Augmented **M**ulti-party **I**nteraction
http://www.amiproject.org

Augmented **M**ulti-party **I**nteraction with **D**istance **A**ccess
http://www.amidaproject.org

State-of-the-art overview

Recognition of Attentional Cues in Meetings

(January 2006)

1 Introduction

Attention refers to the cognitive process of selectively concentrating on one thing while ignoring other things. This is an everyday wording of a definition given by one of the first great psychologists, William James:

> Everyone knows what attention is. It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought... It implies withdrawal from some things in order to deal effectively with others.
>
> (Principles of Psychology, 1890)

Attention is basic for all forms of perception; for outer perception via our senses, vision, hearing, taste and smell, as well as for inner perception, the perception of feelings, emotions, thinking. Attention processes allow us to direct attention to certain aspects in the environment with special care. An impressive example of the capacity to direct attention is the *cocktail party effect*: at a party, in the chaos of noise and voices, we are able to focus and concentrate on voices further away and thus follow a conversation, despite the other acoustic signals that continuously reach our ear. We are also especially good at selectively directing our attention to one or another aspect of an object, such as its color, shape or movement.

The amount of incoming information to the primate visual system is much greater than that which can be fully processed. Only part of this information is processed in full detail while the remainder is left relatively unprocessed. ([18]). There are two prime mechanisms in attention control: our sense system is constantly aware of its environment and changes ask for our immediate attention, second there is the process of directing our attention and selecting the information channel in order to better do what our task requires us to do. The first, bottom-up attentional selection process is a fast, and often compulsory, stimulus-driven mechanism. The other mechanism, top-down attentional selection, is a slower, goal-directed mechanism where the observer's expectations or intentions influence the allocation of attention. Observers can volitionally select regions of space or individual objects to attend. (see [18].)

Although attention is an innate activity of the mind, it shows itself in observable phenomena of body movements, facial expressions and the way people (or animals) act or behave. So it is tacitly assumed that eye gaze, body posture, arm and head movements are phenomena that go with changes in the attentional state of a person. This implies that these phenomena are possible *cues* pointing at the attentional state of a person or a change in his attentional state.

Attention is required for every form of sensual and cognitive perception, and thus for obtaining information. This is the very reason why *eye-gaze*, and head movements are *cues* for identifying the attentional state and focus of attention of someone. Focussing our attention to something or someone in our environment is done by directing our sense organs towards the object that we want to perceive.

Gaze direction is thus a natural sign for focussing attention to something in our visual environment. Laying our hand behind the ear and directing our head in the direction of a sound source is a natural sign of focussing our auditive attention to something that we hear. Becoming aware of this natural function, gaze becomes also a signal that is used to express attentiveness. Gaze patterns, as well as body postures thus become codes, by means of which the interactors show whether they are or are not attentive, whether they are willing to hear, or to interact. Backchannels ([31]) are codes used by hearers to show their attentiveness and signals for the speaker that he can go on. Deictic pointing is a way of referring to something and attracting someone's attention to that something that we want to highlight as being of special interest in a specific situation. Deictic pointing gestures are often

accompanied with verbal gestures especially the use of referring expressions that help to identify the referred object. Gazing at something is the third way in which people try to focus other peoples attention.

Since our perceptions are mostly related to organized activities, which are being performed to achieve specific aims, the attentional cues do not occur unrelated. They show specific patterns. Identification of these patterns of attentional cues and how they correlate with specific activities is important for recognition of these activities from the observable cues.

The term *focus of attention* has different meanings in different research areas. The primary use of the term focus of attention refers to a perceptual variable indicating the object or person someone is attending to ([4]), a description of someone's focus of attention during an activity. At a semantical level the description of someone's focus of attention during an activity involves describing which actions, objects or people someone is attending to. At a syntactical level this could involve describing the spatial and temporal properties of someone's (visual) attention. As such this is not a directly observable category.

As people often orient themselves towards the physical objects or persons they are attending to, the notion of focus of attention has a derived meaning referring to the physically observable behavior of orientation towards an object by means of posture, head orientation and/or gaze. This could be called the *visual focus of attention*. Psychological attention and physically observable attention do not necessarily coincide but are correlated highly. This is a generally held assumption.

