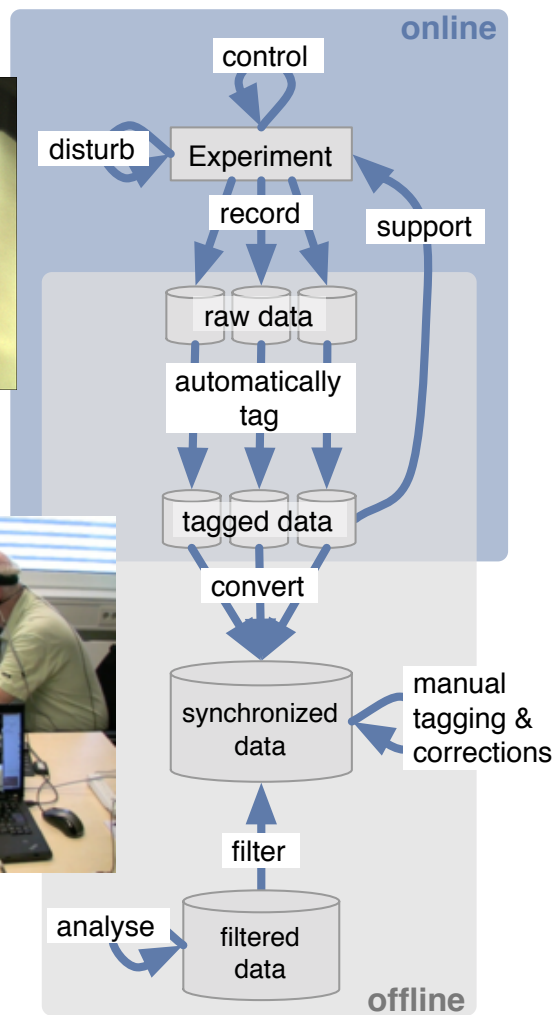


Computer-aided investigation of interaction mediated by an AR-enabled wearable interface

Angelika Dierker



Computer-aided investigation of interaction
mediated by an AR-enabled wearable interface

Der Technischen Fakultät der Universität Bielefeld

zur Erlangung des Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

vorgelegt von

Angelika Dierker

Bielefeld – März 2012

Dipl.-Inform. Angelika Dierker
Angewandte Informatik und Ambient Intelligence
Technische Fakultät
Universität Bielefeld

Gutachter:

Dr.-Ing. Marc Hanheide, University of Lincoln
Dr. rer. nat. Thomas Hermann, Universität Bielefeld

Prüfungsausschuss:

Prof. Dr. Barbara Hammer, Universität Bielefeld
Dr. rer. nat. Thomas Hermann, Universität Bielefeld
Dr.-Ing. Marc Hanheide, University of Lincoln
Dr.-Ing. Kirsten Bergmann, Universität Bielefeld

Abstract

This thesis provides an approach on facilitating the analysis of nonverbal behaviour during human-human interaction. Thereby, much of the work that researchers do starting with experiment control, data acquisition, tagging and finally the analysis of the data is alleviated. For this, software and hardware techniques are used as sensor technology, machine learning, object tracking, data processing, visualisation and Augmented Reality. These are combined into an *Augmented-Reality-enabled Interception Interface (ARbInI)*, a modular wearable interface for two users. The interface mediates the users' interaction thereby intercepting and influencing it.

The ARbInI interface consists of two identical setups of sensors and displays, which are mutually coupled. Combining cameras and microphones with sensors, the system offers to record rich multimodal interaction cues in an efficient way. The recorded data can be analysed online and offline for interaction features (e. g. head gestures in head movements, objects in joint attention, speech times) using integrated machine-learning approaches. The classified features can be tagged in the data.

For a detailed analysis, the recorded multimodal data is transferred automatically into file bundles loadable in a standard annotation tool where the data can be further tagged by hand. For statistic analyses of the complete multimodal corpus, a toolbox for use in a standard statistics program allows to directly import the corpus and to automate the analysis of multimodal and complex relationships between arbitrary data types.

When using the optional multimodal Augmented Reality techniques integrated into ARbInI, the camera records *exactly* what the participant can see and nothing more or less. The following additional advantages can be used during the experiment: (a) the experiment can be controlled by using the auditory or visual displays thereby ensuring controlled experimental conditions, (b) the experiment can be disturbed, thus offering to investigate how problems in interaction are discovered and solved, and (c) the experiment can be enhanced by interactively comprising the behaviour of the user thereby offering to investigate how users cope with novel interaction channels.

This thesis introduces criteria for the design of scenarios in which interaction analysis can benefit from the experimentation interface and presents a set of scenarios. These scenarios are applied in several empirical studies thereby collecting multimodal corpora that particularly include head gestures. The capabilities of computer-aided interaction analysis for the investigation of speech, visual attention and head movements are illustrated on this empirical data.

The effects of the head-mounted display (HMD) are evaluated thoroughly in two studies. The results show that the HMD users need more head movements to achieve the same shift of gaze direction and perform less head gestures with slower velocity and fewer repetitions compared to non-HMD users. From this, a reduced willingness to perform head movements if not necessary can be concluded. Moreover, compensation strategies are established like leaning backwards to enlarge the field of view, and increasing the number of utterances or changing the reference to objects to compensate for the absence of mutual eye contact.

Two studies investigate the interaction while actively inducing misunderstandings. The participants here use compensation strategies like multiple verification questions and arbitrary gaze movements. Additionally, an enhancement method that highlights the visual attention of the interaction partner is evaluated in a search task. The results show a significantly shorter reaction time and fewer errors.

Acknowledgments

Firstly, I would like to thank my supervisor Marc Hanheide for many fruitful discussions and offered opinions, while at the same time leaving always enough room for me to follow my own beliefs. Especially at the end you invested significant time and effort and I wish to thank you for your straightforward assistance in many occasions.

Secondly, I would like to thank Thomas Hermann for providing me with numerous interesting, innovative and novel ideas. I wish to thank you for your time and effort, for letting me work in your group and for the valuable feedback you gave me towards my thesis.

I am very grateful that both project leaders entrusted so much of the project's responsibilities, development and supervision on me. As a result, I was able to contribute my own ideas towards the project and had the opportunity to learn.

Furthermore, I would like to thank:

- Franz Kummert and Gerhard Sagerer for creating such welcoming environments in which to work I enjoyed. In addition, my colleagues in the applied informatics group and the ambient intelligence group for the cooperative working atmosphere.
- My cooperation partners because they helped to increase the projects' capabilities and possibilities.
- Everyone who contributed in whatever way to the ARbInI system, especially Nils Wöhler, David Fleer, Alexander Neumann and Ralph Welsch, who accompanied my research as student workers.
- Everyone who proof-read unfinished text and/or contributed feedback to my work, especially Karola Pitsch, Ulf Großekathöfer, Sascha Griffith, Christian Leichsenring, Anja and Andreas Degenhardt, Andre Dierker, Manja Lohse, Rebecca Förster, Cherrisse Mark and Holger Dierker.
- All study participants without whom my work would not have been possible.

Apart from these, I would like to thank Manja Lohse for her honest and critical feedback and all the fun we had together in our office.

For his work on his thesis and for answering of the reams of questions about his software, special thanks go to Christian Leichsenring. Thank you, for the cooperations, discussions and for being such a good friend.

Generally, I wish to thank all my dear friends for standing by me in spite of my reduced attention in the busy phases. Thank you for luring me away from my work when I needed a break.

Special thanks also go to my parents and family for constantly supporting me and my chosen path, for providing motivation and for forgiving my forgetfulness in remembering birthdays.

And most importantly, I wish to thank my beloved Holger. You gave me so much time, back-up, encouragement and patience that I cannot even put into words but for which I am very, very grateful.

Contents

- 1. Introduction to tasks and processes in experiment control and multimodal analysis** **1**
- 1.1. An example research process 1
- 1.2. Facilitating the research process 2
 - 1.2.1. Recording 2
 - 1.2.2. Tagging 3
 - 1.2.3. Data preparation and conversion 4
 - 1.2.4. Analysis 5
 - 1.2.5. Intermediate summary 5
- 1.3. A closed-loop approach to facilitate experiments 6
 - 1.3.1. Recording 7
 - 1.3.2. Controlling 7
 - 1.3.3. Disturbing 8
 - 1.3.4. Enhancing/Supporting 8
 - 1.3.5. Intermediate summary 8
- 1.4. Research questions 9
- 1.5. Outline 10
- 1.6. Theoretical background 10
 - 1.6.1. Communication, interaction and conversation 10
 - 1.6.2. Selected phenomena in interaction 13
 - 1.6.3. Selected nonverbal behavioural cues 15
 - 1.6.4. Analysis of interaction 18
 - 1.6.5. Augmented Reality (AR) 18
 - 1.6.6. Sonification 20
 - 1.6.7. Abbreviations 21

- 2. The AR-enabled Interception Interface (ARbInI)** **23**
- 2.1. Hardware 24
 - 2.1.1. Video components 24
 - 2.1.2. Audio components 26
 - 2.1.3. Touch components 26
 - 2.1.4. Motion and gesture tracking 27
- 2.2. Software 28
 - 2.2.1. Video data processing 29
 - 2.2.2. Audio data processing 29
 - 2.2.3. Tactile and motion sensor data processing 30

3. Features and Methods of ARbInI	31
3.1. Controlling experiments	32
3.1.1. Component selection	32
3.1.2. AR-based interaction scenarios	32
3.1.3. Controlling the trial process	32
3.2. Disturbing interaction	34
3.3. Recording interaction	36
3.4. Tagging interaction	39
3.5. Enhancing interaction	40
3.5.1. Mediated attention	41
3.6. Conversion: synchronizing and transforming for visualization	43
3.7. Analysing multimodal corpora	48
4. Scenarios	55
4.1. Interaction objects	55
4.1.1. Characteristics & Interaction behaviour of the objects	56
4.2. Criteria for scenarios or tasks	57
4.3. Scenarios and tasks	57
4.3.1. Object games	58
4.3.2. Collaborative and multimodal 3-dimensional data exploration	58
4.3.3. Gaze game	60
4.3.4. Interactive-exhibition design scenario	61
4.3.5. Visual search	65
4.3.6. Animal guessing	65
4.3.7. Smalltalk	66
4.3.8. Prompting by computer	67
4.3.9. Summary	67
4.4. Studies presented in this work	68
5. Side-effects on the interaction	71
5.1. Issues of head-mounted AR and their effects on the wearer	71
5.2. Influence of HMDs on eye and head movements	76
5.2.1. Expectations	76
5.2.2. Method	77
5.2.3. Results	77
5.2.4. Discussion	78
5.3. Influence of HMDs on head movements, speech and task accomplishment	79
5.3.1. Hypotheses	79
5.3.2. Method	81
5.3.3. Results and discussion	90
5.3.4. Summary	107
5.4. Summary and Discussion	108
6. Computer-aided investigation of interaction	113
6.1. Automatic tracking of objects in the field of view	113
6.1.1. Outlook	114

6.2. Automatic annotation of speech times	115
6.2.1. Analysis of the speech data of this work	115
6.2.2. Discussion	116
6.2.3. Outlook	117
6.3. Automatic tagging of head gestures	117
6.3.1. Training and classification	117
6.3.2. Offline evaluation	118
6.3.3. Automatic/Online classification	120
6.3.4. Analysis of relevant axes for detailed analysis of gestures	125
6.3.5. Head gesture corpus	127
6.3.6. Analysis of the corpus data	131
6.3.7. Outlook	132
6.4. Timing of speech and head gesture data	132
6.4.1. Method	133
6.4.2. Outlook	134
6.5. Summary	135
7. Actively influencing interaction with ARbInI	137
7.1. Guiding attention	137
7.1.1. Method	138
7.1.2. Scenario: the “gaze game”	139
7.1.3. Procedure for the study and sample	139
7.1.4. Results	140
7.1.5. Discussion	142
7.2. Disturbing interaction	143
7.2.1. Review of literature	143
7.2.2. Method	146
7.2.3. Initial observations	147
7.2.4. Discussion	149
8. Conclusion	151
A. Appendix	155
A.1. Interactive exhibition design study	155
A.1.1. Questionnaire AR group	155
A.1.2. Questionnaire non-AR group	159
A.1.3. Mapping of markers to interactive exhibits	163
A.2. Enhancing/Supporting: A multimodal display for the focus of attention . . .	164
Todos	167
Bibliography	167

List of Figures

1.1. Schema of an example research process.	2
1.2. Recording	3
1.3. Tagging	4
1.4. Data preparation	4
1.5. Analysis	5
1.6. Intermediate summary	6
1.7. AR features	8
1.8. Schema of the facilitations for an example research process.	9
1.9. Types of communication	11
2.1. Equipment of the participants	25
2.2. AR goggles	25
2.3. Schema: interception & manipulation	26
2.4. Head motion sensors with mounting.	27
2.5. Software setup for the system	28
3.1. Reminder: Schema of the goals developed in Section 1.	31
3.2. Conflicting stimuli	35
3.3. Highlighting of virtual objects	43
3.4. Data conversion tools	45
3.5. Conversion of marker positions	48
3.6. Conversion from logs to ELAN: flowchart of the makefile system	49
3.7. Plotted data file in MATLAB.	51
3.8. Screenshot of MATLAB showing the overlap of two nods in two tiers.	52
3.9. The analysis toolbox	52
4.1. Graspable objects	55
4.2. Examples of virtual objects	56
4.3. 3-dimensional data exploration scenario	59
4.4. Gaze game scenario: marker arrangement and game cycle	61
4.5. Interactive exhibition design scenario: example exhibits and floor-plans	62
4.6. Interactive exhibition design scenario: participants and their views	64
4.7. Visual search scenario: stimuli presentation	65
4.8. Animal guessing scenario	66
5.1. Issues with AR and their effects	72
5.2. HMD generations	73
5.3. Visual search: study conditions	77
5.4. Results: eye-tracking analysis	78

5.5. Design of interactive exhibition: stimuli presentation	83
5.6. Example nod with measure visualisation	85
5.7. Floor-plan with labels	87
5.8. Results: head movement distance covered in space	92
5.9. Results: head movement distance for the AR condition	94
5.10. Questionnaire results: simplicity & comfort	106
5.11. Questionnaire results: naturalness of head movements	107
5.12. Questionnaire results: interference & influence	108
6.1. Tracked objects	115
6.2. Results: Head gesture classification rates	120
6.3. Schema: overlap cases for tags	122
6.4. Position of the MT9 sensor on the head and resulting gyroscope axes	125
6.5. Results: variance for head movement classes	126
6.6. Data file sliced according to one phase of the study and plotted in MATLAB.	133
6.7. Schema: distance calculations for annotations	134
6.8. Screenshot of Matlab plotting overlap between two annotations	134
6.9. Histogram of the measured distances between speech and head gestures.	135
7.1. Gaze game scenario: process overview	139
7.2. Results: search times of “searchers”	141
7.3. Results: Error Rates	143
7.4. Questionnaire results: usage and helpfulness	144
7.5. Results: Body movements of one participant	149
A.1. Mapping of displayed experiments to a set of markers	163

List of Tables

2.1. ARbInI: wearable sensors and displays	24
3.1. Possible methods for the system to interact with its users	34
4.1. Interactive exhibition design scenario: experiments, requirements and sources of interference	63
4.2. Overview of the scenarios in this thesis	68
4.3. Overview of the studies of this thesis and the used system components.	69
5.1. Interactive exhibition design: study conditions	84
5.2. Head gesture annotation: annotators	88
5.3. Results: durations of tasks	91
5.4. Results: head movement distance in space	92
5.5. Results: mean head movement distance	93
5.6. Results: numbers of head gestures per phase	95
5.7. Results: number of head gestures per gesture class	96
5.8. Results: head gesture classification	97
5.9. Results: duration of head gestures	98
5.10. Results: maximum rotational velocity during gesture	99
5.11. Results: number of periods per gesture class	100
5.12. Results: frequency of periods per gesture class	100
5.13. Results: number of utterances	102
6.1. Results: utterance analysis	116
6.2. Results: head gesture classification rates	119
6.3. Results: head gesture classification: confusion matrix	119
6.4. Results: reliability of classification and annotation	123
6.5. Results: most relevant axes per head movement class	127
6.6. Head gesture corpus: animal guessing scenario	129
6.7. Head gesture corpus: interactive exhibition design scenario	130
6.8. Head gesture corpus: artificial head gestures	130
6.9. Results: head movement distances per scenario	131
7.1. Results: disturbances	147
A.1. Translations for the exhibition design study questionnaire – AR group.	156
A.2. Translations for the exhibition design study questionnaire – non-AR group.	160
A.3. Translations for the gaze game study questionnaire.	164

1. Introduction to tasks and processes in experiment control and multimodal analysis

Although there is a long research tradition for the investigation of human-human interaction, there are still numerous open questions that remain to be answered. With today's rapid technological developments, several techniques now are available also for use in interaction analysis. This thesis aims at using such techniques for facilitating the research process. In the course of this thesis, an interface will be presented that allows for a closed-loop control of experiments as well as a framework to facilitate the processing, conversion and analysis of multimodal corpora. For this, computer science methods are applied such as sensor technology, machine learning, object tracking, data processing, visualisation and Augmented Reality. This thesis project is part of the research project *Alignment in AR-based cooperation* (a research project of the CRC 673 *Alignment in Communication*) and several ideas in this thesis have been inspired by Hermann and Sagerer (2005).

This chapter will introduce the goals of this thesis as well as the theoretical background. By means of an example research question, an exemplary research process will be illustrated step by step, show how computer science techniques can contribute to this process and identify those steps where alterations arise because of the contributions. Thereby, Section 1.2 will consider all steps that contribute to the facilitation of the research process and subsequently Section 1.3 will consider the facilitations that can be provided to the experiment by using a closed-loop approach. Section 1.4 summarizes the contributions and presents the research questions regarded in this thesis. Section 1.5 outlines the remaining chapters in this thesis. Finally, in Section 1.6, an introduction will be given to the basic terms that are used in this thesis.

1.1. An example research process

Example A researcher wants to find out which nonverbal signals trigger a listening interaction partner to nod in a dialogue. The researcher presumes features like pauses, voice pitch, eye contact, and head gestures from the speaker to be decisive.

To investigate this, the researcher follows a research procedure from the experiment design to a (verified) hypothesis. This process will be introduced in the following and is also visualized in Figure 1.1 where online steps are depicted with a blue background while offline processes have a grey background.

In our example, the researcher designs a scenario that encourages two participants to show the expected behaviour and conducts an experiment that is suited to answer the research question. During the experiment, the researcher records the interaction of two participants using a

scene camera. These recordings usually are discrete, regularly taken measures (frequency distribution of sound, brightness and colour distribution of video).

After the experiment is finished (thus offline), the sound measures can be used for an analysis of the voice pitch. Additionally, the pauses, nods and eye contact have to be analysed. But the discrete audiovisual measures are not adequate to investigate e. g. nods because these are events taking place in the interaction (the triggered nods performed by the listener but also the presumably triggering nods by the speaker). These events cannot be directly measured in the audiovisual data. Instead, the data have to be prepared and displayed with respect to a timeline (Please note that a 'display' cannot only be visual but also auditory or audio-visual.). Then, the events (nods, eye contact, pauses) have to be found in the displayed data and they have to be tagged manually (a process that is also called coding or annotation) according to an appropriate coding scheme. Once the events are tagged, the researcher can use various statistical methods in order to jointly analyse the events and the pitch measures on the timeline.

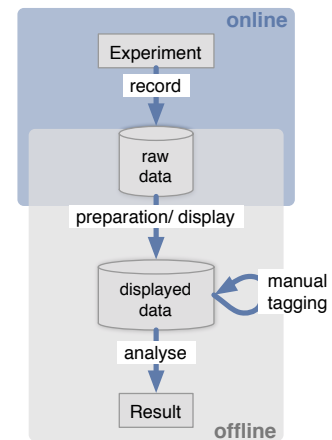


Figure 1.1.: Schema of an example research process.

1.2. Facilitating the research process

The above procedure can be improved at several points with the help of methods from computer science. The following sections will detail the possible improvements in each step (recording, tagging, preparation, analysis) thereby developing step by step a sketch of an *improved* research procedure. The improvements that are possible *during* the experiment will be discussed separately in Section 1.3. Finally, the improved research procedure will be presented completely in Figure 1.8 in Section 1.4.

1.2.1. Recording

In our example, the researcher simply records the interaction audio-visually. For this recording, the researcher has to ensure that *all* perceptions leading to a behavioural event are recorded in order to draw correct conclusions from the data. This can be accomplished by carefully choosing perspective and modalities for the recording.

Perspective In our example, the initial recording happens from the researcher's point of view or, more precisely, from the camera's point of view. No matter how exact the coding, annotation and analysis later are, the data still are recorded from the camera's point of view. But the interaction partners to be monitored, naturally, perceive the world surrounding themselves and their interaction partner's behaviour always from *their* point of view. If we aim at understanding which communicative signals induce what behaviour, we have to take into account the exact perceivable information available for the dialogue partner prior to

the reaction. Otherwise, we are likely to draw incorrect conclusions. Thus, we propose to supplement the scene camera recording with cameras that are mounted on the participants' heads.

Recording the interaction from the participant's point of view particularly enables the researcher to take all external signals into account that have been available for the participant in the preceding time interval. However, internal factors like prior beliefs, knowledge, expectations, etc. naturally cannot be recorded by any of these methods but still have to be inferred.

Multimodality/Immediacy Interaction is a multimodal phenomenon involving a huge variety of signals apart from speech: paraverbal signals (e. g. voice pitch) and non-vocal signals (e. g. (head) gestures, body posture, touch or even olfactory signals). These are transmitted by their respective communication channels (acoustically, visually,...). In our example, the researcher used the video/audio recording for all the analyses. The voice pitch can be directly measured from the recorded sound (bearing our considerations about the recording perspective in mind). But head nods and eye contact can only be extracted from the video in an indirect way since they cannot be easily measured in the video but have to be tagged manually. For a detailed analysis of these signals and their impact on the behaviour, the cues that are deemed to be important should be recorded in a way allowing an efficient analysis of the data.

For this, we have to extend the standard video/audio recording of our researcher from the example so that these behavioural features can be extracted easily. This work's approach is to use (additionally to audio/video) a set of sensors (e. g. one for each feature of behaviour) that are well suited for the recording of these particular features. Together with the audiovisual data, these sensor data can be saved for the later analysis and processed in the following steps. Figure 1.2 sketches this improved recording idea with (as an example) three data types.

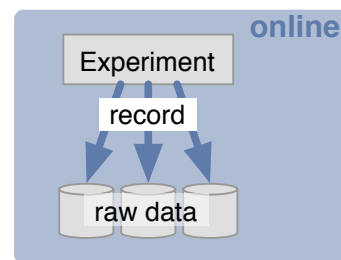


Figure 1.2.: Recording

1.2.2. Tagging

Once the data have been recorded, the researcher can begin with the data tagging. Since the researcher in our example is interested in certain events (head nods, eye contact), these events have to be labelled in the data. This step is – depending on the discipline or the type of the labelled event or information stream – called tagging, coding, annotating, or transcribing. Depending on how many features are to be coded and how precise the transcriptions and annotations have to be, the effort for this phase ranges from costly to extremely costly since the work is particularly demanding and laborious. Like other demanding and laborious tasks, this manual labelling is very likely to be highly error-prone. Moreover, even the best-trained coder using a perfect coding scheme can assign wrong or subjective tags caused by e. g. stress, tiredness or emotions. To remove the resulting coding errors (or at least to be aware of the error rate), the researchers often use a redundant approach so that several different people tag each interaction. However, this greater reliability of the annotations is only obtained at an even higher price of time and effort than before.

Hence, a reliable way to automatically tag behavioural features would be very helpful for quantitative analyses since it would reduce both the amount of work and the complexity of it. Even a method with less-reliable results could accelerate the annotation process: for example, working through the whole interaction is in most cases much more time-consuming than checking the correctness of a set of tags. To enable such an automatic tagging, this work proposes to integrate an additional step into the research process that processes the multimodal sensor data described in the previous step. By applying machine learning methods on the data, they can be analysed for certain features (e. g. head nods, speech pauses) and the resulting classifications can be used as tags. This can be done either online (during the experiment) or offline. Figure 1.3 sketches the automatic tagging where each data type can be analysed for its patterns.

