig. 3. A 3D replica of the IDIAP smart-room

Fig. 4. Humans wearing electro-magnetic sensors

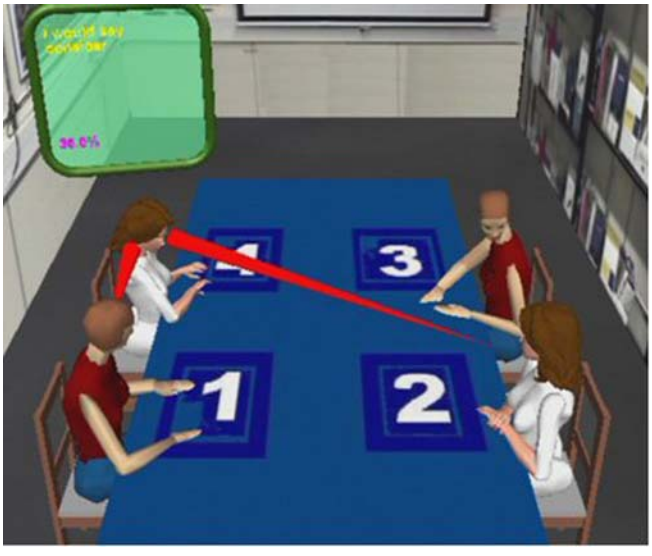

Fig. 5. Reconstructing and augmenting meeting data

were mounted on the heads of the meeting participants for tracking their head movements (see Fig. 4). Especially in meetings, this allows us to record and display in real-time the head orientations of the represented meeting participants. Although there can be differences in head orientation and gaze direction, it nevertheless allows a sufficiently realistic representation of focus of attention behavior (addressing persons, looking at a speaker, looking at notes or looking at the white-board in the meeting room). In Fig. 5, the transformation, using automatic pose and gesture recognition and recognition of head orientation, to the virtual meeting room is shown. In this figure it is also shown how the captured information can be augmented with other information obtained during the course of the meeting.

Clearly, this real-time recognition of meeting activities allows teleconferencing, by which we mean real-time participation by remote participants. Of similar importance is that it also allows simulation of what humans have done in the past, representation of what humans do in a remote meeting and it allows the addition of virtual and embodied meeting assistants. However, we have not designed a fully immersive virtual space for remote participants. Neither facial expressions nor emotions are on display yet.

Intelligence, based on recognition and interpretation of meeting events, is necessary for making an environment smart, for example by introducing intelligent meeting assistants, and by adding intelligence to agents that inhabit the virtual room (see the next section). This is where the annotations of Sect. 2.2 enter and why we need to improve the algorithms that aim at real-time recognition (first-level annotations) and interpretation (second-level annotations) of verbal and nonverbal interactions during

meetings. A virtual room obtained this way can be further augmented with statistics and visualizations and tuned to user preferences.

In Fig. 6 shows an example of gaze behavior during a meeting. The blue domes become more transparent as the participants look in a particular direction for a longer time. This way the bright dots in the dome reveal the most important directions of interest for the individual participants. The size of the black balls reveal the relative level of dominance as predicted by the algorithm discussed in [21]. The size of the cylinder connecting the balls shows the level of interactivity (the amount of subsequent turns in relation to all the turns) between the two persons connected. Obviously, many other aspects can be visualized in similar manners, and the way in which we decided to visualize them is certainly not the only one. More simplified visualizations of for, e.g. talkativeness have been shown by [7].

