Multi-party Interaction in a Virtual Meeting Room

Rutger Rienks, Ronald Poppe, Anton Nijholt, Dirk Heylen and Natasa Jovanovic
Human Media Interaction Group, University of Twente, Enschede, The Netherlands

Abstract
This paper presents an overview of the work carried out at the HMI group of the University of Twente in the domain of multi-party interaction. The process from automatic observations of behavioral aspects through interpretations resulting in recognized behavior is discussed for various modalities and levels. We show how a virtual meeting room can be used for visualization and evaluation of behavioral models as well as a research tool for studying the effect of modified stimuli on the perception of behavior.

Keywords
Multi-party interaction, virtual environments, behavioral recording, automatic data acquisition techniques

1 Introduction
Meetings play an important part in daily life, they are everywhere. A meeting is when antropomorph entities interact [1] and can be seen as a gathering of thoughts where the exchange and generation of information leads to an enhanced level of knowledge improving the performance of the individuals as well as the group [2]. Generally, a meeting is held in order to move group actions forward through decision making by information presentation and collaboration. Ideally such a meeting proceeds efficiently and effectively, is manageable and accessible afterwards. A meeting can be seen as a series of related interactions amongst participants. The behavior of these participants during the interactions is crucial for the resulting outcome.

The HMI group of the University of Twente has a tradition in research in multimodal interaction with embodied conversational agents, research in computer graphics for virtual environments and machine learning techniques for recognition of higher level features (such as dialogue acts, gestures, emotions) from lower level features (such as words, hand arm movements, facial features).

This paper gives an overview of the current research carried out at the HMI group on automatic observations of behavioral aspects in meetings. The remainder of this paper is organized as follows: Section 2 deals with the typically behavioral meeting aspects. Direct and indirect observable meeting behavior is discussed in Section 3 and 4 respectively. In Section 5 we elaborate on fusion and visualization of several recognized aspects of human behavior.

2 Behavior in meetings

We define behavior as the set of external characteristics that an object exhibits as a response to external or internal stimuli that might be observed, taught, learned and measured demonstrating a competency, skill, ability, or characteristic.

The behavior of meeting participants is generally evaluated relative to social norms and regulated by various means of social control. These norms are unstated and generally unwritten. Typical forms of social norms one might encounter in meetings are that one should not yell or scream, that one should let people finish talking, that one should not start private talks, that one should not whisper and that e.g. '*Ad Hominem*' arguments are not allowed. These social norms or conventions define the shared belief of what is normal and acceptable and hence restricts the people's actions. Operand conditioning plays an important role in their establishment and fulfillment. Violations of norms can be punished with sanctions and violators are considered eccentric or even deviant and are stigmatized.

Sometimes a chairman is appointed and given the authority to manage the meeting process. He or she should make sure that social norms are adhered to, follow a predefined agenda and/or maximize the output of the meeting. This chairman is authorized to perform a set of interventions such as selective turn giving and interrupting. These typical actions are triggered, dependent on the displayed behavior of the participants.

The occurrence of unwanted situations such as a rare event with a large disturbing impact, or the repetitive occurrences of events with a smaller disturbing impact are typical examples that could trigger an intervention. A chairman, or in general every meeting participant could for example intervene if someone is disturbing the meeting process by e.g. continuously repeating him or herself without listening to the other participants.

A problem with human chairmen is that they are usually biased towards certain positions on issues or towards certain persons. Instead of appointing chairmen, one could use systems that are able to both automatically observe the behavior of participants and to automatically regulate the meeting. These systems are potentially cheap in use and can, if the observations are stored, be queried for all kinds of user interests (such as number of turns per speaker, total meeting duration etc.).

Human behavior reveals itself through several modalities over time. Meeting participants exhibit characteristics from the first moment they encounter each other. To explore some of these behavioral characteristics, someone could start by observing the happenings. Simple (possibly automatic) frequency counting could suffice in order to get some first impressions. Ethograms are generally used for this task, they are created using labels based on a predefined behavioral dictionary. For meetings, we could for instance be interested in the turn frequency of the participants. This all seems plausible if one is just willing to observe and nothing else. However, as we want to observe automatically and even more, also want to respond appropriately (e.g. to restore order), we need to know what caused the behavior.

The intentions of the exhibited behavior are related to the individual agenda of the participants and the amount of effort they are willing to put into realizing this. A typical agenda could consist of a number of topics for debate and a possible set of constraints such as a limited amount of available time.

This last constraint could cause people to respond more briefly and compliantly.