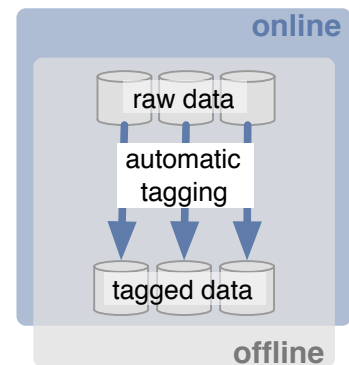


Figure 1.3.: Tagging

1.2.3. Data preparation and conversion

Given that there is a corpus, it is necessary to allow the researcher to efficiently visualize and browse the data after the experiment is finished (offline). Thereby, the researcher can verify, correct and supplement the data with additional (manual) annotations. Since the researcher in our example has now not only video/audio data but also other data at his or her disposal, all available data has to be visualized: the video and audio data, the original data from the sensors, and the classification results. But the data are saved in different file types and storage formats and have to be transformed and synchronized prior to the auditory/visual presentation. This is usually associated with a considerable effort.

The approach of this thesis is to automatically transform, temporally align (synchronize) and prepare all the data: video, audio and sensor data as well as the classification hypotheses (tags) provided by the previous step. The whole multimodal corpus can then be easily presented in one of the existing tools that allow browsing such data.

By adding manual annotations that further describe the participants' behaviour, the researcher gains a comprehensive representation of the interaction situation. Corrections that are undertaken for the classification hypotheses can be used for the re-training of the classification modules of the previous step. Figure 1.4 sketches the automatic transformation and synchronization as well as the manual work for the researcher.

Please note that the automatic tagging described in the previous section is scheduled before the preparation step while the manual annotations are here introduced after the preparation step. This is true since the data preparation and display is *necessary* for the manual annotations while the automatic tagging can already happen during the experiment (online).

This will be crucial for the support feature described in Section 1.3.

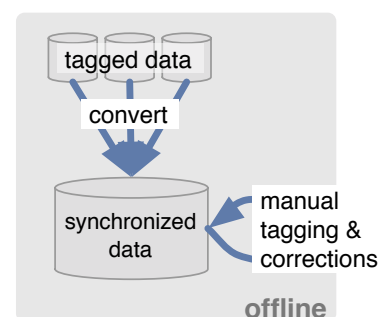


Figure 1.4.: Data preparation

1.2.4. Analysis

The main goal of the previous steps is to prepare the data for analysis. In our example, the researcher believed pauses, pitch, eye contact and nods from the speaker to be possibly triggering nods in the listening interaction partner. In a first step, the researcher might want to count all occurrences of nods and measure the durations of the eye contact phases in order to gain an overview on the data. These measures might then be transferred (that means exported and subsequently imported) to another program in order to apply appropriate statistical methods. Here, numerous established qualitative and quantitative methods would be available. But since the researcher is then working not any more on the corpus, the counting, the transfer and each statistical analysis has to be repeated whenever the corpus is changed (for example for corrections). Moreover, if the process of counting, transferring and analysing would be performed manually, the researcher could not apply the same process to similar research questions (e.g. number of eye contact, durations of speech pauses) in an automatic way.

Thus, it would be much more convenient, to enable the statistics program to import the original corpus directly without having to count the features and export them beforehand. By choosing statistics programs that support working with the command line, we can provide means to script and thus automate the whole process: the counting/measuring, the transfer and the analysis. These command line scripts can then be easily applied to similar analyses thus reducing the work for the researcher substantially.

In a second step, the researcher might want to count (or filter) all nods of the listening interaction partner where the speaker nodded in a certain time range beforehand and where the interaction partners had eye contact beforehand. Additionally, the duration of pauses prior to a listener nod or the voice pitch have to be compared with cases where the listener did not nod. These are complex relationships between more than one coding or data type that are not always easy to detect in the data.

Although such complex relationships could be filtered by hand, the work is much more convenient if the researcher is able to define rules for the phenomena to be investigated. By using the command line, this thesis provides a tool to easily filter complex relationships between the different data types.

In summary, this approach proposes to work directly on the corpus during the analysis (see Figure 1.5) thereby providing techniques that allow scripting the process. Furthermore, they enable the analysis of complex relationships between several data types.

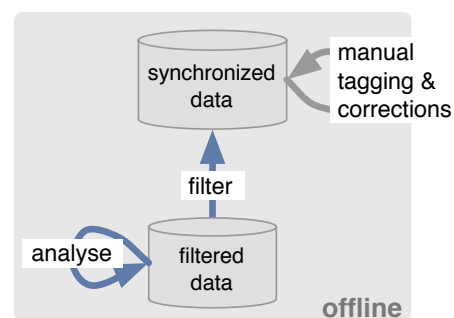


Figure 1.5.: Analysis

1.2.5. Intermediate summary

Let us summarize the goals identified so far (please refer also to Figure 1.6 for a visualization of this summary). The aim of this thesis is to facilitate the analysis of human-human interaction and thereby:

- to enable the multimodal recording of as many of the transmitted signals as possible using a combination of audiovisual recording and sensors (fitted for a direct recording of specific signals)
- to facilitate the tagging of behavioural features in the resulting multimodal data by the use of automatic classification methods
- to synchronize, transform, and provide the recorded data and tags jointly in a way allowing for an efficient control and (if necessary) correction of the tags
- to allow for an efficient analysis of the multimodal data by directly working on the corpus (instead of transferring it), and providing techniques that allow scripting the filtering and the analysis of the data (also enabling complex relationships between several data types).

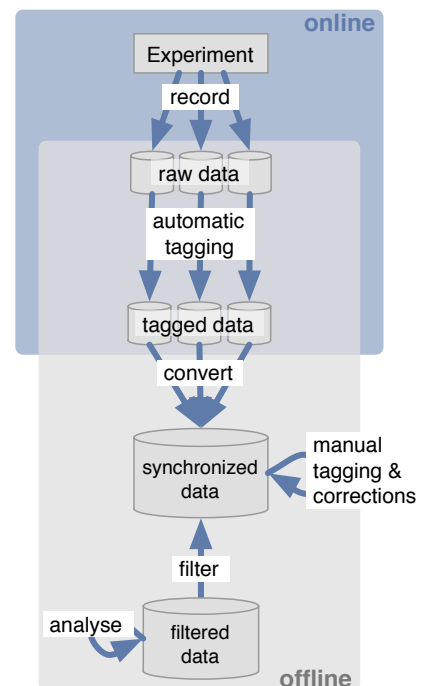


Figure 1.6.: Intermediate summary

The recording of the interaction should thereby happen from the participants' points of view to allow the researcher to put himself/herself in both participants' positions and thus take all perceivable signals into account. To record from the participants' perspective, we propose to use wearable devices: small cameras, microphones and sensors attached to the head. Scene cameras can be used to provide an additional overview or a close view of specific details.

1.3. A closed-loop approach to facilitate experiments

While the previous section described the possible facilitations for the research process from the recording up to the analysis of the data, this section will cover facilitations that are possible *during the experiment*. These additional facilitations are achieved by one alteration: the use of multimodal Augmented Reality. Augmented Reality (AR) is a technique that allows overlaying the reality (the normal sensory stimuli) with additional information (see also Section 1.6.5). The added information is usually virtual and computer-generated and can apply to all senses. For example, a visual scene can be augmented by visual hints; an auditory scene can be augmented by speech or sound.

To perceive the virtual information visually, a user needs a display, for example a head-mounted display¹ that consists of front-mounted cameras and two near-to-eye displays. These devices can be used to de-couple the user from the visual outside world, to interpose some processing (e.g. add the virtual information "augmentation") and then to re-couple by means of the device.

Apart from the visual domain, it is also possible to use AR in the auditory domain. Using microphone head-sets and closed head phones, the users are, analogue to visual AR, de-

¹More precisely, we are referring to video see-through head-mounted displays. See Section 2.1 for technical details.

coupled from the outside soundscape, the sounds can be processed (virtual information can be added) and then are provided again to the users by means of the head phones (re-coupling).

The following four sections will introduce how our goal, the facilitation of interaction analysis, can benefit from this technique and which processing possibilities are available between the de-coupling and the re-coupling. Visual as well as auditory techniques that help to record, control, disturb or support the experiments will be proposed.

1.3.1. Recording

While a head-mounted *camera* can only record approximately what the wearer can see, it might also record signals that actually could not be seen by the wearer. Even more importantly, the camera might fail to record signals that were yet perceived by the wearer. On the other hand, a camera integrated in a head-mounted *display* records everything that the user can see: once the wearer only sees the video on the displays we can determine exactly what he or she can visually perceive (if we assume we block the peripheral perception by blinders). Analogue to the video camera, the video stream provided by the head-mounted displays can be analysed for patterns online and tagged as already described in Section 1.2.2. This again helps the researcher to determine which communicative signals could have been perceived by the user prior to her or his reaction. Of course, not every visual stimulus that is recorded by either of the devices is actually perceived: the human eye and brain applies complex filtering to each image.

Using microphone headsets, we can record directly what each user speaks. By providing the sound and speech to closed headphones, we ensure that the users can hear all the sounds that we want them to hear while simultaneously reducing the outside sounds.

1.3.2. Controlling

The decoupling of the wearer from the visual world allows us to modify the video stream received from the cameras before feeding it to the head-mounted displays. For example, it is possible to give feedback to the interaction partners about the status of the interaction and gain feedback from them in return. By this, we can control the experiment without the presence of the experimenter. This helps to ensure comparable experimental conditions for all participants.

Moreover, the modification offers a whole set of new possibilities for the repertoire of experimental scenarios. For example, we can overlay the (real world) video stream with virtual information using AR. These virtual information can be textual hints, icons or virtual 3D objects and can be either displayed at a certain position in the field of view or displayed on top of markers that are tracked in the scene. When displayed at the position of a marker, the virtual objects are then solidly connected to a physical object (the marker that is e.g. attached to a wooden cube). Two users would both see the same objects (from different view angles). Taking the cubes into their hands, users of the AR system can inspect the objects from different sides or arrange several objects on a table. Virtual objects cannot only be seen but might also be heard when the user interacts with them: objects might emit sound. An easy task could be to arrange the objects according to their size, colour or shape. Using this

kind of objects provides a rich basis for AR-based games ranging from easy to arbitrarily complex tasks fitted to the research question to answer.

1.3.3. Disturbing

The AR-based games explained above also allow showing inconsistent perceptions for both users: For example, for one user one or more objects could be displayed with a different colour, size or type. To disturb the conversation, we can furthermore distort or delay the sound signals that are provided to the users. These inconsistencies and disturbances can lead to misunderstandings during the interaction of both users. More particularly, such a technique could even actively *induce* conflicts, therefore offering a huge variety of ways. This can help to investigate how the participants find a solution to the given task despite of conflicts and how the participants repair misunderstandings. Using normal (real) objects, such manipulations would be physically impossible.

1.3.4. Enhancing/Supporting

Apart from disturbing the interaction between the two users in the just mentioned way, we can also try to actively *enhance* the interaction by adding information into the field of view that would otherwise be not available to the user. Thereby, we could also *support* the task accomplishment. Examples would be to display textual hints about the objects currently in the field of view or to highlight the objects that are currently in the partner's field of view. We can also add sounds that would otherwise not be available to the users. For example, to support the conversation we can provide additional information (e. g. spoken hints, sounds that help to solve the task).

1.3.5. Intermediate summary

As an intermediate summary, the features provided by AR are sketched in Figure 1.7. Thereby, multimodal techniques are used, namely visual as well as auditory. AR newly provides three features that were described above: controlling, disturbing and supporting the experiment. Additionally, it enhances the recording that, however, is also possible without AR as shown in Section 1.2.1 but is enhanced significantly by the technique. In order to allow an interactive influence on the experiment, all these closed-loop features are applied during the experiment, thus online.

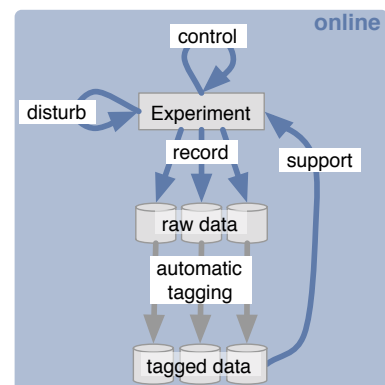


Figure 1.7.: AR features

1.4. Research questions

In the beginning of this chapter, Figure 1.1 was introduced as an example research process. We argued that this process could be facilitated at several steps. The subsequent sections proposed several modifications to this process which were sketched step by step in the Figures 1.2-1.7. Concatenated, these steps result in the new research process that is sketched in Figure 1.8. This process is thereby divided into online steps (that are applied during the experiment) and offline steps (that are applied after the experiment is finished). As the figure shows, the tagging step can be performed either online or offline.

Together, this thesis aims at facilitating experiments with a wearable closed-loop interface. For this, a modular interface is developed that allows intercepting the interaction thereby monitoring and recording the transmitted signals. Furthermore, the interface offers optional audiovisual AR techniques with which the interaction can be controlled and manipulated. The interface presented in this thesis is called the **Augmented-Reality-enabled Interception Interface** (ARbInI). The second aim of the thesis is to support the investigation process of interaction by facilitating several steps on the way from an experiment to the analysis. Tools are provided to assist the tagging, transforming and analysing of the data.

Aided by the ARbInI system described above, this thesis will address the following research questions:

- How does AR-mediated interaction differ from natural interaction? Particularly: How do AR goggles affect the interaction?
- Which behavioural signals are particularly suited to be investigated by the system? What can we learn about the characteristics, duration or timing of those signals?
- How are the AR-based features of the interface able to actively disturb or enhance the interaction? What can we learn about interaction from this?

In order to investigate these research questions, we have to design scenarios that are suitable to evaluate the methods and approaches for the tagging, transformation and analysis of the data as well as the closed-loop features exploiting AR techniques. It is important that these scenarios elicit the respective behaviour in the participants that is to be investigated. This thesis presents a number of scenarios and discusses them with respect to their appropriateness to answer the investigated question. These scenarios are then used for a set of empirical studies that investigate the questions above.

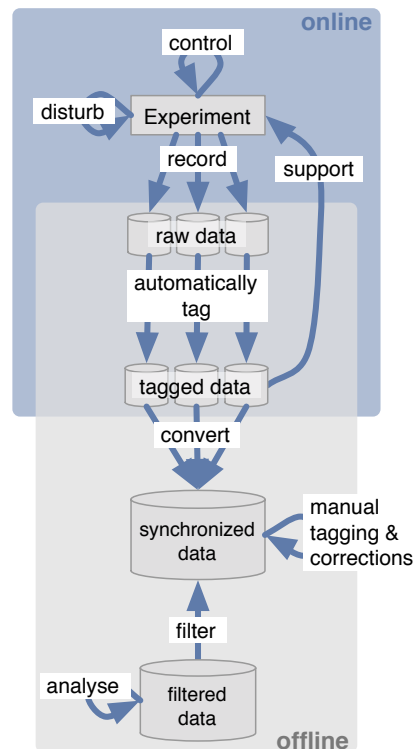


Figure 1.8.: Schema of the facilitations for an example research process.

1.5. Outline

The following Section 1.6 introduces the basic terms that are used in this work. In Chapter 2, the Augmented Reality-enabled Interception Interface will be introduced and its hardware and software components will be described. Chapter 3 will discuss its features and contributions to the goal to facilitate the analysis of interaction as well as the system's ability to modify the interaction in fine detail. Possible scenarios will be discussed that can benefit from ARbInI and Chapter 4 will give a short overview of all studies that are presented in this thesis.

Since the wearable interface presented in this thesis might affect the interaction, Chapter 5 discusses possible side-effects of the used hardware on the AR-mediated interaction. Chapter 6 presents approaches in computer-aided investigation of interaction evaluates the capabilities of the analysis methods by means of behavioural data recorded in different scenarios in this thesis. In Chapter 7, the possibilities of ARbInI to actively influence interaction will be discussed. With the aid of case studies, methods to enhance and methods to disturb the interaction will be discussed and evaluated. Finally, Chapter 8 will draw conclusions to this thesis and give a detailed outlook.

1.6. Theoretical background

This work touches the fields of computer science and linguistics (particularly analysis of interaction). Because of this, there are a couple of terms used in this document that have to be clarified to allow an interdisciplinary understanding. However, this work cannot give a thorough introduction to all terms. Thus, the theoretical background to the most important terms will be introduced only briefly in the following sections. Please refer to the respective cited literature to acquire more detailed insight. The chapter closes with a list of the abbreviations that are used for this thesis.

1.6.1. Communication, interaction and conversation

This section will give a distinction of the phenomena communication, interaction and conversation how they are used in this work. Additionally, key features of nonverbal communication in human interaction will be introduced as well as some terms that are used in the analysis of interaction.

Communication

Communication is an exchange of information between systems. More specifically, O'Sullivan et al. (1983) define it as "a process by which A sends a message to B upon whom it has an effect".² Thereby, the message cannot only be sent by words but also by other means (e. g. gestures, smell) and neither the sending of a message nor the receiving has to be intended (for example if you blush, you often do not intend to blush, or, even if you did not intend

²There are other definitions of communication but a detailed introduction into this topic is not the aim of this work. Please refer to the literature cited in this section.

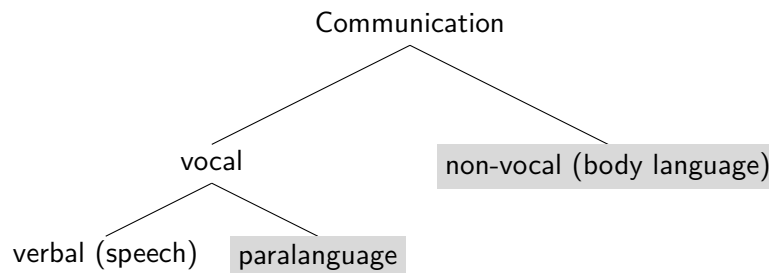


Figure 1.9.: Types of communication. Nonverbal communication consists of both grey-boxed types.

to listen, you might hear the bell ringing). Moreover, it is impossible *not* to communicate (Watzlawick et al., 1971). Even people who say nothing, relax their face and do not move or try not to communicate still communicate (for example that they try not to communicate) (Watson and Hill, 1984). The functions of communication can vary. Watson and Hill (1984) list the following eight functions:

- instrumental (to achieve or obtain something),
- control (to bring someone to behave in a particular way),
- information (to find out or explain something),
- expression (to express feelings),
- social contact (to participate in company),
- relief of being worried (to sort out a problem),
- stimulation (to response to something of interest),
- and role-related (because the situation requires it).

Although these functions seem to suggest that communication is restricted to face-to-face situations, this is not the case. Communication includes also public speeches, writing, mass media (television), advertisements, art, etc. Moreover, communication is not even limited to humans (animals, plants, bacteria, etc. also communicate) (Watson and Hill, 1984).

Forms of communication

There are different types of communication that will be listed in the following. Please refer also to Figure 1.9.

Vocal communication is communication that is produced vocally. It includes not only verbal utterances but also the accompanying paralinguistic information.

Verbal communication refers to the information that is transmitted by words (not including the paralinguistic information).

Nonverbal communication is "communication which takes place other than through words" (Finch, 2002). There are two types of nonverbal communication: vocal/paralinguistic

and non-vocal communication. To the function of non-verbal communication Watson and Hill (1984) state that it “conveys much of what we wish to say and much of what we would wish to withhold”. However, the definition “other than words” has to be seen critically since the verbal and nonverbal behaviour cannot always be separated into distinct categories since there are gestures functioning as words (e. g. sign language) (Knapp and Hall, 2009). Thus, O’Sullivan et al. (1983) proposes that nonverbal communication should not be analysed isolated from speech since both are closely connected. Moreover, the author also notes that nonverbal communication is also influenced by the situation, the content of the conversation and the intimacy between the participants. Please refer to Duncan Jr. (1969) or Ekman and Friesen (1969) for a more thorough introduction to this topic. Please note also that the interpretation of nonverbal communication is culture-specific (e. g. Ekman et al. (1987) Birdwhistell (1970, p. 250ff)). All participants of the studies presented in this thesis were German native speakers.

Paralinguistic communication/ paralanguage means vocal information that is transmitted in parallel to speech that is not words. Mortensen (1972) states that the distinction between linguistic and paralinguistic processes refers to “the difference between what is said compared with how it is said”. O’Sullivan et al. (1983) and Finch (2002) list the following types of paralanguage:

- rate (frequency and regularity of sound, as with a slow hesitant delivery compared with a speedy speech)
- pitch (low-high key intonation, as with bass-soprano)
- timbre (tone and quality of sound)
- volume (intensity of sound, from a whisper to a scream or shout)
- stress (where the words are accentuated)
- giggles, snorts, grunts, sighs.

Non-vocal communication/ bodily communication/ body language refers to communication that is not transmitted vocally. Major sources for non-vocal communication are:

- gaze: eye movement/ contact/ direction (amount of looking at another person's body and face, size of pupil, eyebrow movements) (Kendon and Cook, 1969)
- facial expression (e. g. smiling or grimacing) (Evans, 2003)
- head gesture (Heylen, 2005)
- gesture (e. g. hand movements) and touching Kendon (2004)
- posture (e. g. sitting forwards or backwards), body orientation and body distance (proximity) (Argyle, 1975)
- appearance (skin (including blushing), hair, clothes, smell) (Knapp, 1978)

In sum, communication is a very broad term. But there are more specific terms referring to phenomena as everyday communication, for example the terms ‘interaction’ and ‘conversation’ that will be introduced next.

Interaction

Communication that is happening in face-to-face situations of two or more people is called interaction. The crucial distinction from communication is that interacting people have to be present and that there is a reciprocal reference to each other (Luhmann, 1984; Krause, 2005). Thus, the term covers such communication that happens during greeting, conversation and leave-taking (Knapp and Hall, 2009). During an interaction, the participants can use vocal (verbal and paralinguistic) as well as bodily communication (Goffman, 1959).

The study of interaction addresses norms and strategies of everyday conversation (e. g. turn-taking, see also below) and particularly includes nonverbal communication and social factors in its investigation (Crystal, 2008). Moreover, according to O'Sullivan et al. (1983), the research "should consider not only the present social context but also all of those things that we bring to the situation like role, rule power, socialization, group membership, conformity, motivation, prejudice and perception".

Conversation

Conversation refers to "any spoken interaction (not just informal talk)" (Swann, 2004). This means that conversation is communication that is happening in face-to-face situations of two or more interlocutors and that (in contrast to interaction) language (speech, sign language) is a necessary condition for a conversation.

1.6.2. Selected phenomena in interaction

Following the discrimination above, the face-to-face phenomena that are to be investigated in this thesis are specified within the scope of this work by the term "interaction" rather than with the more broad term "communication" or the speech-focused term "conversation". Throughout the thesis, we will especially investigate nonverbal communication forms, for example in the phenomena of turn-taking, repair or back-channelling behaviour. These phenomena will be introduced shortly in the following.

Turn-taking

Turns are part of the conversational structure: conversation is seen as a "sequence of conversational turns" that is coordinated by rules (Crystal, 2008). The reason for taking turns in speaking is that this reduces the cognitive load of speaker and listener: if we try to speak while listening, the understanding is corrupted (Bonaiuto and Thórisson, 2008). During conversations, people usually leave only very short pauses between the previous and the next turn. Sometimes, they even start before the end of a turn by anticipating from the turn-taking signals (de Kok and Heylen, 2009). Ineffective turn-taking (starting way too early or late) may lead to frustration or to perceiving the interaction partner as rude or dominating (Knapp and Hall, 2009). Sacks et al. (1974) proposed a model for the rules for turn-taking processes focusing on the speech part of the conversation. However, apart from content and syntax, Duncan Jr. and Niederehe (1974) list also paralanguage and body motion to function

as signals for turn-taking. More particularly, the signals are divided into four classes (Knapp and Hall, 2009):

- turn-yielding (signals by the speaker at points where the listener might appropriately act to initiate an exchange of the turn): e. g. gazing at the listener, pitch level, decreased loudness, slowed tempo, extended pause, termination of body movements, relaxing, raising eyebrows
- turn maintaining (signals from the speaker that he/she wants to continue talking): e. g. increased loudness, gesturing, filled pauses, averting gaze
- turn requesting (signals from the listener that he/she wants to start talking): e. g. upraised index finger, audible inhalation, straightening of posture, simultaneous talking (while speaking louder, gesturing and looking away), frequent nodding with verbalizations of agreement “m-hm”
- turn denying (signals from the listener that he/she does not want to talk): e. g. gaze averted, head nods, head shakes, relaxing, silence, smiling, completing a sentence, requests of clarification

Important roles besides the verbal cues, thus, play head movements and gaze (gazing at the interaction partner or away). For example, speakers look to the listener (or the next speaker) and often nod in the end of their turn while they look away and stop nodding when starting a turn (Barkhuysen, 2008). Such signals function also as back-channel behaviour, which will be introduced next.