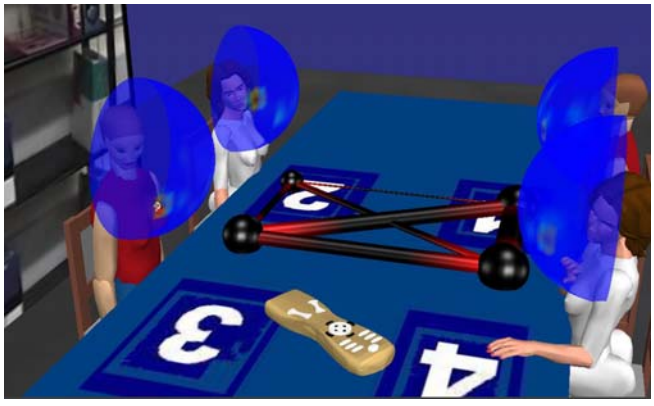

Fig. 6. Displaying gaze and influence of virtual humans

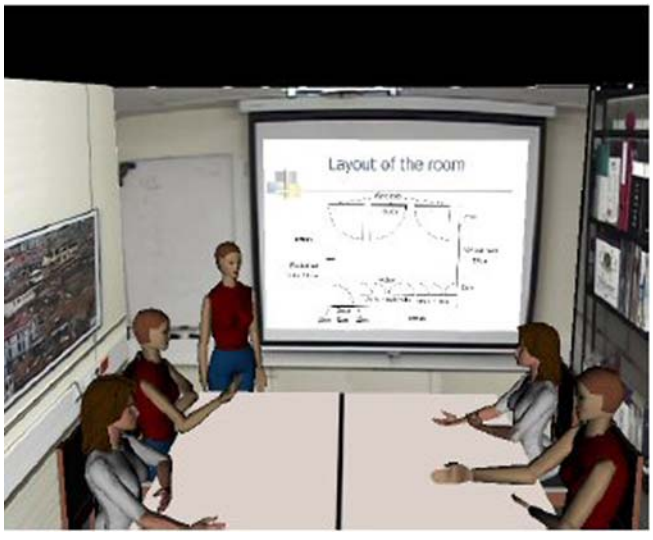

Fig. 7. The virtual presenter

4.3 What roles do virtual humans play?

Virtual meeting participants can mimic what is happening in the physical meeting room, but there is not always a need to mimic everything; moreover, technology or real-time constraints do not necessarily allow that. On the other hand, it is also possible to add information to the behavior of a virtual participant in order to improve communication. Certain characteristics in the behavior, for example, gaze behavior to smoothen turn-taking or make clear who is addressed, can be added. The latter requires automatic addressee detection if we want to do this in real-time. A remote meeting participant may choose to send his avatar to a meeting displaying only listening behavior [12]. The participant is continuously represented, but only when necessary is its owner alerted and takes part, for example, in a mimic mode.

4.4 Meeting assistants revisited

In Sect. 3 we discussed the introduction of meeting assistants. Meeting assistants are agents who know about certain aspects of a meeting and that assist the various participants (including, e.g. a chairman, a note-taker, an off-line visitor and a remote participant). Presently our work focuses on two of these meeting assistants. One is a virtual chairman, an assistant who during a meeting gathers information that is useful for a chairman; in fact, it may act as a chairman. Guarding agenda and time constraints is an obvious task. This also means taking care of the decision-making process; trying to exploit the expertise of the meeting participants, deciding about a presentation, etc. Active software agents assisting in a meeting is discussed in [2, 23]. The second meeting assistant we focus on is an embodied presenter [25]. In the remainder of this section we discuss our work in progress on this presenter (see Fig. 7).

When discussing virtual presenters we can look at on-line presentations where the virtual presenter mimics a human presenter (and maybe add some characteristics in order to improve the performance), presentations where meeting participants present their (PowerPoint) sheets with the help of their presentation agent or off-line presentations where an earlier presentation is regenerated. At this moment, our main concern is to model an embodied agent and we defer questions that are related to off-line, online and (semi-) autonomous behavior to the future. We are looking at existing presentations, available from the AMI project, and are trying to design a script language to represent a presentation as a number of synchronized expressions, that is, create an example

presentation script and rather natural presenter characteristics that allow us to replay an existing presentation in virtual reality. Sheet control, pointing gestures and speech are among the first modalities of a presentation that need to be modelled. Pointing and gestures are other main issues of research for the virtual presenter. Pointing occupies a separate channel in the synchronization language. Apart from constraints on gesture phases (preparation, stroke and retraction) there are synchrony rules that need to be implemented and that take into account the phonological, syntactic, semantic and other rules (e.g. turn taking) that guide the behavior of a virtual presenter. Other issues that are addressed are gaze, hand shape (when pointing) and timing of pointing (moving from a start point to a target area on a sheet). In the future we will look at other gestures and posture shifts that can be employed by a presenter and, most importantly, the possibility of interrupting a presenter [25].