The term *focus of attention* is also used within a (computational) linguistic context. In a theory of discourse structure, developed by Grosz and Sidner, three components of discourse structure are distinguished: linguistic structure, intentional structure, and attentional state. The *attentional state* is considered as an abstraction of the discourse participants focus of attention. This state records the objects, properties, and relations that are salient at a given point in the discourse. (see [5])

Dialogic attention involves listening to a person (auditory mode of dialogic attention) or speaking to one or more persons (articulatory mode of dialogic attention). The focus of dialogic attention identifies these persons, the extent of dialogic attention describes the number of persons within this focus. ([27])

Research in focus of attention and in techniques for recognition of attentional cues has the interest of several research areas, including human interaction, conversational analyses, human machine interaction, user interfaces and attentive systems, embodied conversational agents and virtual environments.

2 Recognition of Attention in Meetings

Meetings are coordinated multi party activities where people collaborate in a shared task. Joint attention is required for effective collaboration. Most of the activities are conversational, participants verbally exchange ideas, have discussions. In order to understand what is going on in a meeting, to be able to write a summary of activities, discussions and decisions made, we need to understand the basic processes that underly these activities and also how these processes show in observable phenomena. Communication is a joint activity, the process of sharing ideas. In conversations a speaker produces meaningful signals by which he tries to direct the attention of the listeners mind to the ideas that he wants to communicate. The speaker selects those codes that he expects are most effective in directing the attention of the interlocutors in the way he wants.

By pointing at something people try to attract someone's attention to something. (spatio deictic referring). For example someone is looking in the direction of the window, then points at it and says:

"Look there is Socrates".

Several issues are involved:

a) that someone is making a movement with his arm, hand, that he moves his body to the one who's attention he tries to attract, his eye-gaze and head-movements, his speech, tone, and wording.

b) that this movement made is a pointing gesture, that he points at something to attract the attention of someone, who is in the same field of perception.

c) how this pointing gesture is related to other simultaneous activities by the same or other actors in the environment

d) how this pointing gesture fits in a sequence of activities over time by the same or by other actors or things in the environment

e) what the target is that the actor refers to

f) what the intention is that the actor has with pointing at this target

For automatic recognition of such a conversational gesture as deictic pointing we need computer vision technology as well as knowledge about the situation and the course in time of the conversational activities in which the act of pointing is embedded. Several modalities play a role in the process of recognition and interpretation.
Since the interpretation of the actions, hand and head movements made by participants in meetings can only be inferred when considering them in synchronized and parallel sequences of actions often Dynamic Bayesian Networks or (multi-layered) Hidden Markov Models are used for the prediction (classification) of the movements in terms of signals that have some particular pragmatic meaning in the context of the meeting.

People in meetings coordinate their actions in collaboration. Research has been performed to recognize meeting activities and segment meetings into sequences of meeting activities such as, note taking, discussion, presentation. These activities show in particular patterns of joint attention or *group focus of attention* and requires modeling simultaneous activities of all participants in a meeting. A meeting location has a number of distinguished locations of interest that participants refer to or focus their attention at in a meeting: white-board, laptop, documents, participants, someone entering the room, and in the AMI design project meetings the prototype of the remote control is a recurrent topic of interest. Head movements is one of the features (together with speech, linguistic and paralinguistic features, body postures and arm movements) that are used for recognition of sequences of meeting activities. (see [1], [7], [15], [20], [19],[32], [2].)

3 Gazing Behavior in Face to Face Communication

A speaker can use gaze to indicate that the party being gazed at is an addressee of his utterance. ([3]) Listeners gaze and direct themselves towards speakers in order to hear and see better what the speaker is saying. At the same time the speakers monitors who is listening and how his speech is received by the hearers. Gaze behavior is also of importance in turn-taking management. There are specific patterns in gaze behavior and hence in head movements related to conversational behavior. Identifying these patterns and recognizing this patterns is part of understanding what is going on in

the meeting in terms of who is being addressed, who is trying to take the floor, who is anticipating floor changes, who is participating in a conversation.

Speakers gazing practices often demonstrate explicitly to co-participants that an initiating action is being directed to a particular party, thus selecting that party to speak next. This shows the gazed-at participant that he or she is the intended recipient, and it shows the participants not gazed at that they are not the intended recipient. For this method to work, then, an intended recipient must see the gaze. Others may also need to see it to grasp that someone (else) has been selected. (see:[14]). As has been shown already in the sixties and seventies, eye gaze serves an important role in guiding the conversation, both at the side of the speaker and the listener (e.g. [10]). At the speaker's side, looking at the listener may serve the function of monitoring the attention level and the processing status of the incoming speech, and help to regulate the flow of conversation. At the listener's side, looking at the speaker serves both the function of providing feedback for the speaker's monitoring activity, to inspect the speaker's behavior (facial expression, posture etc) for information about the speakers attitude and emotion, and to monitor for nonverbal cues for turn-taking. Most of these early findings concerning the use of nonverbal cues in communication are based on the analysis of dyadic conversations. Later research on triadic and multi party conversations has confirmed and extended the early findings. [30] provide evidence that gaze behavior is a reliable predictor of addressee-hood.