Back-channel behaviour

This term was first introduced by Yngve (1970, cited in Duncan Jr. and Niederehe (1974)). Back-channel behaviour covers feedback signals from listeners to the speaker. It shows the attentiveness of the listener to the speaker's turn (e. g. by signalling understanding, lack of understanding, agreement or disagreement) (Duncan Jr. and Niederehe, 1974). The signals used for this phenomenon include according to Knapp and Hall (2009) head movements (nods and shakes), verbalizations (e. g. “m-hm” and “yeah”), postural changes, facial expressions, laughter (Vettin and Todt, 2004), sentence completions, requests and restatements. Back-channels often occur without any pause during a turn (Krauss et al., 1977) and are not regarded as an interruption. Instead, by using appropriate feedback, the listener can actively influence the speaker's utterances (e. g. length, content, clarifications). Bavelas et al. (2002) found that the speaker can moreover elicit back-channel behaviour by looking into the face of the listener at specific points in the utterances. The authors asked participants to tell a story, not knowing that their listeners were engaged into a cognitively demanding task and not actually paying attention to the content. With this experiment, they could show that a lack of appropriate listener feedback caused the speakers to tell their stories significantly less well than in the conditions where the listeners did attend to the content (Bavelas et al., 2000). Verbal back-channel behaviour as “ok” or “uh-huh” occurs often with a decrease in pitch over a certain interval (Ward and Tsukahara, 2000). Head movements, particularly nods, seem to be one of the most important listener feedback signals (Knapp and Hall, 2009). Morency (2009) states that some conjunctive words such as “and”, pauses and filled pauses as “um” encourage head gestures as nods since one idea just ended. On the other hand, directly after such a feature, such head gestures are reduced since a new sub-sentence follows.

Repair

Repair mechanisms are used for the interactive editing of mis-functions in dialogues (Furchner, 2009). If such mis-functions in dialogues occur frequently, this is according to Healey and Thirlwell (2002) “not necessarily an indicator of lower communicative coherence”. The authors state that repairs can instead also reflect the complexity of the topic or the efforts of the interlocutors to understand their partner exactly. Repairs can be for example abortions, re-starts, construction modifications, specifications, explanations, re-formulations or word search processes (Furchner, 2009).

Repair mechanisms rely according to Pickering and Garrod (2004) on two processes: firstly, the interaction partners determine that they cannot straightforwardly interpret the information (e. g. in relation to the prior information), and secondly they initiate a reformulation. This mechanism can also be traversed iterative if the reformulation does not lead to successful interpretation. The reformulation itself can consist of three parts: the repairable sequence, an initiation phrase (e. g. “I mean”) and the alteration (Levelt and Cutler, 1983).

Most repairs (e. g. word corrections) are accomplished within the turn where the repairable occurs (self-repairs) (Schegloff et al., 1977). Repairs by other than the speaker (other-repairs) are usually not accomplished until a turn’s completion (Sacks et al., 1974). There can also occur repairs on the turn-transfer itself if it fails (Sacks et al., 1974).

It is, however, important for the speaker to convey meta-information that a repair is about to occur in order to enable the listener to assign the repair to the correct repairable sequence. Speakers use many non-lexical speech perturbations to signal the possibility of a following repair-initiation like cut-offs, sound stretches, ‘uh’s, etc (Schegloff et al., 1977). Additionally, they use for those signals intonation (Heeman and Allen, 1999), pauses (Jefferson, 1974; Nakatani and Hirschberg, 1993), gaze Goodwin (1980), hand gestures (Chen et al., 2002) and head gestures, particularly head shakes (McClave, 2000; Kendon, 2002). When the repair is initiated by other than the speaker, it can take place verbally with questions categorised according to their specificity: vague questions e. g. ‘huh?’, question words e. g. ‘who, where, when?’, partial repeats plus a question word, and partial repeats of the repairable turn (Schegloff et al., 1977). Additionally, the listener can initiate a repair by delayed back-channelling behaviour that conveys hesitations or questions (e. g. questioning facial expressions, tilted head).

1.6.3. Selected nonverbal behavioural cues

As described above, there are many channels for nonverbal communication. Since this thesis especially considers head movements and eye movements, these will be introduced in more detail below. However, this does not indicate that these are the most important nonverbal cues. Rather, they are important in the interaction phenomena detailed above and are particularly suited to be investigated by the described wearable system.

Head movements

Human head movements are an important cue for interaction. They are performed by speakers as well as by listeners. They signal semantic information (agreement, disagreement), express the mood, emotions or mental load of the performer, are related to internal goals or attitudes of the performer or help to manage the conversational process (see turn-taking above) (Poggi et al., 2010; Heylen, 2005). Particularly, head movements are also used to emphasize (stress or underline) speech (Bull and Connelly, 1985) and are then often synchronised with speech (pitch, loudness, rhythm) (Graf et al., 2002; Hadar et al., 1983). In fact, speakers move their heads almost always during speech while keeping their heads more or less still when listening or during pauses (Hadar et al., 1983). Munhall et al. (2004) could even show that head movements improve the perception of syllables in the Japanese language. However, compared to gestures and facial expressions, head movements have received far less attention in the research community (Heylen, 2005).

There are different kinds of movements that occur frequently during interaction. However, the nomenclature is not fixed in the literature. While there is agreement concerning the names about the head nod and head shake, for other frequent movements various names occur. Moreover, the interpretation of these gestures is culture-specific (Darwin et al., 2002). The following list describes the most frequent head gestures as we will call them in this thesis and their assumed semantics:

Head nod up-and-down (or more precisely forward) movement of the head that can be single or multiple. Is often accompanied with “yes” or “I see” (Poggi et al., 2010).

Head shake a left-and-right movement of the head, single or multiple. Is often accompanied with “no” (Kendon, 2002).

Head tilt tilting the head (around the nose) in the direction of one shoulder while still looking ahead. Can be single or multiple and communicates doubt or hesitation while being often accompanied with “well”, “hmm...” or “that depends” (e.g. DeCarlo et al. (2004); Lee and Marsella (2006); Heylen et al. (2007)). In other work this movement is also called “waggle” (e.g. Cerrato and Skhiri (2003)).

Side-way look a single rotation of the head left or right. In other work, this movement is also called “(side-way) turn” (e.g. Cerrato and Skhiri (2003); DeCarlo et al. (2004)). However, in this thesis, we call this movement “look” since this movement is in our scenario also performed while looking on a table. We believe that the name “look” better describes the movement if the position of the head when starting the movement is not specified.

Jerk backward movement of the head which is usually single (Cerrato and Skhiri, 2003). DeCarlo et al. (2004) distinguishes here backward movements of the head as well as forward movements and presumes that the former signals that one is taken aback while the latter signals to take a closer look on something. Although this movement occurs frequently, it is not considered in this work since it is neither cyclical nor conflicting with one of the considered gestures (see Section 6.3).

The most important or frequent feedback head gesture seem to be head nods that can be produced either as single or repeated gesture (Cerrato and Skhiri, 2003; Allwood and Cerrato,

2003). Head nods are used to emphasize speech and to signal agreement. People nod very often during conversation and usually much more often than they use other feedback gestures, particularly head shakes (Hadar et al., 1985; Allwood and Cerrato, 2003). People even nod if they know that their interaction partner cannot perceive this nod, as for example on the telephone (Argyle, 1988) or during interaction with robots (Sidner et al., 2005; Lohse, 2010). When used as backchannel, the nod often precedes the vocal back-channel “uh-huh” (Dittmann and Llewellyn, 1968).

With today’s developments on embodied virtual agents and robots, head gestures are implemented in these systems in order to allow for a smooth and multimodal interaction with agents and robots. For this, the robots and agents need to perform appropriate head movements at appropriate points in an interaction (e. g. Cassell et al. (1994); Morency et al. (2002); Bui et al. (2004)). Moreover, with the beginning of the systems to use such gestures, the interacting humans also expect the system to be able to interpret head gestures (Pitsch, 2010). Thus, it is important to equip the systems also in recognizing such gestures (Morency et al. (2005); Morency (2006)).

Although there are early studies by Hadar et al. (1983, 1985) and later by Graf et al. (2002) that investigate the properties (velocity, circularity) of head gestures, little attention has been paid to these measures apart from this. The automatic analysis of head gestures is the topic of Chapter 5 and Section 6.3. There, we discuss how this cue can be tracked automatically and present evaluations concerning the velocity, repetitions and duration of head gestures under different experimental conditions.

Eye movements

Eye movements contribute largely to the opinions about interaction partners (Argyle et al., 1974; Cook, 1977). Apart from these interpersonal attitudes, gaze signals have also a huge number of other functions in interaction. For example, gaze employs several functions in conversation management (elicit back-channelling, giving back-channelling), it is closely coupled in terms of timing to the speech and helps with turn-taking as we described above. Furthermore, absence of eye contact conveys information about the cognitive processing (speakers look away to concentrate, listeners or speakers show a thinking face (Goodwin, 1987) to signal thinking). Goodwin (1980) noticed speakers to use restarts (see “Repair” below) and pauses in order to secure mutual gaze. Moreover, the gaze direction can also function as a display of the current focus of attention, which will be reviewed more closely in the following.

Focus of attention Determining the visual focus of attention is important in interaction since it is often seen as an “ability that contributes to understanding what another is thinking, feeling and intending to do” (Brooks and Meltzoff, 2005). Usually, attending to the interaction partner’s eyes derives this information. Humans develop early in their live the ability to pay attention to other people’s eyes. For example four month-old infants are able to distinguish between gaze that is directed at them and gaze that is averted (Vecera and Johnson, 1995). Moreover, Hood et al. (1998) reported that already three-month-old infants were able to follow the gaze of adults. They turned their eyes to the direction earlier when adults had

just looked in the direction. In fact, the eye seems to be a special stimulus that allows for a particularly easy extraction of direction information (Langton et al., 2000).

But not only the direction of the eyes is used to determine the visual attention. Brooks and Meltzoff (2005) found evidence for head orientation being combined with eye gaze as visual cue for infants at the age of 10 or 11 months. Langton et al. (2000) states that the eye and head orientation is furthermore combined with information from the body orientation while the orientation of the head makes a large contribution to the estimation of an interaction partner's direction of attention and can even disturb it. In fact, head movements and eye movements mutually affect each other and overlap with their functions in interaction. This mutual influence is discussed by (Heylen, 2006) in greater detail. For further information about the eye gaze following please refer also to Kleinke (1986); Frischen et al. (2007) and Jaimes and Sebe (2007). The automatic tracking of the focus of attention is topic of Section 6.1.

1.6.4. Analysis of interaction

In analysis of interaction, researchers use often several terms that describe how the recorded data are prepared and structured in such a way that they can be analysed:

Corpus A corpus is a term in linguistics that refers to a body of machine-readable text. It allows the evaluation of features and the comparison of results from one study to another (Biber, 1991). Usually, the machine-readable text contains the speech data from a specific study. A multimodal corpus includes other data apart from text (speech), for example information stream of head movements.

Annotation, Tagging and Classification Annotation is the practice of adding interpretative information to a corpus following a coding-scheme *and* the resulting representation of this information. The annotations are saved electronically attached to the original material (Leech, 1993). The term is usually used synonymously to "tagging". However, to distinguish between automatic and hand-made annotations in this thesis, we will use "tags" as the generic term. Tags will include "annotations" which will refer to tags that are added by hand and "classifications" that will refer to tags that are added automatically by a computer program.

Tier A tier includes several annotations that belong to the same characteristics, e. g. one tier includes the speech information, one tier includes the head movements and a third tier includes button presses. Usually, the annotations of a tier are directly linked to a time interval of a master media file (Hellwig and Uytvanck, 2004).

1.6.5. Augmented Reality (AR)

More than forty years ago, Sutherland (1968) presented the first head-mounted three-dimensional display that could be used for superimposing virtual objects on real world images. It was combined with a head position sensor that allowed the user to move the head and changed the view on the virtual objects consistently to the real world image. This technique was later called Augmented Reality (AR) (Caudell and Mizell, 1992).

AR was defined in 1997 as a technique that

“allows the user to see the real world, with virtual objects superimposed upon or composited with the real world. This means that AR supplements reality, rather than completely replacing it. Ideally, it would appear to the user that the virtual and real objects coexisted in the same space” (Azuma et al., 1997).

This definition, however, was focused on the visual sense by using head-mounted displays but later the authors identified further characteristics for AR (Azuma et al., 2001):

- It registers (aligns) real and virtual objects with each other in 3D,
- runs interactively and in real time,
- although most approaches use head-mounted displays, is not limited to this technique, and
- can apply to all senses (seeing, hearing, touch, smell).

Milgram and Kishino (1994) proposed a clear distinction of AR to other techniques that provide virtual information: In their reality-virtuality continuum, AR is distinguished from augmented virtuality. Augmented virtuality is described as a virtual world that is enhanced/augmented by real-world stimuli. Together, AR and augmented virtuality create the group of mixed reality. And mixed reality lies between the two extrema real life (no virtual information) and virtual reality (all stimuli are virtual) (Drascic and Milgram, 1996).

Immersion is a state that is sought by AR applications and techniques. The degree of *immersion* specifies how much the user of AR has an illusion of reality and that the virtual and real world coexist. This characteristic is influenced by many hardware and software decisions and by a consistent appearance and behaviour of the virtual objects. Ideally, the user cannot distinguish between real and virtual objects.

Techniques

Since the first head-mounted display by Sutherland (1968), which was affixed to the ceiling because of its weight (and which the author himself called “relatively crude”), numerous contributions have been made to enhance hardware (and software). Several surveys give a broad overview on the developments (for example (Azuma et al., 1997, 2001; Papagiannakis et al., 2008)). Today, there are different techniques by which visual AR can be perceived (Zhou et al., 2008):

Handheld displays (e. g. portable computers, tablets or mobile phones) are low-cost and very mobile and are, thus, often used for everyday applications. However, their processing powers are limited and the degree of immersion is very low.

Projection-based displays project the virtual object directly onto a surface in the room and are particularly suited for multi-user applications where the users do not need special displays. High processing power is normally available. However, these displays are usually not portable. The degree of immersion should be higher than in handheld AR.

See-through HMDs offer higher degrees of immersion since the world is perceived through them. Three main approaches are available: video see-through HMDs, optical see-

through HMDs and virtual retinal displays. The differences between video see-through and optical see-through HMDs are discussed in Section 5.1. Virtual retinal displays project the virtual information directly onto the eye's retina using small lasers or light-emitting diodes (LEDs) (e. g. Kollin (1993)).

In this work, we use head-mounted display technology. For these, the trend is to develop less obtrusive and thereby less disturbing devices. The long-term goal is, to design AR systems that are as unobtrusive as sunglasses and as ubiquitous as mobile phones (e. g. Papagiannakis et al. (2008)). A new idea, for example, is to design contact lenses that can be used as display. Here, Lingley et al. (2011) recently published a promising first step that managed to integrate one pixel in a contact lens. Extended to a pixel array and combined with an eye-tracking approach (for example integrated into a contact lens as proposed by Kim et al. (2004)), this could be used for the display of AR some day.

Each head-mounted AR system has to include a tracking technology to permit the registration of virtual objects to the real scene so that moving the system results in a consistent movement of the virtual scene. According to Zhou et al. (2008), most approaches use sensor-based, vision-based or hybrid tracking techniques. An easy to integrate and popular approach, though not perfect, is the ARToolKit library that was presented by Kato and Billinghurst (1999). It works with all display technologies that were mentioned above.

Applications

Today, a wide variety of applications for AR techniques are developed, for example medical (Rosenthal et al., 2001), visualisation and teaching (Alves Fernandes and Fernández Sánchez, 2008), tele-presence (Milgram et al., 1997), games (Thomas et al., 2003; Ulbricht and Schmalstieg, 2003), supportive systems in industrial production environments (Feiner et al., 1993; Ong et al., 2008) or personal assistance systems (Wrede et al. (2006)). A typical example for context-related AR systems is an interactive tourist guide as presented by Reitmayr and Schmalstieg (2004) based on the "Studierstube" AR environment.

1.6.6. Sonification

Sonification is a technique that allows providing information via auditory stimuli thereby using the auditory display (Kramer, 1994). The sonification technique is comparable with visualizations, which provide visual stimuli via the visual display. Scaletti (1994) emphasizes that the crucial difference to other disciplines that produce sounds (e. g. music) is the purpose of interpreting, understanding, or communicating characteristics of the data.

Hermann (2008) introduced a thorough definition. The author lists the following four conditions that have to be met in order to call a technique that uses data as input and generates sound signals *sonification*:

- The sound reflects objective properties or relations in the input data.
- The transformation is systematic. This means that there is a precise definition provided of how the data (and optional interactions) cause the sound to change.

- The sonification is reproducible: given the same data and identical interactions (or triggers) the resulting sound has to be structurally identical.
- The system can intentionally be used with different data, and also be used in repetition with the same data.

There are several differences between the auditory sense and the visual sense that have an impact on how the auditory display can be used. Some of the resulting advantages and disadvantages will be mentioned briefly in the following.

The bandwidths of the auditory and visual sense are different. According to Kurzweil (1990, cited in Lange (2005)), the eyes can process 50 million bits per second while the ears can only process 1 million bits per second. On the other hand, the temporal resolution of the eyes is very low: between 20 and 60 stimuli per second (images or flashes), the eyes cannot perceive the single stimuli any more and instead perceive a video or constant light respectively (Miram and Krumwiede, 1985).

According to Baier (2001), the acoustic organ is specialized for the processing of temporal and particularly rhythmical information. Moreover, the perception of auditory data can be eyes-free while drawing attention towards acoustic signals and changes therein. The ears are particularly suited to notice changes in sounds. Thus, the auditory display is particularly useful for monitoring tasks. Meanwhile the listener can visually focus on something else. Moreover, several sonifications can be attended to even in parallel.

There are also disadvantages for the use of the auditory display. For example, it is difficult for listeners to quantify sound characteristics or absolute values, the sound may interfere with speech communication, the sound channel is a very uncommon channel for perceiving data, and constant sounds can become easily annoying. Please refer to Hermann (2002) for a thorough overview of advantages and disadvantages for the auditory and visual displays. A comprehensive overview of sonification is given in Hermann et al. (2011).

1.6.7. Abbreviations

The following abbreviations are used in this work:

- AR** Augmented Reality
- ARbInI** Augmented-Reality-enabled **I**nterception **I**nterface. This interactive *system* consists of two identical wearable *setups* for two participants. Please refer to Section 2 for an introduction to the characteristics and to Section 3 for the features of the interface.
- DOF** degrees of freedom
- id** identification number
- HMD** head-mounted display
- PCA** principal component analysis

2. The AR-enabled Interception Interface (ARbInI)

Section 1 developed requirements for an interface that helps to analyse human-human interaction and listed concrete goals for such an interface. This chapter describes how these goals are achieved with the proposed system for which the basic idea stems from Hermann and Sagerer (2005). Sections 2.1 and 2.2 give an overview of the approach and describe the used hardware and software components.

This work presents the **Augmented-Reality-enabled Interception Interface (ARbInI)** as an approach to reach the goals. ARbInI is a modular framework integrating various hardware components and associated software modules to investigate human-human interaction. The system offers features to **intercept** (control, support, disturb, enhance and record) the interaction that is mediated by it. Since the system includes also components that allow optionally introducing Augmented Reality (AR) features, we call the system **AR-enabled**. Apart from the AR characteristics described in Section 1.6.5 as **registering of real and virtual objects** and **realtime interactivity**, the key characteristics of the ARbInI system are:

modularity The system hardware consists of a modular set of components that can be individually chosen for each experiment. For each hardware component, ARbInI provides an associated software module. Thereby, the system allows choosing the appropriate hardware and software components for each research question that is to be investigated.

two identical wearable setups The system hardware consists of two wearable setups for two participants. (We distinguish between the (ARbInI) *system* which means the complete experimentation interface and the two participant's *setups* which are those parts of ARbInI that are worn by the participants.) Each participant's setup includes a set of identical hardware components, both sensors and displays. These components are connected to a computer that controls the displays and the recording of the sensor data. Each wearable setup can consist of any set of the sensors and displays listed in Table 2.1.

close coupling Each wearable setup is closely coupled to its user: By the use of specific and appropriate sensors for each cue, the system allows for direct and straight recording of the signal thereby reducing the corruption of the signal to a minimum. Moreover, using multimodal AR features, the interface de-couples the user from the outside world, interposes an optional processing step and re-couples the user to the augmented world by means of multimodal displays.

mutual coupling The two wearable setups that are worn by the two participants are mutually coupled so that the ARbInI system can process the signals of both wearable setups online. Particularly, this can be used to transfer information from one setup to the other (and thereby from one wearer to the other wearer).

	Sensors	Displays
visual	head-mounted camera	head-mounted display
	table camera	
	scene camera	
auditory	microphone headset	closed headphones
	scene camera	
touch	buttons (hand-held device)	vibration (hand-held device)
other	head motion sensors	
	head position tracking (eye-tracking)	

Table 2.1.: Sensors and displays that are available for ARbInI. Any set of these devices can be chosen as a setup. Please note that the head-mounted displays are combined with head-mounted cameras in the same device. Thus it is not possible to use the head-mounted display without the head-mounted cameras. The same holds true for the buttons which are integrated in the vibration device and the scene camera which records video and audio at the same time. Moreover, the eye-tracking system is not integrated in software into the system since it was only used for a single study.

2.1. Hardware

The core of our system consists of one portable computer per participant's setup. These computers process all input from the sensors and control the output of the system using ARbInI's multimodal displays. Moreover, the computers control the progress of the experiment and send the data over network to a database on a third computer. The portable computers are furthermore mutually coupled which allows them to integrate information from the interaction partner (see Section 2.2 for details on the software setup). We use two identical Lenovo ThinkPad T61 computers with Intel Core2 Duo 2.20GHz processors and 2GB RAM. They use an nVidia Quadro NVS 140M graphics adapter and an Intel 82801H chipset with integrated audio.

The participants wear a set of devices: a microphone, headphones and an inertial sensor. Figure 2.1a shows an example equipment for one participant and Figure 2.3 shows a schema describing the sensors and displays in human-human interaction. We will cover all components of the setups in detail below, starting with the components concerning the visual input. The software will be described separately in Section 2.2.

2.1.1. Video components

For the *video input* to our system we use the right FireWire CMOS camera (PointGrey Firefly MV color¹) of the stereoscopic head-mounted display (HMD) that provides a 640×480 pixels video stream with 60 frames per second. The camera is integrated (together with a second, identical camera) in a head-mounted display (Trivisio ARvision 3D², see Figure 2.2).