So what is measured can be attributed to internal stimuli, but what these internal stimuli precisely are, is yet mostly something to guess about. Another aspect occurs when two people are engaged in a conversation. In this case, a balance should be maintained between various levels of communication. Tracy [2] describes these levels of communication as task, or instrumental and face goals. An example of such a balance for a participant is the urge to immediately achieve one's agenda or objective (task goal) on one hand and to act in line with social norms and roles (face goal) on the other hand.

Figure 2 describes the process that we think should take place in a system able to act upon recognized behavior or input in a meeting environment [1].

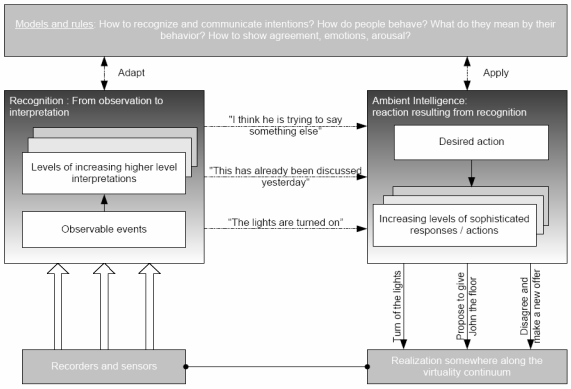

Figure 2. *Observed meeting behavior being recognized, interpreted and regenerated.*

On the lowest level, there are several sensors that observe the meeting. These observations are then to be recognized or labeled automatically using various models of behavior, or behavioral catalogs. The recognized behavior is then transferred to a module that should generate the appropriate actions. This way, various rules of 'accepted behavior' are monitored and in case of violations, the system should decide upon its action. This action selection can be done by either applying predefined actions for each violated threshold or deriving an action real time from which the expected impact is the most appropriate.

To let a system know what it actually observes we make a distinction between direct observable behavior and behavior that is derived or interpreted. We elaborate on this in the next two sections.

3 Directly observable behavior

Directly observable behavior can be immediately measured from the input media without taking the context into account. We measure behavior in each modality (poses, speech, gaze) separately. Speech recognition results in a transcript of what is said but does not yield a semantic interpretation. Body poses can be estimated but are not interpreted until the gesture recognition. Gaze is measured in terms of head orientations. We can obtain these observations through sensors. Within the AMI project, smart meetings rooms are used to collect this data. These rooms are equipped with sensors such as cameras, microphones and in certain scenarios also electronic pens and orientation sensors placed on the participants' heads are used. We will now discuss each of these modalities in turn.

3.1 Body pose estimation

Body pose is an important aspect of behavior. It can be an indicator of involvement (leaning backwards or forwards) and focus of attention (gazing at the speaker, looking at notes). In a multi-party setting, a body pose is estimated for each meeting participant individually. In general, the human body is modeled as segments that are connected with joints. Each joint can have a number of degrees of freedom. A pose is described by a value for each of these degrees of freedom.

Poses can be measured with motion capture equipment but state of the art in computer vision allows for cheap and relatively robust pose estimation without being obtrusive. An overview of recent work on vision-based human motion capture can be found in [4]. Our group has estimated poses from extracted silhouettes and tracked and labeled skin regions [5]

3.2 Gaze detection

Kendon [6] groups the determinants or functions of gaze behavior into five classes: providing visual feedback, regulating the flow of conversation, communicating emotions, communicating attitudes and interpersonal relationships, and improving concentration by restricting visual input. We carried out experiments where electromagnetic sensors were mounted on the heads of each participant in some four party meetings. This information was e.g. used to investigate whether the current speaker could be estimated [7] given the head orientations of all the participants at a specific time. We have shown that in a four person meeting in 79% of the cases a speaker can be predicted on the basis of head orientation of the participants only.

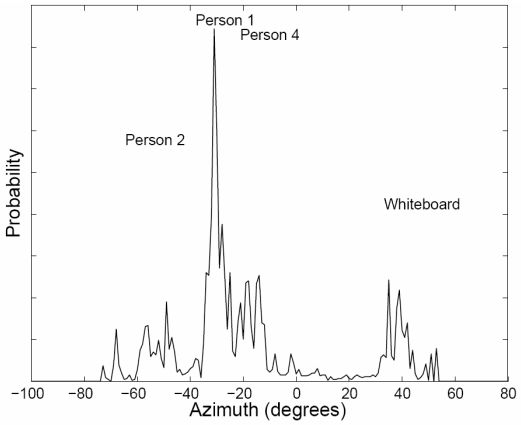

Figure 3. *Distribution of directly observed head orientations along the horizontal (azimuth) plane for person 3 in a four-person meeting setting.*

On the other hand we could derive the participants' attention by creating histograms along their lines of sight. Figure 3 shows the histogram of Person 3 in a particular meeting. A value of zero degrees along the x-axis corresponds with a straight head orientation. The location of the other participants is shown as well as the location of the whiteboard. It follows clearly that person 3 had his head much more often directed towards person 1 when compared to person 4, who was sitting almost opposite. This is information can be useful for addressee detection (See section 4.2).