The main focus of [26] is the exploration of behavioral cues that could potentially be used for classification of the addressee of utterances in a situation where two users interact with each other and with a service system. Facial orientation, utterance length and reactions on system events are considered important cues for focus of attention. In order to evaluate the potential integration of these features, Naive Bayes classifier is used. Classes correspond with attentional states of the users. Both participants look at the system (class A), The speaker looks at the system but the partner looks at the speaker (B),The speaker looks at the partner but the partner looks at the system(C) and Both participants look at each other (D). Results show that facial orientation together with systems events (dialogical context) together are reliable features for predicting the attentional state of the interaction.

4 Research on Head Orientation and Gaze

Most existing work in detecting a user's visual focus of attention makes use of camera-based head pose tracking ([24], [23]) or eye tracking ([29]). The eye-based gaze detectors require a robust eye-tracker, and then typically extract gaze from the position of the pupils relative to the userŠs head. The head-based techniques estimate the focus of attention by determining the orientation of the userŠs head and assuming that the user is looking in the direction in which their head is pointing. These techniques can be accurate, but the fact that they are camera-based means that they are typically not mobile, and can encounter difficulties when the lighting of the scene changes. (see also: [16].)

Head position and eye gaze together and interactively determine whether an observer looking at a picture of a face judges whether the person on the pictures is looking at the viewer or next to him.

Gaze direction is constituted by head orientation and eye orientation ([12]) and can be used as a deictic signal, indicating the current focus of interest ([13]). We have reason to believe that we can use head orientations as a valid substitute for gaze when determining the focus of interest (c.f. [17]). In an experiment with a four-person setting, it was found that in 87.0 % of all cases, the participants rotated their heads and eyes in the same direction

Eye gaze is hard to monitor in a non-intrusive way. Therefore most systems that want to detect visual focus of attention of a person use head orientation. How good can we predict eye gaze from head orientation?

In 88.7% of the time, the focus of interest could be determined solely by the head orientation. Based on the fact that the head orientation component of gaze is so prevalent, it is expected that we see the same systematics that occur in gaze behavior when looking at head orientations alone. This is validated by analyzing a corpus consisting of head orientations and speaker data. ([23]).

How good are outside observers of meeting, i.e. people looking at a meeting through video, in telling what the focus of attention of a participants is? How precise are they in deciding where the head of the participants is directed to? The research described in [21] shows that differences in gaze behavior between speakers and listeners in a multi-party setting also exist when we look at their head orientations. By analyzing a corpus of four-person meetings it appeared that speakers are generally being looked at by more persons than listeners are. In an experiment, conducted in a virtual environment, it was found that observers apply these systematics when asked to identify the speaker when shown only the head orientations of the meeting participants. The virtual environment proved to be a suitable tool for research in perception of human behavior since it allows for good stimulus control.

In [23] an overview of work on tracking focus of attention in meeting situations is presented. A system has been developed that is capable of estimating participants focus of attention from multiple cues. The system employs an omni-directional camera to simultaneously track the faces of participants sitting around a meeting table and use neural networks to estimate their head poses. In addition, it uses microphones to detect who is speaking. The system predicts participants focus of attention from acoustic and visual information separately, and then combines the output of the audio- and video-based focus of attention predictors. The work reports recent experimental results. In order to determine how well we can predict a subject's focus of attention solely on the basis of his or her head orientation, we have conducted Experiment in which head and eye orientations of participants in a meeting were recorded using special tracking equipment show that head orientation was a sufficient indicator of the subject's focus target in 89% of the time. The paper also discusses how the neural networks used to estimate head orientation can be adapted to work in new locations and under new illumination conditions.