¹<http://www.ptgrey.com/products/fireflymv/>

²http://www.trivisio.com/tech_ARvision3DHMD.html

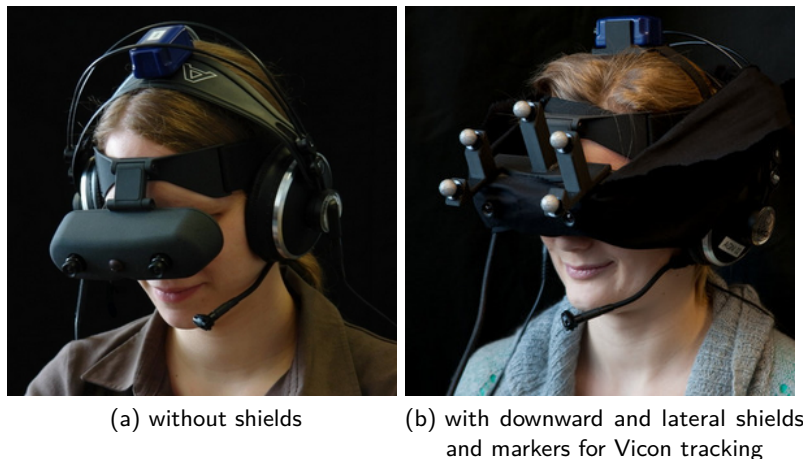


Figure 2.1.: Example equipment of the participants. The participants wear AR goggles, headphones, microphones and inertial sensors. The setup in 2.1b additionally shows black cloth attached to the goggles to shield the user from external visual input as well as passive markers for head position tracking. Photos by team members Till Bovermann and Christian Leichsenring.

AR-goggles: a combination of head-mounted cameras and displays

According to the manufacturers, the resulting field of view covers 42.2° horizontally and 52° diagonally. The device weighs 220 g and its depth amounts to $65 \pm 3\text{mm}$ depending on the focus setting of the lens. The experimenter guides the participants in self-adjusting the mounting of the goggles to the head, the focus setting of the lens, the inter-pupil distance as well as the distance from the eye to the display.



Figure 2.2.: AR goggles.

The computer processes the FireWire video stream and connects the *output video* stream via VGA to the head-mounted display. Two displays then show the (augmented) mono video stream with a screen resolution of 800×600 pixels to both eyes of the user. Although the Trivisio hardware offers stereo vision, we decided to relay on mono vision since the higher computing requirements and other resulting problems (see Drascic and Milgram (1996) for an overview) would have exceeded the resulting benefits.

It can be necessary for some research questions to shield the user from external visual input that would not be monitored by the system. For this, the participants can optionally be coupled even more closely to their setups. In our case we used visual shields (black cloth attached to the goggles and fastened with a knot at the back of the participant's head, see Figure 2.1b) to suppress the external visual stimuli.

Scene camera(s) For the studies described in this work we used up to five scene cameras. For studies using only one scene camera we used a Canon HV30 HD-Camcorder. In the interactive exhibition design study (see Section 5.3), we used two Canon HV30 HD-Camcorders to capture the full scene and the screens of the computers as well as three Panasonic HDC-TM 700EG-K cameras to capture different views on the table and at the participants.

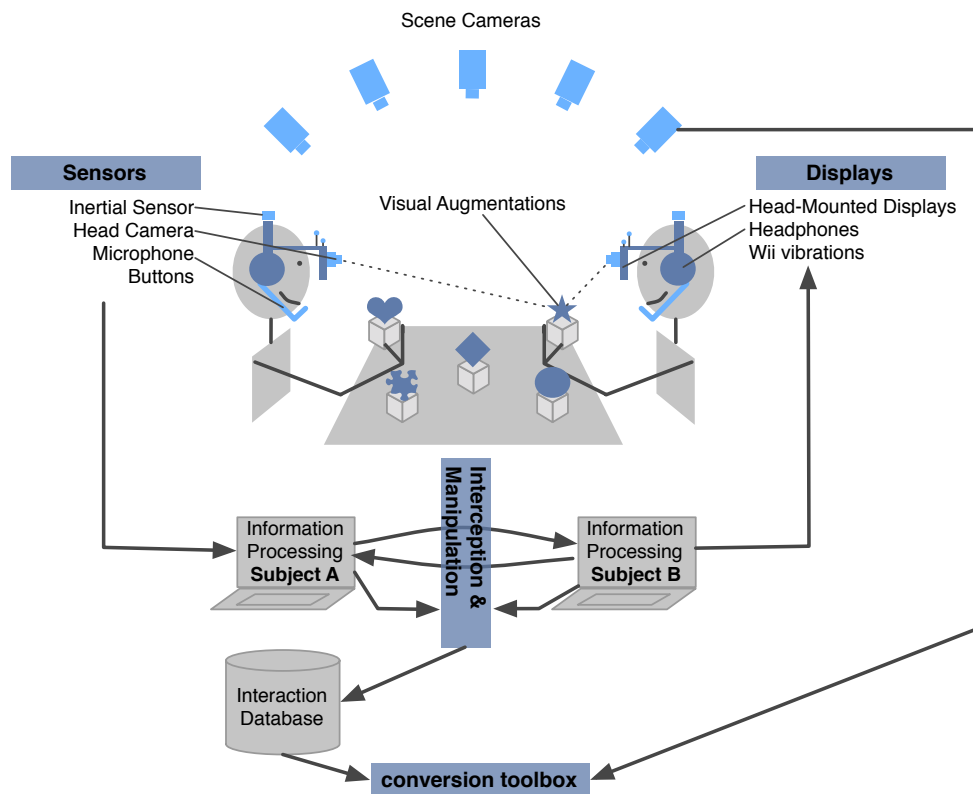


Figure 2.3.: Interception & manipulation. The schema shows a sketch with all integrated components for interception and recording. Derived from a diagram by Thomas Hermann.

Table camera For a recording of the positions and trajectories of the objects on the table we integrated an under-desk camera to the system (PointGrey Firefly MV color).

2.1.2. Audio components

To capture the sound (mostly speech and room sound), the participants wear a microphone headset (AKG MicroMic C520) that is also connected to the computer. The output ((processed) speech, noise and auditory augmentations) can be directed to the participants' headphones (AKG K 271 Studio). We chose closed headphones in order to suppress the wearer from external auditory stimuli. Additionally, all scene cameras record the sound (see above).

2.1.3. Touch components

The participants can be asked to use hand-held devices for the interaction with the system. During some studies, the participants are holding Wii Remotes in their hands in order to control the progress of the task using the buttons. Moreover, the system can provide feedback about the task progress to the participants using vibrations of the Wii Remotes. Thereby, this closes the loop between the user and the system: controlling the trial and giving feedback about the progress of the trial.

2.1.4. Motion and gesture tracking

Inertial sensors For head motion tracking the participants are equipped with a head-mounted inertial sensor (Xsens 3D motion tracker MT9-A/B, outline dimensions: $39 \times 54 \times 28$ mm, weight: 40 g). The motion tracker includes four sensors: The first sensor is an accelerometer (solid state, capacitive readout, 3DOF (degrees of freedom)), the second one a rate-of-turn sensor 'gyroscope' (solid state, tuning fork concept, 3DOF), the third sensor is a magnetometer (thin film magnetoresistive, 3 DOF) and the last sensor is a thermometer. The computer processes the output of the motion tracker and adds timestamps in order to align the data to other data. Since 2010 we also use Wii MotionPlus sensors in parallel to the Xsens hardware to prepare the long-term goal to substitute the Xsens hardware by the Wii MotionPlus sensor. This is planned because of two reasons: (a) The Xsens hardware is over-qualified in terms of the included sensors (we mostly use the gyroscopes) and thus, are bigger/heavier than they have to be. In comparison to this, the Wii MotionPlus sensor is much smaller and more lightweight if removed from the original enclosure and thus, may lead to less obtrusive measurements. (b) The Xsens hardware is already quite old and new sensors are expensive while in comparison the Wii MotionPlus hardware is much more affordable. Both sensors (MT9 and Wii MotionPlus) are mounted to the head with a configurable elastic band and a carrier (see Figure 2.4).

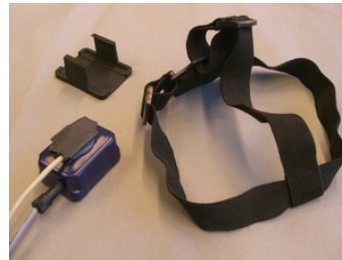


Figure 2.4.: Head motion sensors with mounting.

Head position tracking The HMDs are equipped with infrared reflectors (see Figure 2.1b) that can be used with an outside-in motion capture system e. g. the Vicon motion capture system³. From these data, the line of sight can be reconstructed from the position of the head and its orientation.

Eye-tracking The head-mounted displays are *not* equipped with an eye-tracking technique. A preceding project, the VAMPIRE project⁴, noticed that users wearing an HMD seemed only to use very few eye movements and merely to focus on the middle of their field of view. Specifically, they found that guidance provided in the outer areas was not even noticed by most of the participants (Hanheide, 2006, p. 149). These phenomena can be caused by the restricted field of view or a blurred view in the outer eye area. Other possible reasons could lie in the task as the author discusses. In order to verify these phenomena for the HMDs we used for this work, we combined the HMDs with eye-trackers for a single case study. Section 5.2 covers the study design and its results. In all other cases, we used the HMDs without an additional eye-tracking method. This means that whenever this thesis refers to the field of view, this always means the camera's field of view if not stated otherwise.

³<http://www.vicon.com/>

⁴<http://www.vampire-project.org/>

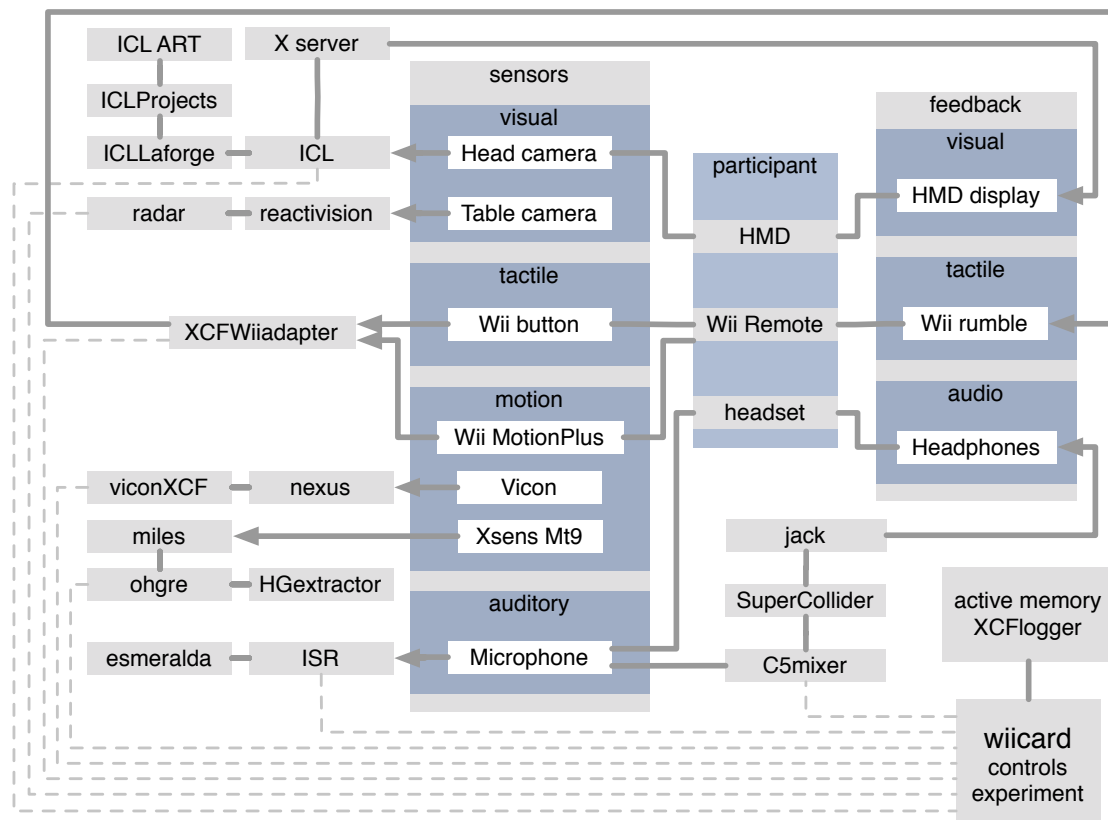


Figure 2.5.: Software setup for the ARbInI online system

2.2. Software

This section explains the software framework of the ARbInI system. The system is a joint effort of the C5 project of the CRC 673 and contains ideas and/or software contributions from several people (alphabetic enumeration which does not represent the amount of work provided by the person): Till Bovermann, Angelika Dierker, Christof Elbrechter, David Fleer, Ulf Großekathöfer, Marc Hanheide, Thomas Hermann, Christian Lang, Christian Leichsenring, Alexander Neumann, Rene Tünnermann, Ralph Welsch and Nils-Christian Wöhler.

Figure 2.5 gives an overview how the software components are connected to the sensors or displays and how the components are linked together. As communication framework we use XCF⁵, a toolkit to build, setup and run distributed systems. We use XCF for publish-subscriber communication between all involved computers: the two computers controlling the sensors and displays and the computer that stores all data (e. g. sensor data, system states) to the database. This communication is accomplished in such a way that all data that are to be exchanged with other computers are converted into XCF messages and then published over the network. The receiving computer can subscribe to these streams. As database for the collected data we use the ACTIVE MEMORY INFRASTRUCTURE (AMI), developed by Wrede et al. (2004b) with a control structure (ACMI) by Spexard et al. (2008). The ACMI

⁵<https://code.ai.techfak.uni-bielefeld.de/trac/xcf>

subscribes on all streams that are to be logged to the database.

The heart of the system is the module called `WIICARD`. This module processes all data and controls the whole framework including the multimodal displays. Specifically, it manages the augmentation of the data by controlling all those software components that process the raw (real) audio and video data, augment it with virtual objects and provide the output to the multimodal display of the user (which is `ICLLAFORGE` for the visual channel and `SUPERCOLLIDER`⁶ and `JACK`⁷ for the auditory channel). Additionally, it controls the progress of the task in a trial. For example in the gaze game task (see Section 4.3.3) the software managed the entire progress of the trial (the correct sequence of conditions and role changes in all 80 cycles). The duty of the experimenter was solely to explain the task to the participants and to start the experiment.

2.2.1. Video data processing

`ICLLAFORGE` controls the visual display and the object tracking. The module is an `ICL-PROJECT` from the Image Component Library (`ICL`)⁸ that processes the image from the head-mounted firewire camera included in the HMD. `ICLLAFORGE` uses `ICLART` (an `ICL`-version of the `ARToolKit`⁹ that handles the firewire cameras) to detect certain black-and-white markers (see Section 4) in the image, maps their marker-ids to the corresponding virtual object (determined by `WIICARD`), overlays the image with the virtual object and directs the image to the `XSERVER` to display. The `XSERVER`¹⁰ embeds the image in the screen image while the embedded part of the image is simultaneously displayed on the HMD. This is especially useful because the experimenter can thus monitor what the participant sees.

A glass table can be equipped with a table camera placed underneath its plate. This can be used for tracking markers on the table that are attached to the objects used for the task. The module `RADAR` is used for processing the table camera stream. It uses the open-source software `REACTIVISION`¹¹ for the marker tracking. Thereby, the object positions and their trajectories can be tracked during interaction.

2.2.2. Audio data processing

The audio stream recorded by the microphone is processed by `ESMERALDA`, which implements a speech recognition (Fink, 1999). The software returns hypotheses of the recognized phonemes. These are converted into `XCF` messages and published.

Moreover, the microphones are connected to the `G5MIXER` written by team member Till Bovermann in `SUPERCOLLIDER` that is able to process the speech, alienate it in a configurable way or add rich sonifications. `SUPERCOLLIDER` and `JACK` are used to provide the auditory augmentations to the headphones.

⁶<http://www.audiosynth.com/>

⁷<http://jackaudio.org/>

⁸<http://www.iclcv.org/>

⁹<http://www.hitl.washington.edu/artoolkit/>

¹⁰<http://www.x.org/>

¹¹<http://reactivision.sourceforge.net/>

2.2.3. Tactile and motion sensor data processing

The XCFWIIADAPTOR controls the Wii Remote and the Wii MotionPlus in performing three different applications: Firstly, the button presses on the Wii Remote are used in WIICARD to control the progress of the trial. Secondly, the vibration function of the Wii Remote is used to give the user feedback about the progress of the task. Finally, the gyroscopes included in the Wii MotionPlus are attached to the user's head and used to record his or her head motion data. All of the mentioned data are published via XCF.

MILES is a module to control the Xsens MT9 inertial sensor and to publish its data. The module configures the format for the data that is provided at the COM port, adds timestamps to it and publishes it via XCF. OHGRE subscribes on this data stream and computes an online head gesture hypothesis that, again, is published via XCF. The HGEXTRACTOR can be used to record head gestures from head motion data. The software asks the participants to nod, shake or tilt their heads repeatedly.

NEXUS¹² is the graphical user interface provided by Vicon to use the Vicon Hardware (a set of cameras and the so-called Giganet). Simultaneously, NEXUS also offers a network interface that provides the real-time data from the system. These data are processed by the VICONXCFADAPTOR that embeds the motion data into XCF messages and publishes them to the network.

¹²<http://www.vicon.com/products/nexus.html>

3. Features and Methods of ARbInI

One purpose of this thesis is to design a software framework that facilitates the research process in the analysis of human-human interaction. Section 1 developed concrete goals for such facilitations. In short, the interface should:

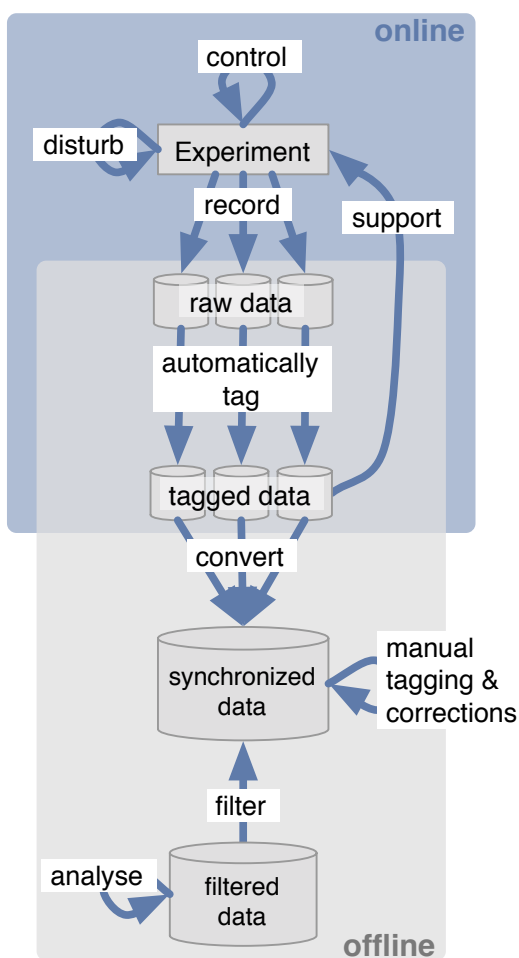


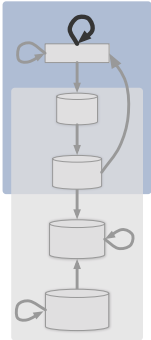
Figure 3.1.: Reminder: Schema of the goals developed in Section 1.

- control the progress of the experiment by giving feedback to the participants (online)
- enable us to actively disturb the interaction of the participants by exploiting AR features (online)
- allow the multimodal recording of as many of the transmitted signals as possible using a combination of audiovisual recording and sensors. The recording of the interaction should thereby take place from both participants' points of view to allow the researcher to take all perceivable signals into account (online)
- facilitate the tagging of behavioural features in the resulting multimodal data by using automatic classification methods (online & offline)
- permit to actively support or enhance the interaction by means of AR (online)
- convert the recorded data and coding hypotheses jointly in such way that allows for efficient display, control and (if necessary) correction of the tags. Moreover it should be possible to manually add further annotations to this interaction corpus (offline)
- allow for an analysis of the multimodal data in an efficient way by directly working on the corpus. (offline)

As indicated in the list of goals (see also Figure 3.1), the features can be divided into two sub-groups: online features, that are available during the experiment and offline features, that are used for the corpus preparation and analysis when the experiment is finished. To reach these goals, we use the closed-loop ARbInI system detailed in the previous section and supplement it with offline tools for conversion and analysis. The following Sections 3.1 – 3.7 will discuss each goal and present the approach of the

ARbInI system to reach it. As a means to keep track of the topic's position in the research process, simplified versions of Figure 3.1 will be displayed whenever a new topic begins. These pictograms are placed in the margin and black arrows will highlight the topic's position in the research process.

3.1. Controlling experiments



This section presents the possibilities of ARbInI to actively control experiments conducted using the system. Firstly, ARbInI offers to select the system's hardware and software components prior to each study. Once the choice of system components has been made, ARbInI secondly offers rich AR-based interaction scenarios and thirdly allows for controlling the experiment.

3.1.1. Component selection

Section 2 introduced modularity as one of ARbInI's characteristics. Prior to an experiment, the experimenter can thereby choose a set of hardware and software components. This modular approach makes sure that only such components are used that are really necessary for answering the present research question. Apart from the research question also other reasons have to be considered for the selection of system components for an experiment: (i) Some sensors or system components are not always available (e. g. components that are shared between research teams). (ii) Additionally, performance issues as processing time and network capacity have to be taken into account. (iii) Some of the system components create moreover a huge amount of data that has to be transferred, stored and analysed. (iv) Apart from this, the load for the participants has to be considered: the more hardware components the users wear the more the system might hinder them in what they do and how they do it. Thus, it is prudent to decide prior to each experiment for a subset of sensors keeping these considerations in mind.

For this modular control about our system components, we use `VDEMO`¹, a script providing a GUI to organize the starting, stopping and logging of arbitrary system components.

3.1.2. AR-based interaction scenarios

Using AR, ARbInI is able to track black-and-white markers in the visual scene. On top of these positions, the system can augment virtual objects that are then anchored to the marker and can be manipulated (e. g. move, rotate, hand over) by both users. Using the component `WIICARD`, the researcher can configure the mapping of the markers to specific virtual objects. This enables a repertoire of experimental scenarios which will be topic of Chapter 4.

3.1.3. Controlling the trial process

Using the close coupling of ARbInI's wearable setups to its users, the researcher can control the process of the experiment. This helps the participants to follow the trial schedule in the

¹https://code.ai.techfak.uni-bielefeld.de/trac/ai/browser/software/vdemo_scripts

intended way. It also allows the experimenter to control the experiment's progress without being present, thereby unintentionally influencing the participants. Our approach here is a closed-loop control that (once started) steers the whole experiment. Two separate steps of interactions with the participants achieve this: giving feedback to the participants about the status of the experiment and getting feedback from the participants about the progress of the experiment. Both will be detailed in the following.

Giving feedback The experiment module `WICARD` includes a component that gives feedback to the participants about the progress of the trial. This feedback for the participants can be given using the visual display of the head-mounted goggles, using sound feedback on the headphones or using a vibration of the Wii Remote that the participants each hold in their hands.

Visual display Apart from displaying the real world image that is captured by the front-mounted camera, the head-mounted display can also be used to augment the image. To achieve this, we register virtual objects to real-world objects (anchors) that become visible every time the real-world object is tracked in the video stream (thus only when the participant looks at the object). Another method is to add information at arbitrary positions on arbitrary time points (thus only when the system is in a specific state). For example, during the gaze game scenario (see Section 4.3.3), the system displays a hint whose turn it is in the task to look for an object and shows a miniature version of the virtual object that is to be focused.

Auditory display Apart from providing the co-participant's speech signals, we can also use the auditory channel for other information using sonification techniques. One idea presented in Section 3.5 is to provide sound feedback when objects enter the field of view and another sound when objects leave the partner's field of view.

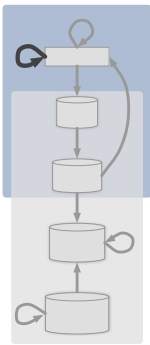
Other displays The participants can be asked to hold Wii Remotes in their hands that can be used as a tangible display, too. For example, in the gaze game scenario, the Wii vibrates whenever it is the turn of the participant to do something (e. g. when he/she can start searching).

Getting feedback On the other hand, the participants control the progress of the experiment by their actions. In the gaze game scenario, we use the buttons of the Wii Remote for this: Every time, the participants finish searching, they press the button so that the next subtask is started by `ARbInI`. Since `ARbInI` also includes speech recognition, marker tracking and head gesture recognition, also other actions apart from button pressing would be possible like the focussing of a specific visual marker for a specific amount of time, verbal commands or a specific gesture. However, apart from the buttons these are not implemented yet. Table 3.1 lists these possibilities both for giving and getting feedback while the non-implemented methods are coloured grey.