5 A distributed virtual meeting room

As is well-known, there exist systems that allow distributed meetings and virtual collaboration and that also allow participants to perceive each other using various modalities and media [3, 9, 10]. In the AMI project, the starting viewpoint is different. There we have a smart physical environment that supports its inhabitants and captures information for off-line browsing and replay. However, as shown in the previous section, it is a rather natural extension to go from such an environment to an application where we have remote participants or where we connect smart meeting rooms. Unlike existing systems, this introduces the capturing, interpreting and mediating of multi-party interaction inside the different locations that are connected. For this reason, we have been working on a distributed version of the virtual meeting room concept as discussed in the previous sections.

The technology used within the distributed virtual meeting room (DVMR) differs substantially from normal video conferencing technology. Rather than sending video data as such, this data is transformed to a format that enables analysis and transformation. The DVMR-server transforms its input to an up-to-date distributed virtual meeting room [18]. Objects in the DVMR can be controlled/moved by its inhabitants. As an example, since many of our recorded meetings are design meetings devoted to the design of a remote control, we created a remote control and put it in the DVMR as an example of how real and remote meeting participants can discuss and manipulate the properties of this remote control. Clearly, visualizing and manipulating objects that are under discussion, whether they represent physical objects or documents and presentations, is an important issue in advanced meeting technology. The meeting participants have a virtual position at the table, and can watch the meeting from that virtual position or, if they prefer, can watch the meeting from a more global point of view.

For the current version of the DVMR the focus is 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, essentially in the same format as 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 hand-held 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 equipped with whatever hardware is needed and available for meetings on the one hand, 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 such as keyboard or mouse, in order to look around, interact with objects and so on. It seems unlikely that in a more realistic setting people 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.

In Fig. 8 we have illustrated the DMVR concept. As shown in the figure, there is the possibility of transforming meeting activities to other media, modalities and appearances before displaying them to meeting participants. Representing human activity through avatars in the DMVR is not necessarily always the best choice. Other information about the progress of the meeting or characteristics of the meeting participants can be displayed as

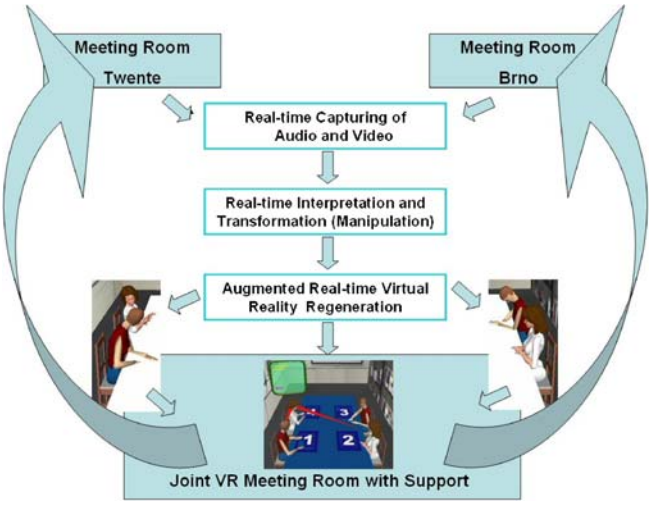

Fig. 8. The distributed virtual meeting room (DVMR) concept

well using various types of visualization, including embodied agents that act as meeting assistants.

6 Conclusions and future work

We have discussed how to go from captured data in a smart meeting room situation to a virtual meeting room. We discussed what needs to be annotated, where in an ideal situation the annotations can be obtained automatically and converted in real-time for virtual reality generation or information retrieval purposes. In current practice there exists a mix of manually and automatically obtained annotations disallowing real-time generation, or there is real-time generation based on imperfect and incomplete data. Whether the latter really is a problem depends on the application. Our research is part of the European AMI project. In this project much more is captured (e.g. emotions) than we have taken into account until now when moving from a physical meeting room to a virtual meeting room. However, realization of (remote) meeting assistants will be our main research topic in the forthcoming years. We will look in particular at agents that provide information about influence and engagement characteristics of meeting participants, agents that try to structure (and visualize) the argumentation going on in the discussions and agents that try to avoid conflicts [23]. In addition to the design of meeting assistants we are looking at interaction issues. Currently, we are working on the integration of speech recognition in the distributed virtual meeting room in order to select or manipulate objects or agents in the environment. Another issue that is being looked at is capturing and mediating gaze behavior between remote meeting participants. Sharing and manipulating shared objects is another issue that requires our attention. Obvious, and also part of our current research efforts, is the visualization of meeting events in virtual reality, interpretations of these meeting events in order to produce semantic preserving transformations, and presentation of these meeting events using various media sources. We also hope to integrate current research efforts on personalization of embodied agents, facial expressions and emotion display into our efforts to obtain meeting environments that allow for more natural meeting experiences.