3.3 Speech recognition

A transcript of what is said by each meeting participant can be obtained by automatic speech recognition (ASR). When issues with multi-speech and background noise can be solved, ASR can generate relatively accurate hypotheses about what is said. These issues are addressed within the AMI project.

Currently, more and more of our research focuses on the extraction of emotional state from speech. This information could be used to detect situations where a participant is enthusiastic or irritated. Other current research topics in this area are automatic topic segmentation and the creation of a meeting dictionary. In contrast to direct observations, specific language models are generally trained to further improve the recognition rates. These sorts of models are essential for the detection of interpreted behavior which is discussed next.

4 Derived or Interpreted behavior

Derived or interpreted behavior is behavior that is not directly observable for a system. Usually recognition systems do not receive more input than an audio and a video signal. To derive relevant aspects of meeting behavior for a chairman for example, one should have models describing how these aspects can be derived from one or more directly observable behaviors. These kinds of models and rules can be extracted by examining large data sets.

4.1 Dominance detection

People who are too dominant in meetings violate the process of collective decision making for which many meetings are intended. We were able to create a system that was able to reach an accuracy of 75% performance when classifying meeting participants as either 'Low dominant', 'Normal Dominant' and 'Highly Dominant' [8]. This classification appeared mainly dependent on the number of floor grabs by a participant and the number of turns someone took during a meeting.

4.2 Addressee detection

Addressing is another important aspect of every form of communication. In small group face-to-face meetings, a speaker can address his utterance to a single individual, to a subgroup of individuals, or to a whole audience. Due to the limitation of the available data for studying subgroup addressing, we limited our current research to the development of a system that automatically identifies whether an utterance of a participant is addressed to the whole group or to just one of the meeting participants. To train our system we used Bayesian Networks. These networks were supplied with a set of utterance, gaze and contextual features in order to automatically identify the participant(s) to whom the speaker is talking. The best performance (82.59 %) was achieved using a combination of all three types of features. [9]

4.3 Argumentation extraction

Once a meeting is over, all that is left are notes that typically contain decisions, action points and perhaps some issues that were left open. Current effort is put in revealing the decisions of a meeting as well as the lines of deliberated arguments in order to provide (automatic) access to representations of conveyed meeting information. The expected behavior of the participants during a discussion can be very interesting for management teams deciding about who to send to which meeting. The desired outcome of this system should eventually provide information about how decisions were made and who brought in which arguments.

An aspect of argumentation that is closely related to behavior is *rhetoric* (How something is told). The ancient Greeks already saw this as the logical counterpart of *dialectic* (What is told). Aristotle defined three main forms of rhetoric: *Ethos* (How the character of a person influences the audience to consider him to be believable), *Pathos (*How emotions affect the message) and *Logos* (How the use of language affects the message). Al these aspects relate to the way people behave or should behave in order to convince someone.

Our ultimate aim is to provide access to representations of conveyed meeting information showing decisions as well as the lines of the deliberated arguments in addition to ordinary meeting notes. Current effort is put into examining what sort of discourse structuring model should be applied to meetings in order to capture the discussion in a useful manner.

5 Regenerating Behavior

If a system has interpreted the observations and recognized one or several behavioral aspects we could use this information for various purposes. Section 5.1 will elaborate upon the visualization of the interpreted data for possible evaluation purposes for both observations and models of behavior. Section 5.2 discusses the possible influence of modified behavioral aspects on the perception of social behavior. For both visualization and evaluation a Virtual Meeting Room (VMR) is used.

5.1 Visualizing recordings

The behavioral aspects that are recognized in each modality can be visualized to evaluate the recognition. One way to do this is to reconstruct the media from which the aspects are recognized, thus to reconstruct the video, the audio and possible other media. This would ideally result in exactly the same video and audio as that we used as input. However, what we recognize is an abstraction of the media itself. For example, from audio we recognize what has been said but we abstract from intonation, pitch and even accent and speed. All these aspects have to be filled in to make our reconstructed media exact duplicates: an impossible task.