Possible interaction methods of the system with its users		
	give feedback	gain feedback
touch	Wii Remote rumble event	Wii Remote buttons
visual	display virtual hints (text, icon or object)	fixation of an object
auditory	sounds (sonification, speech)	speech command (e. g. "next")
gestures	–	head gesture (e. g. "nod")

Table 3.1.: Possible methods for the system to interact with its users (non-implemented methods are coloured grey).

3.2. Disturbing interaction



Why is it our goal to disturb the communication? Section 1.6.2 introduced interaction phenomena as turn-taking, back-channel behaviour and particularly repairs, that are used for the interactive editing of misunderstandings in dialogues. From the researcher's point of view it is extremely interesting to induce misunderstandings that have to be resolved by the interaction partners (e. g. Pickering and Garrod (2004, p.179 f.)). Research questions include "How are such misunderstandings detected by the participants?" and "Which mechanisms do they apply for repair?". The problem is that such misunderstandings do not occur very often. If we do not want to wait for spontaneous ones, how can we induce such misunderstandings?

In traditional experiments (particularly without using AR), it is not possible for the researcher to modify the topic of a dialogue during the experiment without notice of the participants. It is also very difficult or impossible to exchange objects that the participants work with on the table or even let one of the participants perceive a different object than the other. Both dialogue partners always perceive the same object.

Some studies use a confederate/confidante for such purposes. This is a person who seems to be another participant of the experiment but actually is a co-experimenter who is an actor/actress. For example, there are famous studies (especially in behavioural research) as by Milgram (1963) and Schachter (1951) using confederates. However, using a confederate increases the number of experiments, time and effort that has to be invested for the aspired number of participants. And even more importantly, a confederate also increases the influence of confounding variables like experimenter effects as some studies show (e. g. Martin 1970; Narchet et al. 2011).

Video-mediated communication (e. g. O'Malley et al. 1996; O'Conaill et al. 1993; Doherty-Sneddon et al. 1997) generally also allows for alteration possibilities. Here, the dialogue partners each have their own screen that shows the stimuli to them which makes it easy to show different stimuli to them. But, since the two participants use different screens and do not share the same interaction space in such a setup, the participants would not believe that they see the same stimuli (if not from the beginning of the experiment, at the latest from the occurrence of the first misunderstanding). Thus, the participants might question the similarity of the objects more often and would apply explicit questions very early.

Augmented Reality (AR) can also be used to provide contradicting stimuli to the users. Other

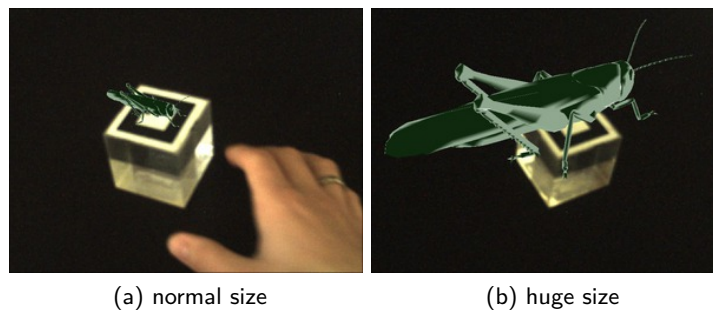


Figure 3.2.: Conflicting stimuli: the virtual grasshopper is shown for participant A normal sized and for participant B huge sized.

than for the video-mediated communication, the participants in AR-mediated communication still share the same interaction space: they can both see the same objects (from different view angles), take them into their hands and hand them to their partner. If the virtual objects are solidly connected to real world objects, the users can interact with the virtual object by manipulating the anchoring real world object. This enables a very natural way of interaction with the objects and might increase the participants' confidence that the virtual objects will follow the same (physical) laws as real objects would. Moreover, from the available AR techniques, the video see-through HMD technique allows for a particularly intense illusion of reality (see Section 5.1 for a discussion of the methods). This illusion of reality and the confidence in the virtual objects, however, are crucial to be able to induce communication conflicts using virtual objects. We believe that otherwise the participants would question the equality of the virtual objects more often and much earlier.

With the help of audiovisual AR, our system enables us thereby to provide contradicting stimuli to the participants while they still share the same interaction space. There are mainly two characteristics, with which the stimuli can be modified in order to disturb the interaction:

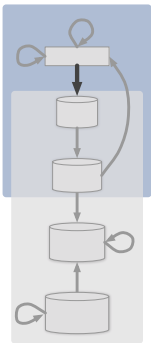
Static characteristics In order to induce misunderstandings, it is possible to use the characteristics of virtual objects that are used for the scenarios in this thesis. We can show all objects identically to both users except one virtual object that may be displayed differently for both participants. Possible differences are altering size or colour of an object or to display a related species, or even a completely different virtual object. These stimuli can be changed at any point in time during the study. For example, in a pre-test where the task was "sort the virtual objects according to their size" (see Section 4.3.1) the users were asked to play with virtual objects augmented on top of real world objects. We implemented a modification of the virtual size for one of the objects (see Figure 3.2) which is presented to one of the participants. Section 7.2 reports about the participants' reactions and their compensation strategies.

Interaction behaviour Apart from static characteristics as size, colour or type of objects, it is also possible to modify the interaction behaviour of the object. We call 'interaction behaviour' everything that can be triggered when the object is in the participant's or the partner's field of view. In a study where the participants learned to use a highlighting of the partner's focus of attention, we disturbed the virtual highlighting itself so that

the highlighting was absent or misleading. Section 7.2 reports about the participants' reactions and their compensation strategies.

Apart from these visual techniques to modify the communication, ARbInI also provides rich techniques to modify the auditory communication channel. When using the full equipment of ARbInI, the users also wear closed headphones to shield them from external auditory input and to provide sound features. The system uses headphones to provide the speech of the interaction partner. In most cases this speech is passed through unmodified but it can optionally be alienated with the C5MIXER (see Section 2.2). This can be used to modify all auditory signals provided by the participant's headphones: the speech signals of the dialogue partner or sound signals provided by the system. All these auditory signals can be distorted, delayed or even replaced by the C5MIXER while simultaneously logging all signals provided to the hearer. As with the visual techniques, the auditory techniques can modify static characteristics as well as the interaction behaviour.

3.3. Recording interaction



In the previous two sections our topic was to control or disturb the behaviour of the participants during the experiment. This section now focuses on the recording of the participants' behaviour. To analyse the behaviour of the participants using ARbInI, our goal is to monitor as many of the signals that are transferred from one participant to another, as possible. Section 1 argued that scene cameras as sole recording devices have crucial disadvantages in terms of the perspective and multimodality of the recording. This will be discussed now in more detail.

Perspective The participants' behaviour during studies is influenced by internal and external factors. Since there is no easy way to record internal factors as prior beliefs, knowledge, expectations, state of mind (DuFon, 2002), these will not be topic of this work. As external factors, the participants' behaviour is also greatly influenced by their perceptions. In order to investigate which perceptions might have triggered a certain reaction from a participant, we have to take all stimuli into account that were perceivable for the participant prior to the reaction. Scene cameras are unsatisfactory here since the researcher cannot always judge if a stimulus was perceivable by the participant or not. This is due to the fact that the scene camera always takes a third-position perspective on the interaction, which means that it shows the interaction not from (one of) the participant's point of view but from another point of view which is from a more or less deviant view angle. This view angle makes it difficult for the researcher to estimate the visual focus of attention or to decide which stimuli were perceivable for the participant. Particularly, if only one scene camera is used, it is usually positioned above or sideways in order to capture both participants and the whole scene. This means that the scene camera records stimuli that are actually not perceivable by the participant and thus might bias the analysis of the researcher. At the same time, there might be stimuli that are not recorded by the scene camera because of its perspective although they influence the behaviour of the participant. This again might bias the analysis of the researcher. Thus, instead of the exclusive use of scene cameras, we also record the interaction from the participant's point of view in order to make sure that *all* stimuli that were perceived by the participant prior to a given reaction are recorded.

Multimodality Interaction is a multimodal phenomenon including vocal signals like speech and voice pitch as well as non-vocal signals like (head) gestures or body posture (see Section 1.6.1 for an introduction). Most of these signals are either transmitted (and thus also perceived) over the auditory or visual channel, and thus would be recorded with an audiovisual recording as it was described above. Nevertheless, we believe that the recording can be furthermore enhanced. Extending the recording with a set of sensors – each fitted for the recording of one specific signal – we are able to record the interaction signals in a way that allows for an efficient subsequent analysis (see Section 3.4).

It is not sufficient, though, to restrict the monitoring to the signals *received* by the participants. Instead, it is also necessary to monitor all signals that are *transmitted* by the interaction partner during interaction. Signals transmitted by other people, the room or (specifically in AR-mediated interaction) signals provided by the system itself (e. g. all virtual stimuli) also have to be monitored for a full pattern of the interaction. Altogether, there are three possible approaches for the recording of the communication signals: recording at the receiver of the signal, recording at the transmitter of the signal and recording somewhere between both participants.

Recording at the receiver of the signal As already argued, a recording at the receiver of the signal is appropriate for the visual cue since we want to record exactly the visual signals that could have been perceived by the participant. For this, ARbInI uses head-mounted video cameras integrated in see-through displays.

It might seem reasonable to apply this also on the auditory signals as well. Technically, it would be possible to attach (for each participant) a microphone with omnidirectional recording close to each ear (e. g. attached to the headphones) in order to sustain the spatial distribution of the sound sources in a way similar to the acoustic perception of the participant. However, at the current state, we record the sound at the transmitter of the signal:

Recording at the transmitter of the signal For sound recording, we use microphone headsets that (mainly) record the vocal cues. One reason is that the most important sounds that are produced in our scenarios are the speech sounds of the co-participant. The interaction objects do not produce particular sounds themselves and if they do, they are played for each participant differently on headphones (as the auditory highlighting described in Section 3.5). The room sounds may even distract the participants. Thus we chose a recording at the transmitter of the signal with microphone headsets combined with headphones in order to shield the participants from the room sound. Besides, another advantage of this transmitter-based recording is that we can apply speech recognition on these data enabling the automatic annotation of speech times per participant (see Section 3.4). As a side-effect, we can easily modify the vocal signal before providing it to the hearer (as described in Section 3.2).

Recording between transmitter and receiver Generally, a recording at the transmitter or at the receiver should be preferred over a recording in between. This is due to the fact that in every signal transmission, some noise is added to the signal (Shannon and Weaver, 1962). Thus, by recording in between, we record neither exactly the signal that was transmitted nor the signal that was received but possibly something else. Nevertheless, a recording in the middle between both participants is still appropriate if

it helps the researcher to obtain an overview of the complete recording situation or a specific view. For example, this is the case with the frequently used scene camera. It records audiovisual signals from a specific perspective thereby allowing the researcher to gain an overview of the interaction. In our approach however, the scene camera provides several behavioural signals, that are recorded also with receiver or transmitter-recording and thus are redundant.

Summary of the recorded data

Sensory data There are visual, auditory, motion data from the sensors as well as button presses:

Visual: At the sensorial site, the camera included in the HMD could be used to save a full video stream. However, this would be very costing in terms of disk space as well as processor- and network load. On the other hand, the computers providing the video stream that is shown on the HMDs, mirror the same video stream on their displays. This enables the experimenter to easily supervise the trial and its progress. Video-taping the computer screens, it is thus possible to record the video data shown on the HMDs during the trial without straining the performance of the computers further.

Auditory: The speech data could technically be saved as binary to the memory, too, but we abstain from this – again because of network and processor performance reasons. Since our focus is not on the speech itself, it is sufficient to use the speech data from the scene cameras for further speech-based analysis, instead.

Motion & User feedback: From the Xsens motion tracker, we save all data it provides and add timestamps to the timeseries. We use the 3 DOF gyroscope data furthermore as input for a head gesture recognition (see Section 6.3). From the Wii MotionPlus we save the 3 DOF time-stamped gyroscope data and use it as alternative input for the head gesture recognition, too. The head gesture hypotheses derived from this recognition is saved to the memory, too.

Apart from the MotionPlus sensor providing gyroscope data, the Wii Remote also provides acceleration data and the user feedback data (button presses). These data are saved to the memory again.

As optional add-on to the ARbInI we used a Vicon system. The data provided by the Vicon system is saved to the memory and can be used to gain the gaze direction of the participants in 3D. This could be useful to enhance the gaze direction display from discrete to continuous display (see Section 6.1 for further information).

System states There are two types of data that are recorded as system states:

Controlling: The controlling feature (see Section 3.1) transmits inter-process communication, configurations (e. g. mapping of markers to virtual objects), synchronization events, experiment progress, and interaction events (button presses, provided feedback signals).

Enhancing/Disturbing: Every time the system tracks one of the markers in the field of view of the HMD camera, it augments the associated virtual object atop this marker. These locations of the markers in the field of view are recorded. Furthermore, it is used to compute a coloured highlighting of those objects for participant A that are currently

in participant B's field of view (see Section 6.1). These highlighting states are recorded. Moreover, the enhancing feature (Section 3.5) and the disturbing feature (Section 3.2) provide data concerning configurations (mapping of markers to virtual objects), visual and acoustic highlighting, and modified characteristics.

Side-effects of recording While recording transmitted and received communication signals, it is important not to disturb the communication itself. This means that the effects of measuring should be minimal so that the system or the participants to be studied may not notice the measurements at best.

For the *system state* data, not only developmental output might be interesting to log. In event-based systems like ARbInI, it is possible to record the system states using the events that are initially designed for the inter-process communication. The ARbInI system uses an xml-based approach for internal processing that was proposed by Wrede et al. (2006, 2004a) and is called XCF. Using XCF, we can easily log the internal information flow without interfering with the system.

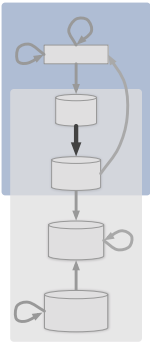
For the *sensory data*, Section 5 will discuss side-effects of some of the head-mounted hardware on the interaction. Studies investigate the influence of the ARbInI system on the interaction and will allow further review on their implications. We argue that – in several cases – the benefits from the method outweigh its influence on the interaction. Apart from this issue, all relevant data is published by the respective modules that have been detailed in Section 2.2 to the network using XCF messages (see also Dierker et al. (2009a)):

Other modules can subscribe to specific tags of the messages so that they can receive data without disturbing the data flow and without having to process *all* data. In order to record all data to create a multimodal corpus, we use the ACTIVE MEMORY INFRASTRUCTURE (AMI), developed by Wrede et al. (2004b) with a modified control structure (called active control memory interface – ACMI) by Spexard et al. (2008). The ACMI subscribes on all publishers that are to be saved to the corpus.

ARbInI integrates a set of classification methods that provide automatic tagging of the sensorial data with hypotheses that will be topic of Chapter 3.4. Since this is handled by XCF and since the resulting hypotheses are sent via network using XCF again, the communication is not affected by this. However, the more processing we do on the data (add augmentations, classifications) the more the lag of the whole system increases. In our case the lag is easily measurable monitoring the provided frames per second (FPS) of the video data. In our experience, it is mainly affected by the number of ARToolKit markers tracked and rendered in the system: for example, for 8 markers, the system provides 22-27 FPS while providing only 15-18FPS for 16 markers tracked and rendered. The measured FPS value varies depending on the angle and the distance from the marker.

3.4. Tagging interaction

The previous section proposed strategies to record interaction signals as efficient and noise-free as possible. This section will briefly introduce possibilities how these signals can be processed further.



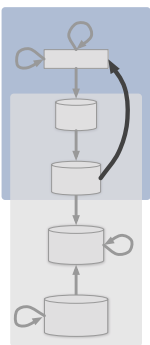
ARbInI integrates a set of classification methods that provide automatic tagging of the sensorial data with hypotheses. As argued in Section 1.6, apart from speech also other signals play an important role in interaction. Head movements and eye gaze are used to a significant extent to structure the interaction, particularly in the management of complex everyday phenomena as turn-taking, back-channelling and repair. Moreover, the focus of attention contains information about the visual attention of human interaction partners. Thus, our automatic tagging attempts to focus on the following signals (which will be further detailed in Chapter 6).

Focus of attention The video data are processed by tracking software that determines the positions of certain markers in the field of view (Mertes, 2008). These marker positions can be used to gain an estimation of the visual focus of attention of the user. This is described in further detail in Section 3.5 and discussed in Section 6.1.

Phoneme classification The data from the microphone is directly processed by a speech recognition software by Fink (1999). The software provides a phoneme hypothesis that is saved to the memory. At present, we use these hypotheses for an analysis of the timing characteristics of the speech (which participant is speaking when).

Head gesture recognition The data from the motion sensor are analysed for sinusoidal patterns with an ordered means models approach (Wöhler, 2009). Please refer to Section 6.3 for a detailed introduction and analysis of this technique.

3.5. Enhancing interaction



The previous section provided automatically tagged data. The aim of this section is to develop methods to use this data for enhancing interaction and supporting the users in their collaborative task-completion. Although there already exist several systems that enable a *collaborative AR* which allows multiple users to share a common mixed reality (e. g. Billinghurst and Kato (2002); Reitmayr and Schmalstieg (2004)), such wearable AR systems have a significant drawback: while they introduce virtual stimuli and information to augment the real world, they also reduce the user's perception of this world. Despite new achievements in hardware design such as high quality see-through displays allowing less intrusive embedding of virtual content in the real world, all devices at least partially shroud the eye area of the wearer. Although this is usually no direct problem for the wearer him- or herself, it has a negative effect in human-human collaboration scenarios. In collaborations, humans can benefit to a significant extent from direct visual eye contact (see also Section 1.6.3). Several interaction-relevant cues require eye contact between the interaction partners; a most prominent one considered in this section is *mutual attention*.

The ARbInI system is suited to particularly *support* collaborative tasks in shared spaces by establishing an AR-based mutual coupling between the two users to facilitate joint attention (see Section 2). In the study we present in Section 7.1, two users have to jointly solve a well-structured task with regard to the manipulation of virtual objects in a real-world table setting. Both are equipped with wearable AR setup and we explicitly enhance their interaction abilities with a multimodal mediation of their mutual foci of (visual) attention. In previous work, for instance by Kalkofen et al. (2007), AR techniques have already been

employed to guide a single user's attention in a context-aware manner. In our work instead, we *closely couple* two AR systems, exploiting one's field of view as contextual information for the augmentation of the other. The following sections will show that this significantly improves the collaboration both quantitatively and qualitatively.

3.5.1. Mediated attention

Attention is a mechanism for the allocation of limited perceptual resources. It means "selecting one event over another" (Baars, 2007). In other words, if we pay attention to something (e. g. a pattern or a sound), we increase the processing accuracy of the respective perceptions while we at the same time decrease the perception of those stimuli that are not included into our attention. This increasing/decreasing has been confirmed on the neuronal level, as Kandel (1996) summarizes: neurons show increased fire rates during attention and their firing rates are reduced in the non-attentive regions. We have to distinguish between stimulus-driven attention (exogenous attention) and goal-driven (endogenous attention) (Wikipedia, 2012; Egeth and Yantis, 1997). In the context of this section, the focus here lies on the stimulus-driven attention since the question is how and if we can achieve that the user of our system attends to the information stimulus that is provided by the system.

Several mechanisms of attention are well studied in the visual domain, for instance using eye-tracking methods in controlled experiments (Koesling, 2003). In the context of human-human and mediated cooperation, attention touches different aspects: (a) the mechanisms used by interlocutors to allocate their perceptual resources (e. g. focus on a visual region of interest or attending a certain signal stream in the soundscape), (b) the methods and signals used by the cooperating users to draw, shape or direct the other's attention. Joint attention can be summarized as an active, bilateral and intentional coupling of attention (Kaplan and Hafner, 2006). It may be assumed that joint attention supports, or even enables cooperation, particularly in the case where the interaction partners' internal representations differ regarding their current context.

Interaction partners use a multitude of strategies to best employ their limited perceptual resources, particularly in cooperation. For instance, we are capable of interpreting other people's focus of attention from observing their head orientation, gaze, and often the body posture and arm gestures (see the theoretical background in Section 1.6.3). Pointing and other indexical gestures are commonly used to guide another person's focus of attention and we often are not aware of the complexity of these mechanisms since they are subconsciously and routinely used.

In the light of these mechanisms, the questions arise how and under which circumstances new forms of technical mediation can actively contribute to support joint attention, thereby forming a kind of *artificial communication channel*? For example, can signals be displayed that accelerate the process of joining attention? In general, since natural communication channels are very good and humans even coevolved phylogenetically with them, as stated above, technical systems need to find niches where they can contribute. We currently see two such niches: (1) the compensation of disadvantages the technology introduced itself if it is necessary to use it, and (2) exploiting the attention bottleneck, e. g. retaining and providing data that the user in principle would have been able to perceive but actually did not because the attention was concentrated somewhere else.

The following sections will present two augmentation strategies that ARbInI uses: firstly, the visual augmentation of elements in the other's field of view, and secondly, the auditory augmentation by means of sonification (see Section 3.5.1.2) depending on whether objects are visible by the partner.

3.5.1.1. Vision-mediated attention

Visual augmentations can manipulate the users' attention in a variety of ways. For selecting the best method it would be necessary to fully understand the mechanisms that guide the user's attention in visual exploration. Since the interplay of these mechanisms differs from situation to situation, a single simple answer may not exist. However, by categorising attention as a mixture of subconscious and conscious processes (such as the explicit searching for certain patterns), we can at least suggest some visual augmentation types.

For instance, a localised adaptation of saliency (e. g. local image filters such as changing the contrast or brightness, or applying low- or high-pass filters) will change the underlying basis for our existing visual attention processes and lead, for instance, to quicker (or slower) detection of the thereby pronounced (or obscured) object. Such techniques that highlight or augment a specific area in the field of view are often referred to as *magic lenses* and have been successfully employed to guide attention in AR (Mendez et al., 2006). Alternatively, temporal patterns such as blinking at certain locations are a strong and salient cue to guide the eyes (or in our case the head), yet such elements need to be used carefully since they may disturb more than they help because they are also strong distractors from otherwise relevant information (Posner and Cohen, 1984). Another type of augmentation effect would be the localised magnification of regions, using a local nonlinear distortion (like a fovea) that lets highlighted image regions cover more space in the field of view of the user.

Such highlightings are interesting – however, they are computationally expensive and radical in the way they break with the user's normal visual perception. We therefore use a more basic yet effective form of visual augmentations that are more easily implemented, offer good control and a good experimental examination of how mediated attention affects AR-based cooperation: we augment grey-coloured virtual objects on top of physical objects using the ARToolKit marker system (Kato and Billingham, 1999) and control the colour (hue) of the virtual objects for one user according to the object locations in the field of view of the *other* user and vice versa. More precisely, the colour changes from yellow (peripheral) to red (in the centre of the partner's field of view) (see Figure 3.3). To enable the system to be useful in situations of temporarily divided attention, e. g. one user looks at an object a moment after his partner has looked away, the colour highlighting fades in and out with configurable times and envelopes. The fade-in is useful to prevent a quick glance or a sweep over an object from letting it after-glow as if the focus of attention had rested on it for a substantial amount of time.

Other types of vision-mediated attention have been suggested and implemented by Mertes et al. (2009), such as the direct indication of the field of view as vision cone or projection onto a surface are conceivable and might be intuitive.



Figure 3.3.: Highlighting of virtual objects according to their position in the partner's field of view. Grey objects are not in the partner's field of view, red objects are in the centre, yellow objects in outer regions of the partner's field of view.

3.5.1.2. Sonification-mediated attention

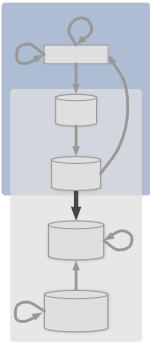
Sonification, the non-speech auditory display of information by using sound is an alternative and complement to visualization (see the introduction in Section 1.6.6). From everyday experience we are quite familiar with the fact that sound often directs our attention, e. g. towards approaching cars when we cross a road, towards somebody calling us, or towards the boiling water kettle in the kitchen. In these situations, our ears guide our eyes and therefore we regard it as an interesting approach to explore the possibilities to use sound to couple the attention of cooperating users.