5 Annotating Focus of Attention

In order to train and evaluate the quality of attention recognition techniques hand annotated video corpora are made, in which the focus of attention of each of the participants is labelled continuously. A fixed list of possible targets is used to identify the focus. In annotating focus of attention we stick to the visual focus of attention of individuals, defined by the head orientation or eye gaze. So if someone is looking at a person but thinking about his upcoming holiday we will only label where he is or she is looking at.
For research in addressing behavior in face to face meetings focus of attention of participants in scenario based recorded meetings (collected in the M4 and AMI projects) was annotated. The coding is based on observations of the gaze and head turning of participants. The target set of interests for this research on addressing are meeting participants. Reliability of marking the changes in the gazed target (segmentation) was about 80%, and reliability of target labeling showed a kappa value of 0.95. Experiments with Bayesian Network models show that information about the focus of attention of participants, speakers as well as listeners contributes to the reliability of addressee prediction. ([8, 9])

Telling in what direction or what the participants are looking at is not only relevant for reliability of coding of the focus of attention in meeting behavior but also for technology mediated meeting participation, where remote participants rely on similar video and audio technology as annotators.

6 Applications

Modeling and tracking a person's focus of attention is useful for many applications: Intelligent supportive computer applications could use information about a userŠs focus of attention to infer the user's mental status, his/her goals and cognitive load and adjust their own responses to the user accordingly. For multi-modal human computer interaction, the user's focus of attention can be used to determine his/her message target. For example, in interactive intelligent rooms, focus of attention could be used to determine whether the user is to control the refrigerator, the TV set, or he/she is talking to another person in the room. ([22])

Recognition of the attentional state of communicating and collaborating agents is a requirement for attentive systems, which observe user activity to anticipate user needs as well as for Attentive User Interfaces, user interfaces that manage the user attention deciding when to interrupt the user, the kind of warnings, and the level of detail of the messages presented to the user ([28]).

Systems are build that integrate perceptual attention into multi-party, multiconversational dialogue layers [25]. A computational model of the dynamics of attentional state should model the perceptional aspects and the way the senses react to the perceived signals, as well as the activities, goals and intentions in which the actors are involved. In [11] a computational model of controlling the focus of perceptual attention for embodied agents is proposed. It provides the potential to support multi-party dialogues in a virtual world. It demonstrates that embodied agents can respond dynamically to events that are not even relevant to the tasks and shift their attention among objects in the environment.

Finally, Horvitz et al. present an overview of principles and methodologies in research on integrating models of attention into human-computer interaction. ([6]).

7 Conclusion

We provided a short, and necessarily incomplete, overview of research in the area related to the recognition of attentional cues in meetings. The analyses concerns (a) the behaviors of participants in face to face conversations, as well as in face to face meetings, related to various meeting activities (b) the analyses of this behavior by outside observers of meetings, notably annotators of meetings, and (c) the impact for remote meeting participation.

The recognition of attentional cues by machines requires recognition and tracking of human bodies and body parts, head movements, arm and hand positions, and thus builds on state of the art technology in computer vision. We do not give an overview of this research area.

We have only reviewed the main lines of research related to the issue of recognition of attentional cues. Insights gained by this research can help our understanding of what is going on in meetings. The study of meeting behavior, the roles that the various communication channels play in conversations is of prime importance for specifications of the requirements that technology mediated meetings should satisfy. The locations of participants in the meeting, the positioning of video and audio recorders, screens and sound boxes in meeting rooms, they all influence the way (remote) participants and outside observers perceive various modes of interactions that occur in the meeting. The identification of patterns of attentional behavior, head movements, gestures, eye gaze, can only be done after detailed and very careful observations of people in situations that are as realistic as possible. Conversational analysts have been doing this type of research already for many decennia. Often this research is based on observations by researchers who did real-time annotations. The multi-modal meeting corpus recorded in the AMI project makes it possible to gain more insight in what is going on in meetings. We have pointed at the relevance of the issues and research reviewed in this overview for

the understanding of outside observers, annotators and remote participants, who observe and annotate meetings by means of video and audio technology.
Many issues arise that need further investigation. To name one issue: are perceived changes in the audio field a cue for changes in focus of attention of participants in the meeting? What are the consequences for the perception and experience of meeting participation of the reconstructed audio field for remote participants.
A lot of experiments have been performed to study the importance of gaze and mutual gaze in remote meetings, and the impact of the lack of a visual channel on conversations and meetings. Often these experiments are necessarily performed in controlled situations in order to be able to exclude influences that may interfere with the conditions under study. Making it often quite risky to infer results from this experiments to the real situations in which the knowledge should be applied. It is a great advantage that we now have the opportunity to study meetings in settings as realistic as possible in which the technology to support meetings is being used. This is by far the best way to get more insight in the impact of these technologies and to further the development of this technology and to get it tuned to the practice of meetings in the best possible way.