Acknowledgement We have described joint work of the Human Media Interaction (HMI) group of the University of Twente. This work was partly supported by the EU 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-166). Many people helped to realize the virtual meeting room and the first prototype version of the distributed virtual meeting room (DVMR). Job Zwiers and Jan Peciva of the University of Brno were responsible for realizing ideas existing in our research group for many years and explained in papers since 2000. Earlier research in our HMI research group on multi-agent platforms, on embodied agent animation and on recognizing human bodily movements and mapping these movements on virtual meeting participants has been integrated in the prototype version. Ronald Poppe, building on earlier work in the group, tuned his software for gesture recognition to this particular application. An initial stripped-down version was demonstrated in July 2005 in Edinburgh, where DMVR connection between Edinburgh and the University of Twente was explored.

References

1. Bales, R.: Social Interaction Systems: Theory and Measurement. Transaction Publishers, Somerset, NJ (2001)
2. Barthelmess, P., Ellis, C.A.: The neem platform: An evolvable framework for perceptual collaborative applications. J. Intell. Inf. Syst. **25**(2), 207–240 (2005)
3. Benford, S., Bowers, J., Fahlén, L.E., Greenhalgh, C., Snowdon, D.: User embodiment in collaborative virtual environments. In: Proc. ACM Conf. Human Factors in Computing Systems, May 1995, CHI, vol. 1, pp. 242–249. ACM Press, Denver, CO (1995)
4. Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., Voormann, H.: The nite xml toolkit: flexible annotation for multimodal language data. Behav. Res. Meth. Inst. Comput. **35**(3), 353–363 (2003)
5. Deutscher, M., Jeffrey, P., Siu, N.: Information capture devices for social environments. In: P. Makropoulos, B. Eggen, E. Aarits, J. Crowley (eds.) Proceedings of Second European Symposium on Ambient Intelligence, LNCS 3295, pp. 267–270. Springer, Eindhoven, The Netherlands (2004)
6. DiMicco, J.: Designing interfaces that influence group processes. In: Doctoral Consortium Proceedings of the Conference on Human Factors in Computer Systems (CHI 2004), pp. 1041–1042. Vienna, Austria (2004)
7. DiMicco, J., Hollenbach, K.J., Bender, W.: Using visualisations to review a group's interaction dynamics. In: Proceedings of CHI, Work in Progress Section, pp. 706–711. Montreal, Canada (2006)
8. Frécon, E., Nöu, A.: Building distributed virtual environments to support collaborative work. In: VRST, pp. 105–113. ACM Press, Taipei, Taiwan (1998)
9. Greenhalgh, C., Benford, S.: Massive: a collaborative virtual environment for