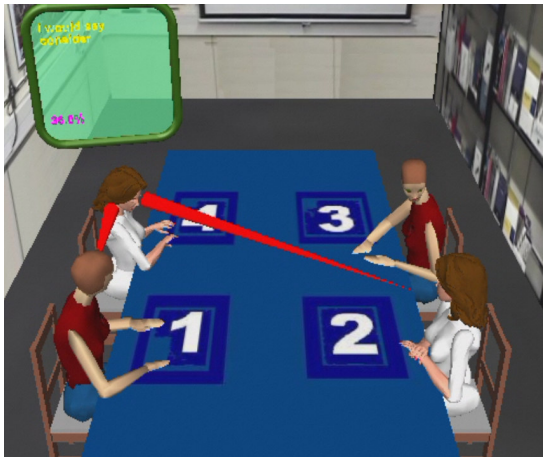

Figure 4. *The virtual meeting room showing gestures, head movements, the speech transcript, the addressee(s) of the speaker and the percentage a person has spoken the entire meeting.*

Instead, we can make a reconstruction of our recordings where only the semantics are preserved. For example, we could have a visualization where the gestures are performed differently but have the same meaning (e.g. pointing). We will now describe our VMR it enables us to create exactly these kinds of visualizations.

The VMR is a 3D virtual environment and a replica of a real meeting room, see Figure 4. A benefit of this virtual environment is that different modalities can be controlled. Each modality can be omitted or adapted individually. Another aspect is that the VMR allows visualization of recognized behavioral aspects that in turn can be evaluated. The meeting room can be viewed from all possible angles and it is even possible to be immersed in the VMR by means of a head mounted display (HMD).

5.2 A research environment

The ability of virtual environments to remove or manipulate modalities is very useful for research into social behavior. It allows investigation of which modalities are important for the perception of behavior and what the relation is between modalities. In the experiment we described in Section 3.3, we looked only at head orientations: speech and gestures were omitted.

Another research possibility is to transform modalities [10]. Instead of visualizing in an abstract form what behavior is recognized, the recognized behavior is transformed in order to evaluate the human perception of this behavior. For example, in a gaze experiment, the gaze behavior of the speaker can be mirrored and the perception of the addressee(s) can be measured.

6 Future Work

Meetings can nowadays be assisted with a huge variety of tools and technology, ranging from completely passive objects like a microphone to completely autonomous actors such as virtual meeting participants. Along this line it will not be long before meetings can be held where participants can participate remotely in an immersive virtual meeting space. Next to the virtual replicas of the actual persons, there could be all sorts of active software agents assisting the meeting [11]. Virtual participants (e.g. a virtual chairman) are examples of these possible meeting assisting agents.

When we talk about a virtual chairman, in the most ideal case we would like to have a software-driven virtual representation of a human that is indistinguishable from the representation of a real human. Therefore he or she should not just look like a real human, but also behave like a real human. In an ordinary meeting the chairman has to manage the meeting process in order to maintain the meeting atmosphere, follow a predefined agenda and/or maximize the output of the meeting. Typical mechanisms of the virtual chairman to steer the meeting process are e.g. selective turn giving, interrupting and summarizing.

All these mechanisms depend upon the underlying models of behavioral aspects describing how to interpret observations. These models, including the ones described in this paper will all be crucial for a successful realization. To further improve them in the near future more experiments are to be conducted aiming to find out which various modalities influence the perception and interpretation of social behavior and how they do it.

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-101).

References:

1. Rienks, R.J., Nijholt, A., and Reidsma, D. (2005). Meetings and Meeting Support in Ambient Intelligence. *In Ambient Intelligence, Wireless Networking, Ubiquitous Computing*, Artech House, Norwood, MA, USA.
2. Moran, T.P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., Van Melle, W., and Zellweger, P. (1997). I'll get that off the audio: a case study of salvaging multimedia meeting records. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 202–209. ACM Press.
3. Tracy, L.K., and Coupland, N. (1990). Multiple goals in discourse: An overview of issues. *Journal of Language and Social Psychology*, 9:1–13.
4. Moeslund, T.B., and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268.
5. Poppe, R.W., Heylen, D., Nijholt, A., and Poel, M. (2005) Towards real-time body pose estimation for presenters in meeting environments. *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, Plzen, Czech Republic
6. Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica* 32, 1–25.
7. Rienks, R.J., Poppe, R., and Poel, M. (2005) Speaker Prediction based on Head Orientations, *Proceedings of the Benelearn 2005*, Enschede, The Netherlands, pp. 73-79
8. Rienks, R.J., and Heylen, D. (submitted). Automatic Dominance Detection in Meetings Using Support Vector Machines.
9. Jovanovic, N., Op den Akker, R., and Nijholt, A. (submitted) Addressing in face-to-face meetings.
10. Bailenson, J.N., Beall, A.C., Loomis, J., Blascovich, J.J., and Turk., M. (2004) Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators and Virtual Environments*, 13(4):428–441.
11. Ellis, C.(S.), and Barthelmess, P. (2003) The Neem dream. *Proceedings of the 2003 conference on Diversity in computing*, 23–29. ACM Press.