For this approach of using sonification for interaction support we see various possible methods and thus far have only scratched the surface of what we expect to become a promising field in auditory AR technologies. To implement our ideas of mediated attention via sonification as an additional channel to the visual highlighting explained above, we continue with the same approach of using objects (tracked via ARToolKit) that may be in the interlocutors' field of view (see Section 3.5.1.1). The simplest sonification strategy is to play a short sound event (e. g. a click sound) whenever an object enters the partner's field of view. Using different sounds for entering and leaving (e. g. 'tick' – 'tock'), each user receives some awareness about the overall head activity without the need to look at the partner's head.

So for the study presented in Section 7.1 we use sonification only as a marginal information stream displaying by sound whenever objects enter or leave the partner's field of view. This is because at this stage we are primarily interested in the overall influence of mediated attention on performance in cooperation.

3.6. Conversion: synchronizing and transforming for visualization

The previous sections worked on recording interaction in an appropriate and efficient way, on automatically generating tags to the monitored data and using these classifications for



disturbing and enhancing interaction during the experiment. After the experiment, a researcher may now want to review these classifications in order to analyse or correct them. To allow a comprehensive analysis, this review should take place with respect to the original data. Moreover, since ARbInI creates so many different data types and a huge amount of data, it is very important to find a convenient method for synchronizing all data and transforming all data types into a format that can be jointly displayed. Finally, it should be possible to correct the hypotheses from the classification in an efficient way and it would be nice to allow the researcher to add further (manual) annotations or transcriptions.

These considerations led to the idea to automatize the use of one of the well-established tools for browsing and creation of manual annotations on audiovisual data to jointly display our whole dataset. Since our group² mostly uses ELAN³ for such annotations, we decided for this tool. ELAN is available without fee, is XML-based and ELAN offers rudimentary timeseries support that we can use to display discrete, raw data from sensors. Alternatively to ELAN, other XML-based annotation tools with similar abilities also could be used (e. g. Anvil⁴ or Interact⁵).

How can we import our data into ELAN? The annotations in ELAN are structured in several layers, so-called *tiers* (see Section 1.6.4 for explanations of the words corpus, annotation, tag and tier as they are used in this thesis). We take the scene camera video(s) as the temporal baseline. During the experiment, the ARbInI system creates an audiovisual event (a window opening on a computer screen and a sound that is played) that is recorded in the video and simultaneously logged as an event to the memory. The videos from the scene cameras are aligned using synchronization events (clapper board) and are synchronized then with the memory data using an audiovisual event that is logged to the memory at the same time. Thus, the videos can be synchronized with the other data. As already described in Section 3.3, the augmented video stream that is displayed on the HMDs is shown on one of the scene cameras as well as the sound is captured using the scene cameras.

In order to display all our data, the idea now is to automatically synchronize all data sources (e. g. synchronisation events, system states that determine the current task in the experiment, sensor data, classifications) and then to generate the xml-based ELAN file (.eaf). The solution was strongly inspired by a script from Marc Hanheide, transforming system log events from a robot into ELAN annotations and synchronizing them with the video. This initial approach was rewritten leading to a modular makefile⁶-based system applicable for all sorts of systems. These scripts now transform not only the log events into ELAN annotations but also include the discrete data (time series data) and several configurations. There are different categories of data to be displayed that will be described in the following. For an overview, please refer to Figure 3.4a for a sketch of the makefile system and to Figure 3.4b for a screenshot of a resulting example ELAN file bundle.

audiovisual data The video and audio data can be linked into our ELAN .eaf using the native video support of ELAN.

²<http://aiweb.techfak.uni-bielefeld.de/>

³<http://www.lat-mpi.eu/tools/elan/>

⁴<http://www.anvil-software.de/>

⁵<http://www.mangold-international.com/en/products/interact.html>

⁶<http://www.gnu.org/software/make/>

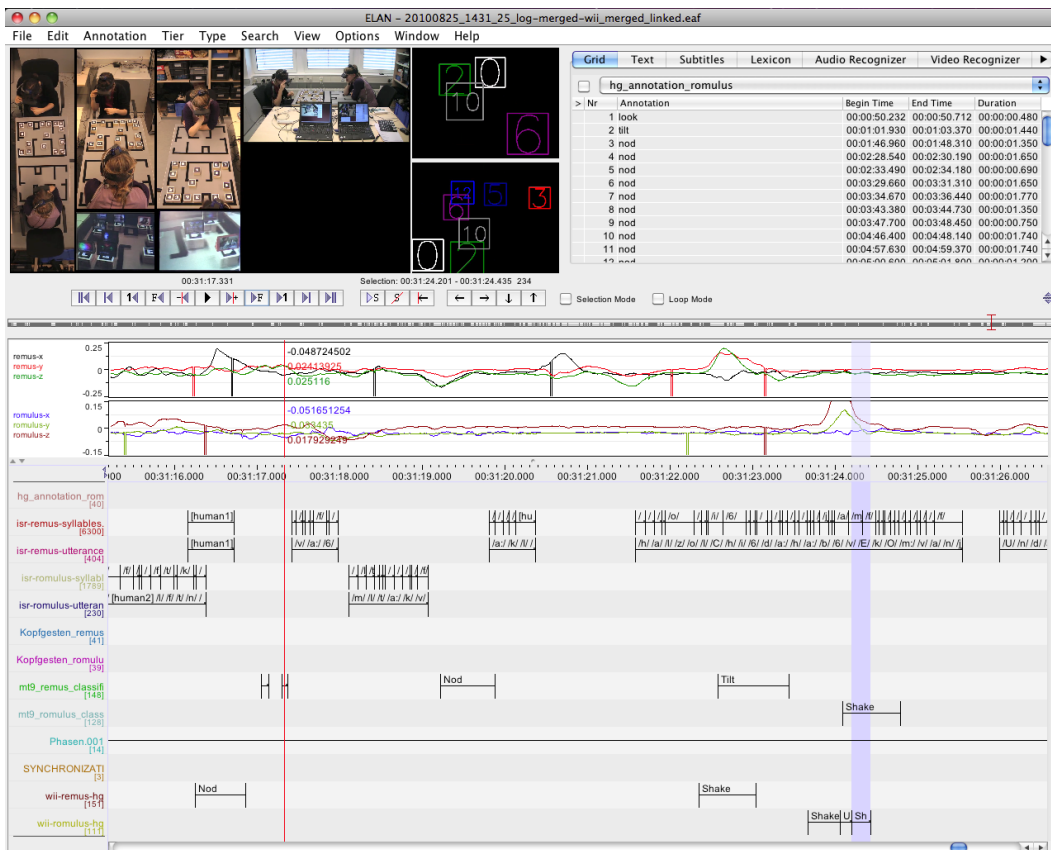
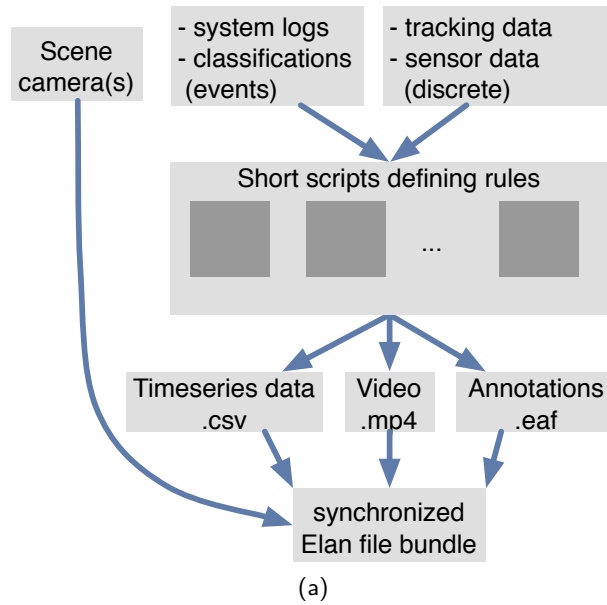


Figure 3.4.: (a) Schema of the conversion tools: in order to apply this method your data you simply have to create a short script defining rules how to display your data in ELAN, adjust the overall makefile, run it and synchronize your scene camera(s) with the resulting ELAN file bundle. (b) Screenshot of ELAN with imported experiment data.

logging events These can be logged system events e. g. the synchronisation events or logged experimental states like the actual task or which participant's turn it is in the task. Such events consist of a timestamp, a type (e. g. "Example Text"), an emitter (the system component that sends this message to the log) and the content (e. g. "Participant A's turn to search") as it is shown in the following XML code.

```
<?xml version="1.0" ?>
<EMITTER_NAME>
  <TIMESTAMP>1282306717896</TIMESTAMP>
  <TEXT>Example Text</TEXT>
</EMITTER_NAME>
```

ELAN annotations are interval-based which means that every annotation consists of a begin timestamp, a duration, an end timestamp and the annotation value. How can these events be transferred to ELAN? We pursue the same strategy that was developed by Marc Hanheide in his script where he transferred logging events to annotations following these rules:

1. single event: we assign a default duration to a single event to make it visible as annotation in ELAN
2. corresponding pair: A pair of two corresponding events represents the beginning and end of an annotation, the duration can be easily calculated.
3. state-change event: An annotation continues until another event of the same type is received. Resulting annotations are directly consecutive.

By this, we can build tiers from each emitter and transfer each event from this emitter into a single ELAN annotation and assign it to the tier. In result, we get a structure of tiers and annotations that can easily be displayed and modified in ELAN.

classifications The new data derived from the automatic tagging/coding methods (see Section 3.4) is discrete data with probability values per timestamp for each possible gesture⁷. Additionally, a resulting classification hypothesis is provided for each timestamp (nod, shake, tilt, nothing), depending on for which gesture type the threshold was under-reached (see the following Example for a classified nod).

```
<?xml version="1.0" ?>
<HYPOTHESIS>
  <GENERATOR>mt9_remus_classification</GENERATOR>
  <MIKRO_TIMESTAMP>1282306712050056</MIKRO_TIMESTAMP>
  <TIMESTAMP>1282306712050</TIMESTAMP>
  <CLASSIFICATION>Nod</CLASSIFICATION>
  <PROBABILITY class="Shake" value="24.0943214687" />
  <PROBABILITY class="Nod" value="11.7300707407" />
  <PROBABILITY class="Tilt" value="19.1744881824" />
</HYPOTHESIS>
```

Although we could display these data in the timeseries window as well, we consider a transformation to annotations more appropriate since we are mainly interested in the final hypothesis than in the exact probabilities for the three gesture types since this

⁷Gestures that are known to the classification system.

is used in the subsequent analysis. This enables us to interpret the data in the same way as data derived from manual annotation and compare these two tagging types with each other, which will be done in Section 6.3.3. In this case, we can interpret the changes from one classification value to the next as a logging event and then apply rule number 3. Since we are interested in successful classifications only, we can furthermore ignore annotations from uncertain classifications (“nothing”), keeping only *nod*, *shake* and *tilt* annotations.

timeseries data In the case of the inertial sensors (see Section 2.1), the raw data includes a timestamp column as well as 3-9 data columns *per participant* depending on which data streams are used from the inertial sensors. These files can be added as *Linked Secondary Files*. In order to display these sensor data in a *TimeSeriesViewer*, all tracks have to be configured: every track (that is a column in the data) has to be named, the data range has to be specified and – in order to allow a distinction between the single tracks – a colour has to be set. Moreover, it is useful to display the tracks in different track panels, one for each participant. Even for a relatively small data file with 4 columns all these configurations would result in a significant amount of work (and quite a number of mouse clicks in the GUI) to configure the whole file. Moreover, this has to be done for every participant pair. To prevent this, the makefile automatically generates the configuration file (-tsconf.xml) for our respective .csv file, thereby setting for each track its name, data range, panel and colour and linking this file into our .eaf file.

tracked objects The coordinates of all tracked ARToolKit markers in the users’ field of view is calculated and saved to the memory (see the following XML code).

```
<?xml version="1.0" encoding="UTF-8"?>
<ARTCOORDS>
  <HYPOTHESIS>
    <GENERATOR>ARTCOORDS_2</GENERATOR>
    <TIMESTAMP>1282307705754</TIMESTAMP>
  </HYPOTHESIS>
  <MARKER id="8" posx="244.910126" posy="-51.781258" posz="
    1156.713379" screenx="0.725830" screeny="0.435985"/>
  <MARKER id="12" posx="268.635681" posy="142.951096" posz="
    884.981689" screenx="0.827198" screeny="0.749081"/>
  <MARKER id="8" posx="-143.296463" posy="173.642563" posz="
    797.160217" screenx="0.294562" screeny="0.836412"/>
</ARTCOORDS>
```

This information is transformed into a video using `FFMPEG`. The resulting video shows the (at that time point) visible marker ids on a black background as shown in the example in Figure 3.5. Section 6.1 will discuss the benefits of this approach.

For each sensor or system component, we developed a configuration script determining rules how the log messages are to be transferred into annotations (or .csv/.tsconf-files). When called by the makefile, this configuration script generates all annotations from the respective log messages and associates them to a tier labelled as the sensor or system component that generated the respective log message. For each tier the makefile generates a single .eaf file. In a final step, the `EAF-MERGER` synchronizes and merges all single .eaf files into one

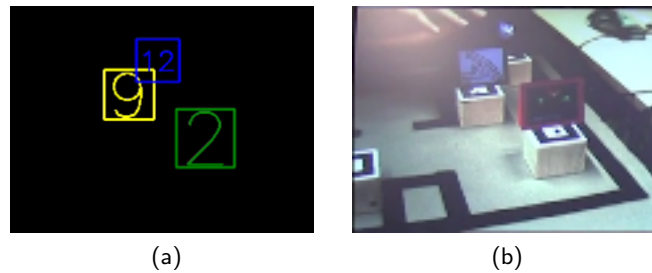


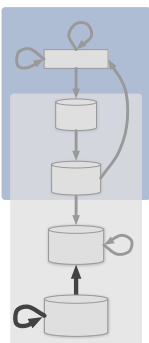
Figure 3.5.: Conversion of marker positions (a) The positions of the marker ids are depicted on black background. (b) Original view through the participant’s HMD.

aggregate .eaf file linking the configuration files. These single and complete .eaf files have the advantage that we can browse the single .eaf files as well as the complete file. This is useful, for example, if the corpus grows too big for an efficient display in ELAN or if some data are corrupted. Then we can still display the smaller single .eaf files and find out what might be wrong. Moreover, the `EAF-MERGER` allows merging any set of .eaf files into aggregate .eaf files, which is particularly useful since ELAN provides to our knowledge no such feature itself. Writing a simple configuration script and adding it to the makefile, we can easily integrate new system components. The makefile handles the merging and synchronisation automatically. Figure 3.6 illustrates the described process from the log file(s) to the complete .eaf file and shows all included scripts.

In conclusion, we developed a way of efficiently generating visualized and browsable corpora by exploiting ARbInI’s recordings as a means of automatic annotation, temporally aligning them with a video and jointly presenting them in an annotation tool. Here, manual annotations can be added, thus allowing a comprehensive analysis. This modular approach has already been applied to other projects⁸, further extending the capabilities of the tool.

3.7. Analysing multimodal corpora

The previous section presented a method to automatically synchronize all data that is provided by ARbInI in the previous steps. Moreover, the data is converted into ELAN file bundles in order to allow for a joint visualization and thereby offering the possibility to browse the data and to correct or add annotations manually. Once all data are in ELAN, all corrections have been made and all manual annotations are added to the corpus, the next step in the research process is to filter the data for the interesting patterns. Subsequently, these filtered data has to be analysed efficiently. However, annotation tools as ELAN usually offer only limited capabilities for filtering and statistical analysis. More particularly, there are the following challenges:



automation As argued in Chapter 1, researchers usually have to perform a set of steps over and over again (e.g. exporting and importing data, counting patterns) in order to bring them to their analysis. Additionally, they often perform very similar analyses for different data types. However, since annotation tools are usually not automatable (or

⁸e.g. by the SoziRob Project <http://aiweb.techfak.uni-bielefeld.de/projekt-sozirob>

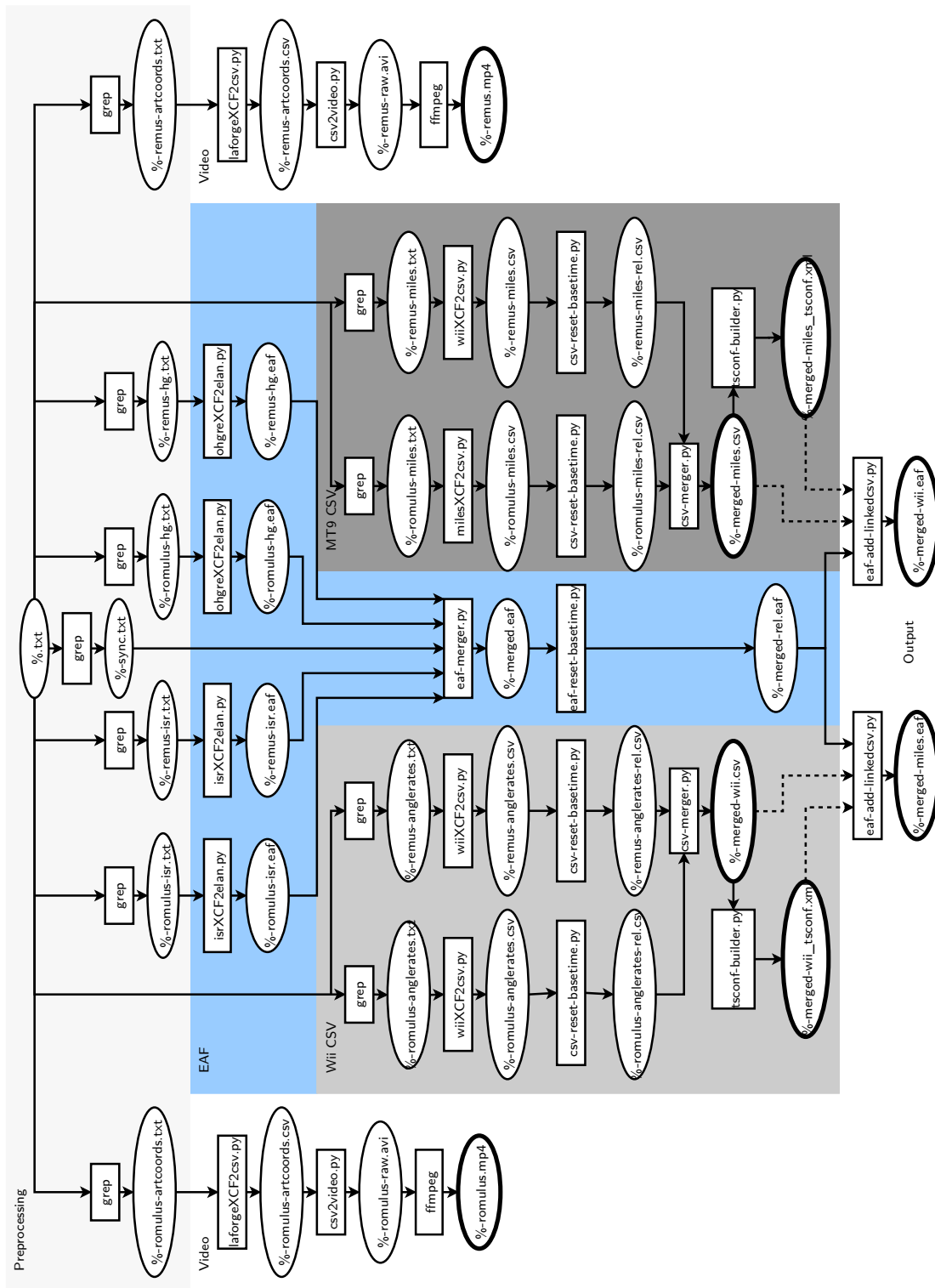


Figure 3.6.: Flowchart of the makefile system that translates all system and memory logs to ELAN-loadable file formats (.eaf,.csv,.mp4) and links and synchronizes them into multimodal ELAN corpora. Derived from a diagram by David Fleer

scriptable), all these steps have to be performed manually.

filter complex relationships It is often the aim of research on interaction to understand complex relationships between several types of interaction signals under study. This is particularly true in work with multimodal data (as in the case of this work which integrates sound, timeseries data and annotations of multimodal signals). An exemplary research question was discussed in Chapter 1: “which nonverbal signals trigger a listening interaction partner to nod in a dialogue?”. However, in order to investigate such relationships between behavioural signals, researchers have to perform several small analyses. For example, it seems interesting to analyse the timing between the nods of the listener and (a) nods of the speaker, (b) voice pitch of the speaker, (c) speech pauses of the speaker, and many more. Such multimodal analyses are often performed manually. Since this is laborious, the size of the analysed corpora is often very small.

As an approach to these challenges, we here present an analysis toolbox. Marc Hanheide and Manja Lohse started this toolbox project under the name SALEM (**S**tatistical **A**nalysis of **E**LAN files in **M**ATLAB, Hanheide et al. (2010)). The initial toolbox is available for download⁹.

In the following, the characteristics of this analysis toolbox will be described briefly. Subsequently, the contributions provided by this thesis will be detailed. Additionally, Section 6.4 will describe an exemplary analysis process. The toolbox provides the following advantages to our filtering and analysis processes:

parsing The toolbox allows to directly import the entire ELAN file bundle preserving the synchronization. This saves the researcher the exporting step from ELAN.

scriptable MATLAB provides a command line interface and thereby offers to automatize all steps of the analysis. By saving these scripts, the researcher can easily re-calculate the complete analysis (e. g. if the corpus has been modified) and apply similar analyses to other data. This enhances the comparability of the results.

statistics and overview The toolbox provides descriptive statistics customized for the analysis of multimodal corpora. Moreover, it calculates the overlaps or distances between arbitrary annotation data and allows to plot the complete ELAN file bundle for overview.

slicing The toolbox enables the researcher to reduce the data by means of specific criteria. For example, when slicing arbitrary intervals of the study, single tiers, or specific annotations (e. g. all ‘nod’ annotations or a phase of a study ‘task1’), the corpus is reduced to all data that overlaps with the annotations matching the criteria.

timeseries support The toolbox allows analysing timeseries data included in the ELAN file bundle thereby calculating characteristics as duration, extreme values and frequency of periods.

For this thesis, several extensions have been integrated into the toolbox that particularly support the analyses necessary for this thesis:

Parsing Since we aim at analysing the whole multimodal corpus including the sensor data, the parsing script `ELANREADFILE` was extended to include files possibly linked to the `.eaf`

⁹<http://aiweb.techfak.uni-bielefeld.de/node/2431>

file like timeseries data files (.csv) and videos. These files (and their offsets deriving from their synchronisation in ELAN) now are saved to the ELAN struct parallel to the tiers struct. Additionally, we import an eaf-basetimestamp file (.txt) if it is located in the same directory determining the base timestamp from the experiment¹⁰. While all ELAN timestamps are given in milliseconds since the beginning of the experiment, this base timestamp can be used to calculate the original timestamp (exact time of an event during the experiment in milliseconds since 1.1.1970).

Coloured plotting The plotting function `ELANPLOT` was extended to enable coloured plotting. For each tier the number of annotations and the number of different annotation values is calculated. Each different value is assigned a colour uniformly distributed over a colour map. On the right side, the numbers of overall (and differing annotations) are listed.

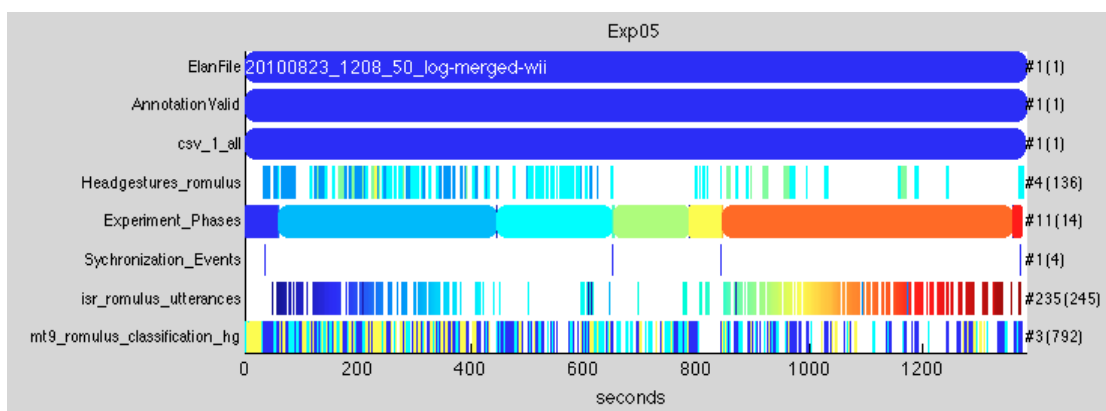


Figure 3.7.: Plotted data file in MATLAB.