teleconferencing. ACM Trans. Comput. Human Interact. **2**(3), 239–261 (1995)
10. Greenhalgh, C., Benford, S.: Virtual reality tele-conferencing: Implementation and experience. In: Proc. Fourth European Conference on Computer Supported Cooperative Work (ECSCW'95). North Holland Press, Stockholm, Sweden (1995)
11. *http://www.ai.sri.com/project/calo/*
12. Maatman, R., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Proceedings of the International Conference on Interactive Virtual Agents (IVA), Kos, Greece, LNCS 3661 (2005)
13. Masoodian, M.: Human-to-Human Communication Support For Computer-Based Shared Workspace Collaboration. Ph.D. thesis, University of Waikato (1996)
14. McCowan, I., Gatica-Perez, D., Bengio, S., Moore, D., Bourlard, H.: Towards computer understanding of human interactions. In: Proc. of the European Symposium on Ambient Intelligence (EUSAI), LNCS 2875. Eindhoven, The Netherlands (2003)
15. McGrath, J.: Groups: Interaction and Performance. Prentice Hall, Englewood Cliffs, NJ (1984)
16. Nijholt, A.: Gulliver project: performers and visitors. In: J. Hemsley, V. Cappellini, G. Stanke (eds.) Digital Applications for Cultural and Heritage Institutions, pp. 285–292. Ashgate Publishing Ltd, Hampshire (UK) (2005).
17. Nijholt, A., Op den Akker, R., Heylen, D.: Meetings and meeting modelling in smart environments. AI & Society, J. Human Centered Syst. **20**(2), 202–220 (2006)
18. Nijholt, A., Zwiers, J., Peciva, J.: The distributed virtual meetingroom exercise. In: A. Vinciarelli, J. Odobez (eds.) Multimodal Multiparty Meeting Processing, Workshop at the 7th International Conference on Multimodal Interfaces (ICMI), pp. 93–99 (2005)
19. Reidsma, D., Opden Akker, H.J.A., Rienks, R., Poppe, R., Nijholt, A.: AI & Society. J. Human Centered Systems (ed. by R. Fruchten), to appear
20. Reidsma, D., Hofs, D., Jovanovic, N.: Designing focused and efficient annotation tools. In: L. Noldus, F. Grieco, L. Loijens, P. Zimmerman (eds.) Measuring Behaviour, 5th International Conference on Methods and Techniques in Behavioral Research. Wageningen, The Netherlands (2005)
21. Rienks, R., Heylen, D.: Automatic dominance detection in meetings using easily detectable features. In: S. Renals, S. Bengio (eds.) Proceedings of the MLMI, LNCS 3869. Springer, Berlin Heidelberg New York (2005)
22. Rienks, R., Heylen, D., Van der Weijden, E.: Argument diagramming of meeting conversations. In: A. Vinciarelli, J. Odobez (eds.) Proceedings of Multimodal Multiparty Meeting Processing, Workshop at the 7th International Conference on Multimodal Interfaces, pp. 85–92. Trento, Italy (2005)
23. Rienks, R., Nijholt, A., Barthelmess, P.: Pro-active meeting assistants: Attention please! In: M. Asako (ed.) Proceedings of the 5th workshop on Social Intelligence Design, pp. 213–228. Osaka, Japan (2006)
24. Turoff, M., Hiltz, S.: Meeting through your computer. IEEE Spectrum **14**(5), 58–64 (1977)
25. Van Welbergen, H., Nijholt, A., Reidsma, D., Zwiers, J.: Presenting in virtual worlds: Towards an architecture for a 3D presenter explaining 2D-presented information. In: Proceedings of the Intetain, pp. 203–212. Springer, Berlin Heidelberg New York (2005)
26. Vertegaal, R.: Look Who is Talking to Whom. Ph.D. thesis, University of Twente (1998)
27. Voida, S., Mynatt, E., MacIntyre, B.: Supporting collaboration in a context-aware office computing environment. In: UbiComp 2002 Workshop on Collaboration with Interactive Walls and Tables, pp. 105–113. Gotenborg, Sweden (2002)
28. Waibel, A., Steusloff, H., Stiefelhagen, R.: Chil – computers in the human interaction loop. In: Proceedings of NIST ICASSP Meeting Recognition Workshop. Montreal, Canada (2004)
29. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., Lathoud, G.: Modeling individual and group actions in meetings: a two-layer hmm framework. In: Second IEEE Workshop on Event Mining: Detection and Recognition of Events in Video, In Association with CVPR (2004). AMI-11




ANTON NIJHOLT is full professor and chair of the subdepartment Human Media Interaction of the Department of Computer Science of the University of Twente. His main research interests are multi-party and multi-modal interaction, virtual environments and social and intelligent agents. Before joining the University of Twente he held positions at the Vrije Universiteit Brussels and several other universities in the Netherlands and Canada.

RUTGER RIENKS is a third year PhD student in the Human Media Interaction research group of the University of Twente. His activities focus on the extent to which computers can replicate human abilities to perceive and comprehend multiparty interaction. He has published on meeting modeling in general; on the role that technology can fulfill in the meeting domain and has shown possibilities for applications on various dimensions of the meeting process.

JOB ZWIERS is associate professor at the Human Media Interaction group of the department of Computer Science of the University of Twente. He has a background in physics and theoretical computer science. After working several years on verification of distributed systems, he shifted his interest towards human computer interaction, computer graphics, and multi-agent systems.

DENNIS REIDSMA is a third year PhD student in the Human Media Interaction research group of the University of Twente. Currently, his activities focus on the problems and issues that arise in creating large, multiply annotated corpora, such as development of annotations schemes and tools and investigating reliability of annotations. Related to this is his interest in the application of knowledge about human interaction gained from these corpora in embodied agents.