References

[1] Marc Al-Hames, Alfred Dieleman, Daniel Gatica-Perze, Stephan Reiter, Steve Renals, Gerhard Rigoll, and Dong Zhang. Multimodal integration for meeting group action segmentation and recognition. In *Proceedings MLMI'05, Edingburgh, Scotland*, 2005.

[2] A. Dielmann and S. Renals. Multistream dynamic Bayesian network for meeting segmentation. *Lecture Notes in Computer Science*, 3361:76–86, 2005.

[3] Charles Goodwin. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York, 1981.

[4] D. Gopher. *The Blackwell dictionary of Cognitive Psychology, chapter Attention*. Basil Blackwell Inc., 1990.

[5] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12:175–204, 1986.

[6] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication. from principles to applications. 2004.

[7] McCowan Ian, Daniel Gatica-Perez, Bengio Samy, Moore Darren, and Bourlard Herve. Towards computer understanding of human interactions. *Proceedings of EUSAI 2003, LNCS 2875, (E. Aarts et al. ed.)*, pages 235–251, 2003.

[8] N. Jovanovic and R. op den Akker. Towards automatic addressee identification in multi-party dialogues. In *5th SIGdial Workshop on Discourse and Dialogue*, pages 89–92, 2004.

[9] N. Jovanovic, R. op den Akker, and N. Nijholt. A corpus for studying addressing behavior in face-to-face meetings. In *6th SIGdial Workshop on Discourse and Dialogue. Lisbon, Portugal*, 2005.

[10] Adam Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[11] Youngjun Kim, Randall W. Hill, and David R. Traum. Controlling the focus of perceptual attention in embodied conversational agents. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 1097–1098. ACM Press, New York, NY, USA, 2005.

[12] Chris L. Kleinke. Gaze and eye contact: a research review. *Psychological Bulletin*, 100(1):78–100, 1986.

[13] Stephen R.H. Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *Quarterly Journal of Experimental Psychology*, 53A(3):825–845, 2000.

[14] Gene H. Lerner. Selecting next speaker: the context-sensitive operation of a context-free organization. *Language in Society*, 32:177–201, 1998.

[15] I. McCowan, Gatica-Perez D., S. Bengio, and G. Lathoud. Automatic analysis of multimodal group actions in meetings. Technical Report RR. 03-27, IDIAP, Martigny, 2003.

[16] D. Merrill and T. Selker. The attentional mixer, internal tech report, context-aware computing group. Technical report, MIT Media Lab, 2004.

[17] Kazuhiro Otsuka, Yoshinao Takemae, Junji Yamato, and Hiroshi Murase. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of International Conference on Multimodal Interface (ICMI'05)*, pages 191–198, Trento, Italy, 2005.

[18] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.

[19] Stephan Reiter and Gerhard Rigoll. Multimodal meeting event recognition fusing three different types of recognition techiques. Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI), Martigny, June 2004.

[20] Stephan Reiter and Gerhard Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005.

[21] Rutger Rienks, Ronald Poppe, and Dirk Heylen. Differences in head orientation between speakers and listeners: experiments in a virtual environment. *IJHCS*, 2005.

[22] R. Stiefelhagen, J. Yang, and A Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, Vol.13, No. 4, 2002.

[23] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI'02)*, pages 273–280, Pittsburgh, PA, 2002.

[24] Rainer Stiefelhagen and Jie Zhu. Head orientation and gaze direction in meetings. In *Extended abstracts on Human factors in computing systems (CHI'02)*, pages 858–859, Minneapolis, MN, 2002.

[25] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings AAMASŠ02*, pages 15–19, 2002.

[26] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proc. of ICMI*, 2005.

[27] R. Vertegaal. *Look who's talking to whom. Mediating Joint Attention in Multiparty Communication and Collaboration.* PhD thesis, University of Twente, 1998.

[28] R. Vertegaal. Attentive user interfaces. *Communications of the ACM*, Vol. 46(3):33–36, 2003.

[29] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Why conversational agents should catch the eye. In *Extended abstracts on Human factors in computing systems (CHI'00)*, pages 257–258, The Hague, The Netherlands, 2000.

[30] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of the conference on Human factors in computing systems (CHI'02)*, pages 301–308, Seattle, WA, 2002.

[31] V. H. Yngve. On getting a word in edgewise. *Papers from the sixth regional meeting of the Chicago Linguistics Society, Chicago: Chicago Linguistics Society.*, 1970.

[32] Dong Zhang, Daniel Gatica-Perez, Samy Begio, Iain McCowan, and Guillaume Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Workshop on Event Mining in Video (CVPR-EVENT)*, Washington DC, July 2004.