Timeseries support The loaded timeseries data additionally is transferred to annotations representing the time intervals of the ELAN file where timeseries data is available. Thus, these are plotted as well when using `ELANPLOT`.

We included an `ELANTIMESERIESSLICE` method that extends the `ELANSLICE` capabilities to include timeseries data in the slicing. This function optionally plots each timeseries slice and saves it to a single file. When slicing gestures annotated by hand for example, this save method allows us to feed the resulting gesture slices to the gesture classifier in order to enhance their training.

`ELANASSIGNTIERSWITHCSV` allows assigning tiers (e.g. head gesture annotation by hand or head gesture classification) with csv files or parts (columns) of the csv file.

`ANALYSEHEADGESTURES` is a script for analysis of annotated or classified head gesture events. It uses a specific tier to find for each annotation the corresponding interval in the assigned timeseries data. The resulting intervals are analysed using `FINDEXTREMA` according to their number and frequency of periods in the gesture as well as their maximum values for each extremum.

¹⁰Though the ELAN XML format includes a timestamp, it is unfortunately reset every time the file is opened in ELAN and is, thus, is not usable as experiment base timestamp.

Other scripts The script `CREATEANNOFROMGAPS` creates new annotations with either specified names or sequential names in a specified tier according to the gaps given in the arguments. This script moreover allows the user to close all gaps between specific annotations. With this, we can for example annotate (and thus slice) all intervals in which a participant is *not* speaking and thus might be listening to the interaction partner.

`ELANCORRELATETIERS` compares the annotations in two arbitrary tiers with each other. For this, it computes the overlap of tier1 with tier2 (optionally plots it, see Figure 3.8) and moreover compares the annotation values of the overlapping annotations with each other. The function is used in this thesis to compute correlation values between automatic gesture tags from the classification module with the manually gesture annotations (see Section 6.3.3). The resulting data can be used to compute confusion matrices providing the probabilities for each wrong classified gesture. Additionally, the function is used to compute distances between adjacent annotations of two tiers (see Section 6.4).

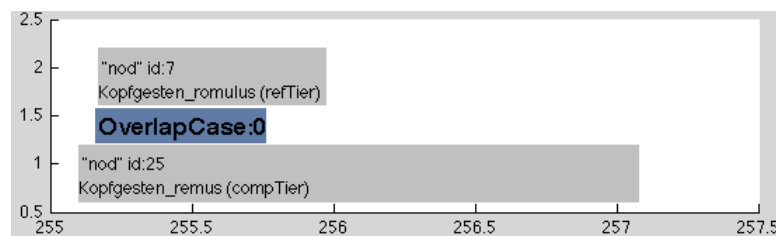


Figure 3.8.: Screenshot of MATLAB showing the overlap of two nodes in two tiers.

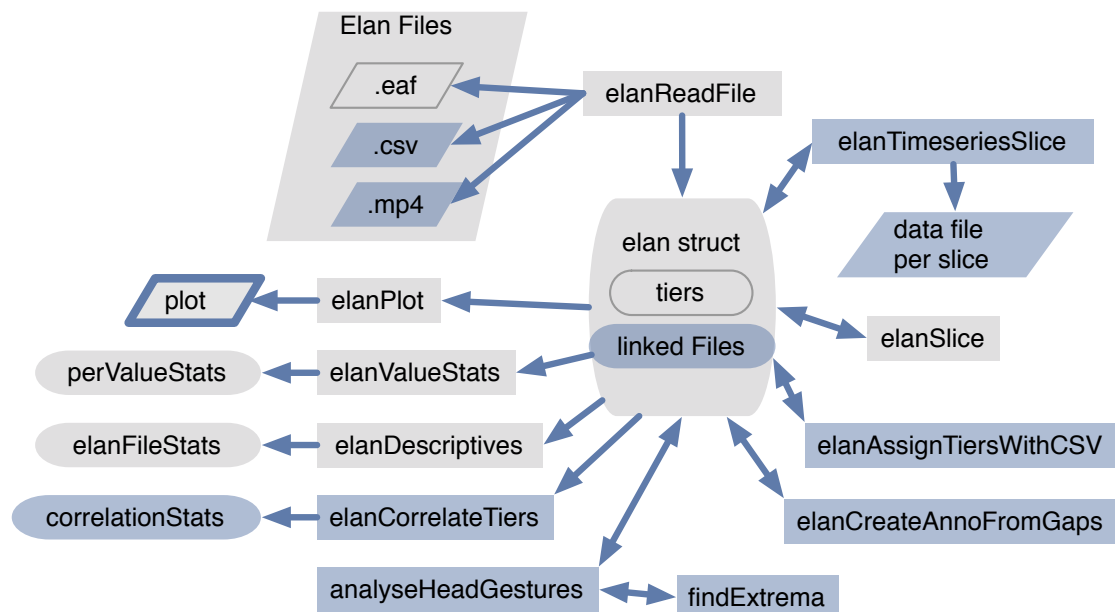


Figure 3.9.: The analysis toolbox SALEM. MATLAB scripts are depicted with rectangles, input/output with parallelograms while MATLAB variables are represented with rounded boxes. Boxes with blue background represent functionalities provided by this thesis while boxes with grey background represent functionalities implemented by Marc Hanheide.

Contribution An overview of the whole toolbox is shown in Figure 3.9. In the flowchart, MATLAB scripts are depicted with rectangles, input/output with parallelograms while variables are represented with rounded boxes. Objects with blue colour represent functionalities provided by this thesis while boxes with white background represent functionalities that were implemented by Marc Hanheide in the initial version of the toolbox.

In conclusion, the analysis methods extends the SALEM toolbox at several points and particularly adds rich timeseries support to load, process, analyse and plot the timeseries data in detail. It offers to analyse multimodal data quantitatively.

4. Scenarios

While the previous chapters described the goals of this work and explained the methods with which the goals are to be achieved, the present section will give an overview of the scenarios that are used for the studies that will be presented in Chapters 5, 6 and 7. Firstly, this section will describe the objects that were used as stimuli in many of the studies. Subsequently, quality criteria for scenarios will be developed. Then, the scenarios will be introduced, divided into AR scenarios and non-AR scenarios. Finally, Section 4.4 will present an overview of the studies that were conducted for this work.

4.1. Interaction objects

The stimuli that were used for some of the scenarios consist of three parts: (a) a (tangible) real world object usually in form of a cube to allow for convenient handling, (b) a black-and-white marker called ARToolKit marker that is fixed or glued to the real world object (see Figure 4.1a) and (c) a virtual object that is shown on top of the marker. In this way, the real world object can be seen as the anchor of the virtual object: every time the anchor is manipulated (e. g. moved, rotated) the virtual object is manipulated in the same way.

There are different kinds of real and virtual objects that were used during this work. For tangible real world objects, we used wooden or acrylic cubes and so-called TUImod objects developed by Bovermann et al. (2008) (see Figure 4.1a). First, we used the acrylic cubes but these were found to reflect infrared light and were thus not suitable for use with the Vicon tracking system. In order to compensate for this, we equipped TUImod objects with ARToolKit markers. But these have the disadvantage that for a larger amount of virtual objects we would need plenty of TUImod material, which would have been cost-intensive.

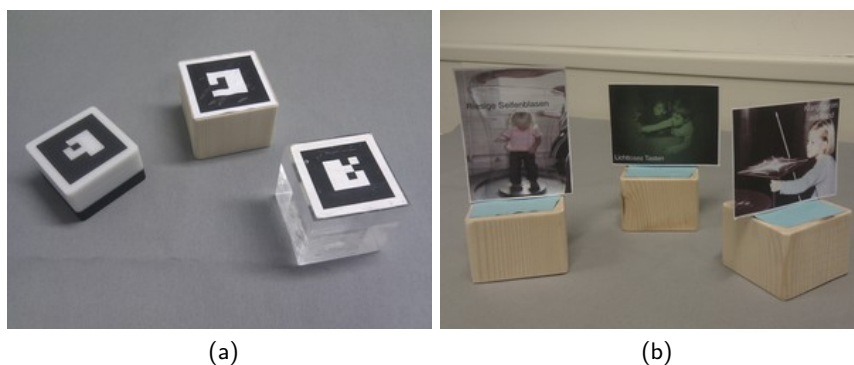


Figure 4.1.: (a) Real world object anchors with markers. From left to right: TUImod object, wooden cube, acrylic cube. (b) Pictures printed on paper for non-AR interaction.

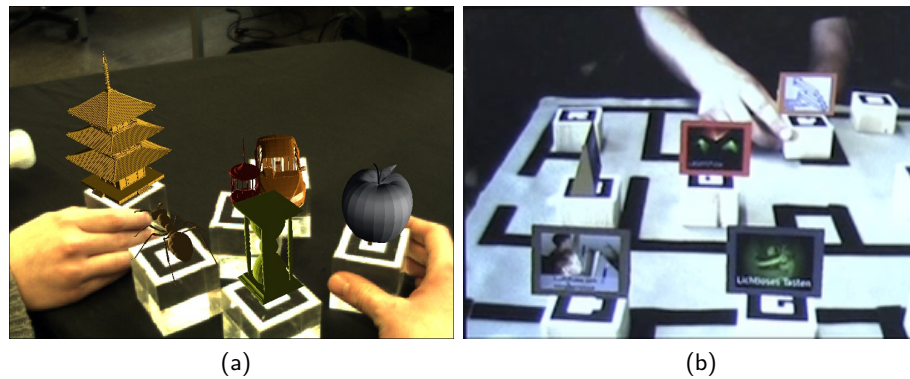


Figure 4.2.: Examples of virtual objects: (a) virtual 3-dimensional objects: car, ball, sand-clock, ant, apple, pagoda. (b) images shown on sides of wedge-shaped 3-dimensional objects.

Finally, for larger amounts of objects we glued the markers to simple wooden cubes that are easily available and do not reflect infrared light. However, compared to the TUImod objects that are very robust, the disadvantage of gluing markers to wooden or acrylic cubes is that the marker has to be renewed from time to time because it wears out.

The virtual objects that are displayed on top of the markers are 3-dimensional objects (see Figure 4.2a) or pictures that are displayed on the lateral surfaces of wedge-shaped virtual 3-dimensional objects (see Figure 4.2b).

4.1.1. Characteristics & Interaction behaviour of the objects

The characteristics of the virtual objects or the set of objects chosen for the study can be configured by the experimenter (see Section 3.1). Possible characteristics that can be configured are the colour, size or type of the object. Apart from the mentioned characteristics, the objects can also play sounds. Moreover, the characteristics can be changed during the trial whenever the participants trigger an alteration. Implemented triggers for such a change are (a) the objects' entering or leaving of the field of view, (b) the objects' position in the field of view and (c) the objects' position in the partner's field of view (see Section 7.1). (d) a button that is pressed by a participant (e) regular or random interference by the experimentation system or explicitly by the experimenter. Another trigger that could be easily implemented is the position of the object on the table (which can already be tracked by an under-desk camera if using a glass table). For this, the table could be divided into fields that trigger certain behaviour of the objects, for example that sounds that are produced by objects in the same room are added in order to allow the user to estimate the field's loudness.

Such interaction objects are used in most scenarios of this work while some also alter the objects' behaviour during the task. In the following, we will first develop criteria for the quality of scenarios for our system and then introduce all scenarios that were used.

4.2. Criteria for scenarios or tasks

In order to monitor interaction signals during AR-mediated interaction, a well-designed scenario is required where the expected signals are actually transmitted. In detailed discussions, our research team¹ developed several criteria regarding the acceptance, the outcome and the feasibility for such a well-designed scenario:

Acceptance: To ensure a high acceptance from the participants, the scenario and the task must not be boring but should rather be interesting or challenging. This is important since the motivation to find a good solution for a task might increase with the motivation of the participants. Nevertheless, the task should be easily explained by the experimenter and understood by the participants. It is important in every study to make sure that the participants understand the task in the intended way. For this, the scenario requires a clear task description.

AR has several side-effects that are perceived as more or less crucial by the participants (see Section 5.1 for a detailed overview and analysis). Some participants even experience discomfort like headaches or nausea. Thus, in AR studies there is a possibility of participants aborting the trial because of such discomforts. It would be great if the participants could enjoy the task in order to compensate for some of the discomforts. Such enjoying the trial might reduce their uncomfotableness or might even prevent them from stopping the trial before it is completed.

Apart from this, the sought scenario should also make use of (at least *some*) AR features because otherwise the participants would wonder why they have to be equipped with so many wearable devices.

Outcome: Since our aim is to analyse interaction signals, the scenario has to offer a task for at least two participants. During this task, the participants are required to collaborate or discuss. More importantly, the scenario has to encourage a rich exchange of interaction signals in order to allow an investigation of the transmitted interaction signals. Thereby, not only verbal but also non-verbal signals should be encouraged.

Feasibility: Finally, the scenario obviously has to be technically viable. When including AR features, it is moreover important to ensure that AR issues like lag or registration errors are reduced as much as possible.

4.3. Scenarios and tasks

With the above criteria in mind, our research team discussed several scenarios and tasks. The following sections present these scenarios that have been implemented (at least to a certain extent) in order to test their effectiveness in our setup.

¹Participants of the discussions: Marc Hanheide, Thomas Hermann, Christian Leichsenring, Christian Lang, René Tünnermann, Till Bovermann and Angelika Dierker.

4.3.1. Object games

In this scenario, two participants are seated at a table and are working with a set of virtual 3-dimensional objects connected to real world cubes. Examples for appropriate virtual objects are shown in Figure 4.2a. There is a huge amount of tasks that is possible with such objects. Simple examples are “cluster the objects into groups that seem to be appropriate to you” or “sort the objects according to their size”. These tasks can be performed either individually or collaboratively. Additionally, the experimenter can allow or forbid speech during the task. Moreover, the virtual objects can be shown differently for the collaborators as was proposed in Section 3.2. For example in one pre-test, a grasshopper was shown with a small size for one and a large size for the other participant.

Discussion with respect to the criteria

Such object games are feasible with our system, the task is easily explained and the acceptance is good. The scenario uses AR features for the display of the virtual objects. In particular, showing conflicting characteristics of the objects (as in the example with the large/small grasshopper) would not be possible without AR. However, the duration of the task is very short. For a sufficient size of interaction corpus, we would need either a huge number of participants or a huge set of virtual objects that can be sorted subsequently. The huge amount of participants would be time-consuming since the adjustment of the sensors and AR goggles to the participant would need much more time than the entire trial. Additionally, the participants would not get used to the technique during the trial and thus might act less normal than participants in later phases of the trial (see Chapter 5 and particularly Section 5.3 for more discussion about the consciousness of the technique). A huge set of virtual objects is difficult to achieve since there are limited numbers of 3-dimensional virtual objects available for free. The purchase of a huge set of objects would be costly and the construction would be time-consuming. Moreover, such a repetition of the same task might bore the participants and reduce the exchange of interaction signals since the participants are likely to simplify their interaction step by step and might use shortcuts instead.

Because of this, the object games were used in pre-tests and as a familiarization phase prior to the gaze game that is described below. With this scenario, the participants can easily familiarize with the system’s AR features (real-time interactivity, alignment of virtual objects with real world objects, visual and auditory techniques, see Sections 2 and 3) as well as the technique’s limits (lag, tracking errors, unusual hand-eye coordination, see Section 5.1). Since the virtual objects are not displayed in their real size (the size they would have in the real world), one goal for the participants is to agree collaboratively if they want to sort the objects according to their size in the real world or to their virtually displayed size.

4.3.2. Collaborative and multimodal 3-dimensional data exploration

Two participants are equipped with AR goggles and their head is tracked in the room. They are sitting or standing next to each other or vis-à-vis. In the space in front of them, a dataset is displayed virtually in 3 dimensions (see Figure 4.3). The participants can examine the data individually from different view angles and can collaboratively discuss about their

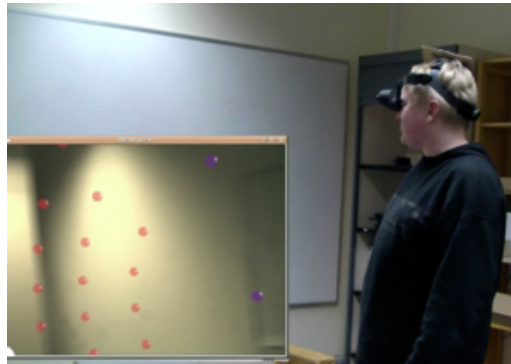


Figure 4.3.: Exploring data in space. In the user's view through the AR goggles (shown in the lower left window) coloured data points are displayed in the space. The user can manipulate his view on the data by his own head-movements. Picture from David Fleer.

characteristics. Although the participants cannot touch the data, they can rotate, shift or zoom the data and can even step into them.

The idea is to facilitate collaborative data exploration. Usually, computer screens are used for such tasks with the computer mouse as interface. With the mouse, the navigation of the data (rotation, shifting, zooming) is not intuitive since the 2-dimensional information has to be mapped to the 3-dimensional manipulation. Thus, novice users usually do not work smoothly with such a task. Moreover, in collaboration it is much more difficult to keep track of the navigation for the person that is *not* manipulating the scene with the mouse compared to the one, that is using the mouse. With the use of AR, however, the navigation and exploration should be much more intuitive and it should be easier to keep track of such navigations because the users can observe each other's manipulations. The scenario has been implemented in a seminar student project (Heinrich et al., 2010) and has been inspired by single-user data exploration implementations, for example by Lee (2008).

Apart from this, it would be possible to attach a marker to the hand or a tool and thereby to interact with the data points (e.g. to highlight data points, remove or draw borders between them). Moreover, it is also possible to combine this exploration technique with other datamining techniques. With this, certain data points could be highlighted in the data set or sonification could be used for the exploration as it is done for example by Bovermann et al. (2006) thus deriving a multimodal display.

Discussion with respect to the criteria

The scenario was so far only used in single-user pre-tests. Nevertheless we believe the scenario to be highly acceptable by possible participants since we found in pre-tests that participants were interested in the work with virtually displayed data. The scenario clearly uses visual AR features and, if combined with sonification, even multimodal AR. In discussions about the data, the users would be able to collaborate very well with their partners and thus we expect a rich exchange of interaction signals including also nonverbal signals. However, the problem might be to find tasks and datasets that are easy enough to understand for users that are not used to such data exploration techniques but are still complex enough to offer longer discussions between the two participants.

4.3.3. Gaze game

In the gaze game, a pair of participants plays cooperatively, sitting next to each other at a table. Six markers are placed on the table on top of which virtual objects are shown to the two participants through their AR goggles. The marker positions at which the objects are displayed are shown in Figure 4.4c.

The goal of each cycle of the game is to search that object on the table that the interaction partner is currently focusing on (gazing at). Each of the two players alternately has the role of the *gazer* or the *searcher*. Each game cycle (see Figure 4.4b) starts with the gazer pressing a button. Thereby, one of the virtual objects that are displayed on top of the markers on the table is displayed in a 2-dimensional version in the corner of the head-mounted display (see Figure 4.4d). Subsequently, he or she has to find the corresponding object on the table and fixate it (phase i). When this is accomplished, the gazer presses a Wii Remote button to trigger a vibration in the searcher's Wii Remote. For the searcher, this is the signal to find the same object as fast as possible (phase ii). Once the searcher thinks the object is found, he/she indicates this with another button press that triggers a vibration in the gazer's Wii Remote. No speaking or gesturing is allowed up to this point in each cycle. The time between these two button presses is taken as a performance measure. The two players are then allowed to speak again and are asked to verify whether they were indeed looking at the same object (phase iii). A final button press by the gazer finishes a cycle, begins a new one and triggers a new object to be displayed. At this moment, the placement of the objects on the table is randomly changed to prevent the players from learning the positions by heart. After ten cycles the roles of gazer and searcher are switched.

To support the participants furthermore in the game cycle, the system additionally displays text asking confirmation whether the participant is ready for the next game cycle to begin (see Figure 4.4e) or announcing a change of the gazer/searcher roles. This scenario can be used with additional auditory as well as visual AR features as is explained in Section 3.5. In short, the objects entering or leaving the field of view of the interaction partner trigger a sound and the objects are coloured according to their position in the partner's field of view.

This scenario was used to evaluate the effectiveness of the augmentations described in Section 6.1 and to test whether the interaction can be enhanced by such a technique (see Section 7.1). For this, each participant pair performed 40 cycles of the gaze game: 20 cycles without and 20 cycles with the enhancement feature switched on.

Discussion with respect to the criteria

Although the gaze game cycle is complex and the explanation is not very easy, the game (once understood) works very well and is easy for the participants since the experimentation system supports the process of each cycle and helps the participants to perform the right actions at the appropriate point in time. Thereby, the experimentation system records autonomously the durations of each cycle which reduces the amount of work for the analysis to a minimum. In our experience, the acceptance of the scenario is good in the beginning but declines to the end. We believe this to be due to the scenario being little varying.

The scenario makes rich use of AR features: auditory as well as visual AR features are used.

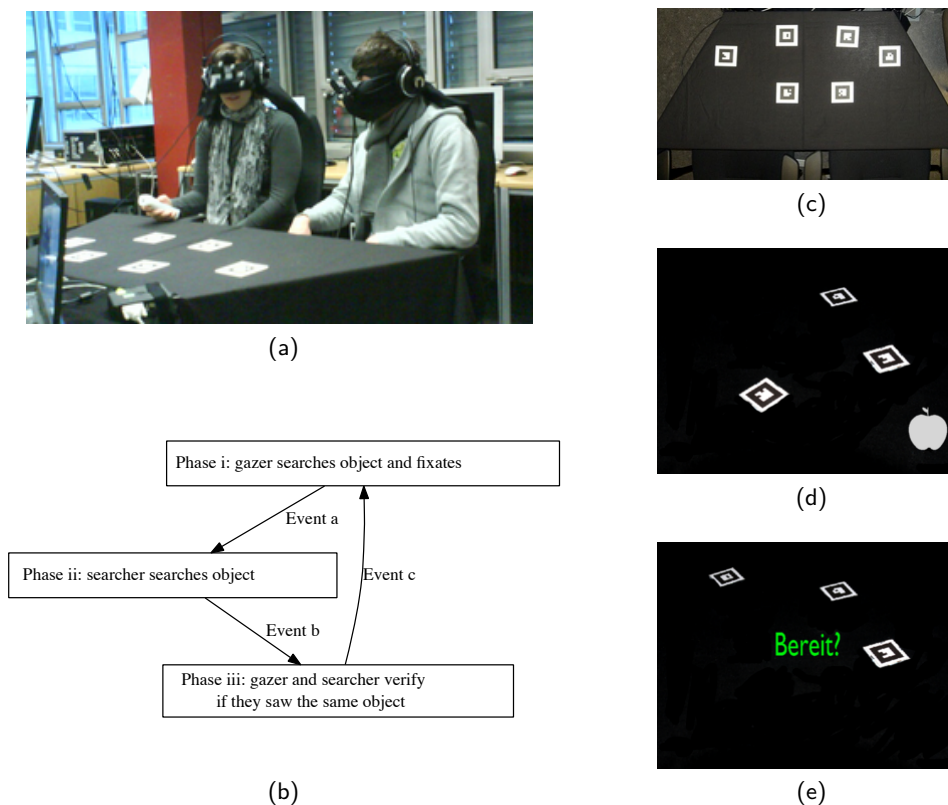


Figure 4.4.: Gaze game scenario: (a) Two participants during the study. (b) Phases of one cycle of the trials which are started/ended by Wii button events (see text for explanation). (c) Arrangement of markers during the gaze game. The long side of the table measured 140cm, the opposite side 70cm and the depth was 61cm. (d) 2-dimensional versions of the object to be fixated (here apple) are shown in the lower right corner to the gazer. (e) “Ready?”. Textual hints are shown at specific points in the cycle. Here, the participant is asked to confirm that he/she is ready to begin with the next game cycle.

Each object’s appearance or disappearance triggers a sound, the objects are coloured according to their position in the partner’s field of view and virtual hints are shown to the participants at specific game cycles showing the object to be searched for, announcing a role change or asking whether the participant is ready for a new cycle. Moreover, a random permutation of the objects after each game cycle would not be possible without AR. However, since the gaze game consists of short repeated cycles, there is no rich exchange of interaction signals between the participants. Additionally, the participants reduce the exchanged signals further during the trial because of efficiency reasons. Moreover, in our experience the participants rarely used non-vocal signals.

4.3.4. Interactive-exhibition design scenario

In this scenario, two participants are sitting face to face at a table. A floor-plan of a museum with different rooms for an exhibition is placed on the table (see Figure 4.5c). The participants are asked to plan an interactive exhibition. There are pictures of exhibits that are to be

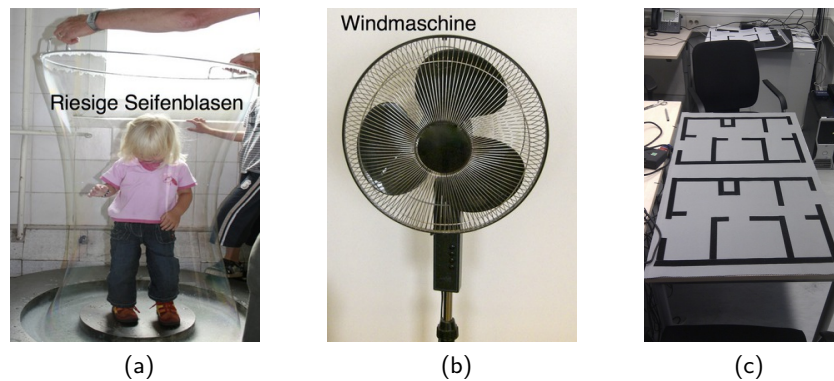


Figure 4.5.: Pictures depicting interactive exhibits for the scenario with titles in German (a)The picture with the (translated) title *huge soap bubbles*. (b) The title is *wind engine*. (c) The floor-plans used in the Scenario.

distributed on the floor-plan. The exhibits would (in a real museum) interactively teach (physical) characteristics of the world to its visitors. For example, one picture shows a child inside a huge soap bubble representing an experiment to learn about surface tension. Another experiment allows the visitor to learn about airflow by means of a huge fan (Figures 4.5b and 4.5a). All pictures of the experiments are labelled in German.

The pictures of exhibits are to be arranged on the floor-plan in such a way that they do not interfere with each other. On the one hand, most of the pictured exhibits *expect certain conditions* of their surroundings. For example, the experiment with the huge soap bubbles needs a room where wind is unlikely. Likewise, the huge fan might be damageable by water. Thus, it should be placed in a room with no water. On the other hand, the experiments also can be the *sources of interference*. For our example this means that the huge fan emits wind and thus should not be located next to the soap bubbles experiment to avoid the destruction of soap bubbles. Moreover, visitors might play with the water of the soap bubbles experiment and thus, electric power (like it is used by the huge fan) should be nowhere in the near.

Pre-tests and discussions revealed the requirements that are listed in Table 4.1 as well as their emissions. Furthermore, Figure A.1 in the appendix lists the titles of the experiments with their translations to English as well as with their mapping to the markers. From this table we can derive the following sources of interference between the exhibits:

- lighting
- sound
- smell
- wind
- wetness
- vibrations

14 of the overall 16 pictures of the exhibits have been derived with kind permission for the use during studies from Phänomenta Peenemünde², Phänomania Essen³ and Phänomenta Lüdenscheid⁴, the remaining two by the author.

²<http://www.phaenomenta-peenemuende.de>

³<http://www.phaenomania.de/essen>

⁴<http://www.phaenomenta.de/Luedenscheid>

Title of the experiment	requires	emits
Airflow around obstacles	room without wind	smell/fume
Colour Mixtures	controlled lighting	light
Extinguishing a candle by drumbeat	room without wind	sound, smell
Feel around in the dark	absolute darkness	
Humming stone	silence	
Huge soap bubbles	room without wind	water
Lasershow	darkness	light
Listening	silence	
Optical illusion: Swivel disk	lighting	
Optical illusion: Triangle in House	lighting	
Optical illusion: Arrows	lighting	
Plasmadisk – electric discharge		
Smelling tree	neutral smell	
Soundfigures of sand	room without wind, dryness	sound
Water-sound dabbling bowl	lighting	sound, water
Wind engine	dryness	wind

Table 4.1.: Titles of experiments that have been used in the scenario (the given title is a translation). The other columns list the experiments' requirements and possible sources of interference between the experiments.

The scenario was used in a study as an individual as well as a collaborative task. In the *individual* task, the 16 objects were divided into two predetermined sets of 8 objects. Each set is assigned to one of the participants who are then asked to find an arrangement for his or her 8 exhibits in his or her museum floor-plan. Although the participants are sitting at opposite sides of the table and are performing the task at the same time, they cannot see each other during the task due to a visual shield and are asked not to speak while finding their solution. In the *collaborative* task the participants are asked to find a placement for all 16 objects in one floor-plan. Thereby, they are encouraged to collaborate with each other discussing the requirements and sources of interference with their partner.

Additionally, the scenario was used with and without AR. In the *AR scenario*, both participants are equipped with video see-through AR goggles. The pictures of the exhibits are presented using virtual objects (see Figure 5.5a). The virtual objects are depicted on top of ARToolKit markers that are attached to wooden cubes that measure $5.5 \times 5.5 \times 4$ cm (see Figure 5.5b). In the *non-AR scenario*, the participants do not wear AR goggles. Here, the pictures are printed on cards with the dimensions 8×6 cm. These cards are vertically affixed to wooden cubes (see Figure 5.5c).

As an example, Figure 4.6 shows two participants during the collaborative phase in the AR scenario. The computer screens show the view through the goggles. More details on the stimuli and procedure as well as the results for this study can be found in Section 5.3. See also Dierker et al. (2011).



Figure 4.6.: The interactive exhibition design scenario: Two participants are positioning interactive exhibits in a museum plan. The computer screens show (enlarged) the participant's current view through the AR goggles.

Discussion with respect to the criteria

The acceptance of the scenario seems to be very good. Most of the participants stated the task to be interesting and fun (see also the evaluation of the questionnaire results in Section 5.3.3). Since the task is explained and understood easily but nevertheless challenging and difficult to master, there are many things to discuss between the collaboration partners. This yields a rich exchange of interaction signals including also nonverbal communication like head gestures. Additionally, the duration of the task is long enough to ensure satisfactory amounts of data without having to repeat a similar task. The task uses AR features to show the pictures of the exhibits to the participants. Additionally, the attention focus display (see Section 3.5) can be used in this scenario to help the participants to keep track of their partner's focus of attention.

However, some participants asked why they had to use AR during the task. There are many features that can be added to this scenario that were discussed. These were not used yet since the focus of the study in Section 5.3.3 was to compare the effect of the system on the interaction. Thus, the conditions needed to be comparable (without advantages over another).

For real benefits of the AR features in this scenario, the main idea is to provide information to the participants using AR about the sources of interference for the interactive experiments that are represented by the virtual objects. For example, the system could let the users know if a certain object can be placed in a room in the museum plan when asked for a specific source of interference (e. g. sound) or it could highlight all objects that are already located in the room that interfere with the new one. Alternatively, the soundscape of a room could be played to the participants' headphones every time they place a new object into a room.



Figure 4.7.: Stimuli presentation for the visual search scenario: (a) search target “Find the 2!”, (b) number grid

4.3.5. Visual search

During the visual search scenario, one participant is sitting in front of a 275 × 205 cm white projection area. The participant is shown a target (e. g. ‘Find the 2!’) projected in the middle of the white surface. By a button press, the participant can confirm that he or she is ready. Triggered by the button press, an 8 × 6 number grid of single-digit numbers (0..9) appears on the projection area that includes the target digit (‘2’) exactly once. The participant is now asked to find the target in the grid as fast as possible and then press the button again to confirm the completion of the trial. Every participant traverses a series of such trials. The target numbers and their position are varied between each trial.

The visual search scenario was used to compare the eye movements of participants wearing a head-mounted display to lateral blinders and unrestricted view. For this, all participants wore an eye-tracking system. Please find further details about the study in Section 5.2.

Discussion with respect to the criteria

The visual search scenario is easy to explain and to understand. The scenario does not use AR features although the participants wear AR goggles in one of the conditions. Moreover, the scenario is not collaborative since only one participant is performing the task at a time. Thus, we do not expect to record an exchange of interaction signals. Instead of recording signals, this scenario is tailored well for the analysis of the specific question how the eye movements vary between the three different restriction conditions.

4.3.6. Animal guessing

In the animal guessing game, two or more persons are sitting facing each other. The game starts with one participant thinking of an animal. The other participant has to find out which animal his or her partner has in mind by asking him/her questions. The first participant is only allowed to answer with ‘yes’ or ‘no’ (or in case of need with ‘that depends’). Both participants are asked to take turns in guessing their partner’s animals (see Figure 4.8).

This scenario was used to acquire a detailed head gesture corpus. The corpus was then



Figure 4.8.: Data acquisition in an animal guessing task. Both interacting participants are wearing a head-mounted motion sensor and are taking turns in guessing the animal the other is thinking of. Pictures taken by Christian Leichsenring.

used to achieve two goals (see Section 6.3): The first goal was to learn more about head gestures and to find average parameters for head gestures as well as extreme values for head movements. The second goal was to train the head gesture recognition software that we developed in order to facilitate the analysis of the head gestures that were found in our studies. The software can now provide a head gesture hypothesis based on the acquired corpus data.

In order to gain more head movements with the same scenario, it is possible to ask the participant who thinks of an animal not to answer with ‘yes’ or ‘no’ during the game so that he or she instead relies on head movements. Sometimes, these head gestures are then paired with vocal but non-verbal back-channels like ‘mmh’ or ‘u-huh’.

Discussion with respect to the criteria

The scenario is very easy to explain and (since many participants have already played it as a children’s game) easily understood. The participants seem to have fun playing the game. Still, after some time, the participants seem to lose interest in the game. Since the game is a non-AR scenario, we do not use any AR features. During the game, the participants exchange both verbal and nonverbal communication signals. However, since the game exclusively consists of questions and yes-or-no answers, the interaction signals that are exchanged are presumably different compared to other interaction situations like spontaneous conversation, a discussion or a negotiation situation where we would expect a wider range of different signals. Which kind of interaction signals is desirable has to be considered with respect to the research question.

4.3.7. Smalltalk

In this task, two participants are sitting vis-à-vis and are asked to talk about a topic they like. If they do not like to choose a topic by themselves, the experimenter suggests they could use the time to learn to know each other or to talk about if they had ever been to a so-called interactive museum where the visitors can interact with the exhibits and thereby learn something about the world. The goal of this task is to encourage a spontaneous conversation.

This task was used as a familiarization phase in the interactive exhibition design study (see Section 5.3) for the participants to get used to the sensors attached to their heads while learning to know each other. It was used with as well as without AR goggles. The minimum duration of this task in the study was 5 minutes and the task was interrupted after this duration at a time where the conversation seemed appropriate to be interrupted.

Discussion with respect to the criteria

The smalltalk task is explained and understood easily. Although some participant pairs experience a short interval with awkward silence, most of them do not and find plenty to talk about. Thus, we can conclude, that the task is not perceived boring and the acceptance is good. Additionally, it is a collaborative task that can lead to a rich exchange of verbal and nonverbal communication signals from spontaneous interaction. However, the task does not make use of AR features. Even for the participants that wear AR goggles during its duration the AR goggles simply connect the video stream through. In conclusion, this scenario encourages spontaneous interaction and nonverbal signals but does not need AR features and thus the participants might question the goal. To prevent this, we introduced the task as a familiarization task, as a preparation for the following tasks.

4.3.8. Prompting by computer

In this task, one participant is equipped with a head motion sensor and is sitting in front of a computer screen. The computer screen requests the participant to nod, shake or tilt his or her head repeatedly. Thereby, the participant is asked to provide nods of different intensity and duration. This task was used as one approach for the head gesture data acquisition.

Discussion with respect to the criteria

This prompting by the computer is easily explained to the participants. However, the task is neither interesting nor fun for the participants and has to be repeated several times to provide the researcher with a sufficient amount of data. Additionally it is not collaborative and does not make use of AR features. Thus, it is not suited for the investigation of natural nonverbal behaviour in AR-based collaboration. Nevertheless, the task is a very easy way to gain data for the training of the head gesture recognition. However, it has to be kept in mind that this data is not taken from natural interaction and thus might bias the recognition reducing the classification results.

4.3.9. Summary

The previous sections introduced 8 scenarios and discussed them with respect to the criteria. Table 4.2 gives an overview of the discussed scenarios and the criteria they meet. As the table shows, several scenarios do not meet several criteria (especially the last four columns). Why were they used in spite of this? When planning a study it is not only important to choose a scenario that meets the highest number of criteria. Instead, it may be even more

Criteria	Object games	Coll. data exploration	Exhibition design	Smalltalk	Gaze game	Animal guessing	Computer prompting	Visual search
collaborative/2 participants	+	+	+	+	+	+	-	-
exchange of nonverbal signals	o	+	+	+	-	+	-	-
makes use of AR features	+	+	+	-	+	-	-	-
interesting/fun	o	+	+	o	-	-	-	-
enough data without repetition	-	+	+	+	-	-	-	-
explained easily	+	o	o	+	+	+	+	+

Table 4.2.: Overview of the scenarios of this thesis and their fulfilment of the criteria developed in the beginning of the Chapter.

important to choose a scenario that is suitable for answering *this specific* research question. While the first 4 columns show scenarios that focus on eliciting nonverbal behaviour by using AR features, the remaining columns are used in this thesis to answer specific questions (gaze game and visual search) or to gain a head gesture corpus to train the recognition with (animal guessing and computer prompting). With the considerations above in mind, we regard the collaborative data exploration scenario and the interactive exhibition design scenario as particularly promising for the thorough analysis of interaction phenomena since they allow a rich exchange of nonverbal signals, are collaborative and make use of AR features. However, to focus our resources on one of these scenarios (since both are particularly complex), only the exhibition design scenario was implemented completely. It will be evaluated in more detail in Section 5.3.

4.4. Studies presented in this work

Table 4.3 gives an overview of all studies presented in this thesis. For each study, a subset of the available system components was used which are listed in the Table. The reasons to use only a subset instead of all system components were varying:

Availability some of the sensors were not available for every study. For example, the Vicon system is shared amongst many research groups and thus seldom available. The Wii MotionPlus gyroscopes sensors were only prepared as a replacement for the Xsens MT9 sensor since one of them broke down during the study. The Eye-Tracker belongs to a different group and was only available to this project during the cooperation. Although developed during this thesis, the under-desk-tracking support was not usable during this thesis any more since there was no glass desk available for the study.

Performance some of the sensors create huge amounts of data. These data have to be stored during the trial without affecting the study and analysed subsequently. Thus,

we limited the number of sensors used during the study to the ones necessary to the research question.

Load since most of the sensors are attached to the participant's head, we had to take into account the maximum load that is worthwhile for the study.

Scenario every scenario focuses on certain aspects of the interaction. We chose the sensors to appropriately answer the research questions.

This chapter introduced the scenarios and tasks that were used in this thesis. The following chapter will evaluate the effects of our HMDs on interaction while using the visual search task and the interactive exhibition design scenario. Chapter 6 examines the capabilities of ARbInI's methods for facilitating the analysis of interaction and Chapter 7 investigates how mediated interaction can be enhanced or disturbed by using ARbInI's closed-loop experimentation system.

System components	Head gesture data acquisition		Side-effects on the interaction		Enhancing/ disturbing interaction
	<i>animal guessing</i>	<i>prompting by computer</i>	<i>visual search</i>	<i>smalltalk & interactive exhibition design</i>	<i>object games & gaze game</i>
Scene Camera	✓	–	✓	✓	✓
HMD	–	–	✓	–/✓	✓
Headphones	–	–	–	–	✓
Microphone	–	–	–	✓	–
Xsens MT9	✓	–	–	✓	–
Wii MotionPlus	–	✓	–	✓	–
Wii Buttons	–	✓	–	–	✓
Wii Vibration	–	–	–	–	✓
Vicon	–/✓	–	–	–	✓
EyeLink II	–	–	✓	–	–
Recording	✓	✓	–	✓	✓
# participants see Section	5(6)×2 Sec. 6.3.5	7 Sec. 6.3.5	6 Sec. 5.2	12(13)×2 Sec. 5.3	11(13)×2 Sec. 7.1

Table 4.3.: Overview of the studies of this thesis and the used system components. The last row lists the section of this thesis, where the study is presented. The second last row gives the numbers of participant pairs that were used for the analysis and the number of pairs that attended the study.

5. Side-effects on the interaction

The previous chapters proposed the use of the *AR-enabled interception interface* (ARbInI) for the facilitation of communication research. The hardware and software necessary for such a system was presented as well as the possible features that are achieved using this system. Particularly, the benefits of Augmented Reality (AR) and head-mounted displays (HMDs) for our setup were discussed. It is important, however, to also acknowledge the restrictions current AR systems impose on the interaction. This allows researchers to be better able to choose a specific AR system over its alternatives. This is especially true in those cases where they would most likely benefit from the chosen system despite its currently inevitable side-effects.

This chapter investigates the issues that the use of AR via HMDs still has and discusses possible influences on the behaviour of the users. Beginning with an overview on related work on this topic, this chapter proceeds by presenting two studies analysing certain effects of the AR goggles used in *our* setup on the user's behaviour. Based on the findings, the final discussion summarizes some issues that are particularly affecting the behaviour of the users and proposes some alterations that seem to be both valuable and accessible.

5.1. Issues of head-mounted AR and their effects on the wearer

AR has been introduced more than forty years ago (Sutherland, 1968) and has been highly revised and enhanced since then. But there are still a significant amount of challenges on the way towards unobtrusive devices without side-effects.

Characteristics of the used AR system The participants of the studies presented in this thesis were equipped with *video see-through HMDs*. Video see-through denotes that the users wear goggles that enable them to perceive the real world via video images. These images are captured by a front-mounted camera included in the goggles. AR objects are superimposed on the video stream. More specifically, using the vocabulary of Patterson et al. (2006), a *closed system* is used. This means that a direct perception of the outside world is not possible because it is blocked by the goggles. A peripheral view is only possible if the participants bypass the system. The system is furthermore *bi-ocular* (which means that the same image is presented to both eyes of the user, see below for a distinction of 'bi-ocular' from 'bin-ocular' and 'mon-ocular'). As a head-mounted see-through system, it is also *world-stabilized* (which means that the image changes with the user's head position).

There are several surveys comparing the issues for such kind of displays with others, for example by Azuma et al. (1997, 2001); Patterson et al. (2006) as well as Papagiannakis et al. (2008). The authors state that HMDs over the past years have been enhanced in size, robustness, resolution and weight. Additionally, the software used for AR applications

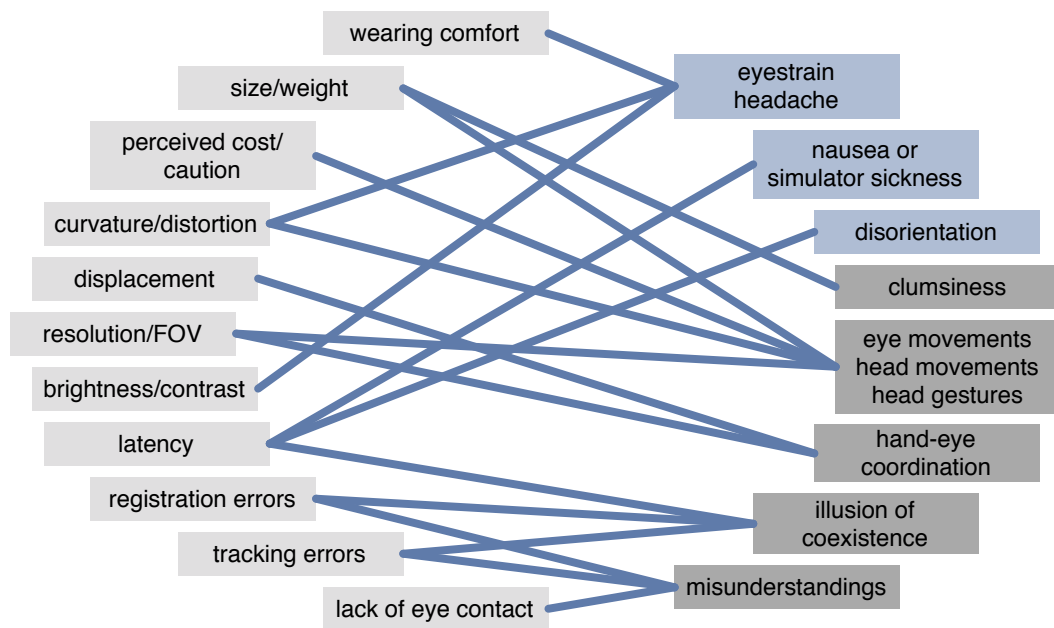


Figure 5.1.: Visualization of the discussed issues with AR and their effects. This diagram shows the problematic characteristics on the left side and their possible effects on discomfort (blue) and behaviour (dark grey) of the users. Both, discomfort and behaviour might affect the task performance.

improved over the years. But in spite of these improvements in the hardware and software used for head-mounted AR, there are still several drawbacks that are perceivable by the wearer. These issues can possibly influence the wearer's behaviour and performance. In the following, the typical characteristics of today's head-mounted AR systems will be discussed one by one thereby collecting the issues that are introduced by each characteristic. Figure 5.1 additionally summarizes the issues and their effects on discomfort and behaviour of the wearers.

bi-ocular Compared to mon-ocular and bin-ocular systems, the bi-ocular system uses a bin-ocular display that shows the *same* image to both eyes. According to Ellis et al. (1997), this kind of display causes significantly more discomfort, both in eyestrain and fatigue than mon-ocular or stereo displays. But also the other two possibilities of display bear problems: Mon-ocular displays, for instance, can be used either with occluding or non-occluding the remaining eye. If not shielded, there is the risk of interocular rivalry and perceptual instability, as Pölönen et al. (2010) stated. Open or semitransparent mon-ocular HMDs on the other hand might damage the illusion of reality (Azuma et al., 1997). Bin-ocular HMDs on the other hand (where both eyes see different images) can create misalignment problems and image distortions between both eyes views (Patterson et al., 2006). Additionally, bin-ocular vision needs the simultaneous processing of two full image streams instead of only one that bears further problems of latency.

Since the illusion of reality is crucial to our setup (see Section 3.2), this work used bi-ocular vision. Thus, it balances between a tolerable latency while preserving the highest possible illusion of reality thereby accepting that the method might give rise to discomfort. We believe this discomfort to be negligible since most participants did not



Figure 5.2.: Hardware improvements. The HMD hardware was reduced and improved in weight and user comfort over the past years and extended to use a set of sensors. left: oldest version, right: version used in this thesis.

notice that they were seeing the same image with both eyes and were even surprised when they learned this after their participation.

video see-through In contrast to video see-through devices, devices that use the *optical* see-through method usually (a) suffer from less distortion, (b) there is no eye offset since the natural inter-pupillary distance is not affected and (c) the real-world is not affected in resolution in contrast to video see-through HMDs as Rolland et al. (1995); Azuma et al. (1997, 2001) argue.

The authors also list several advantages for using *video* see-through devices: (i) they provide a wider field of view for the AR technique and the whole field of view can be used for displaying AR effects. (ii) Video see-through devices enable the researcher to obscure real world objects completely. (iii) Since the video stream is processed and displayed jointly with the AR features, there is less temporal mismatch between the real and virtual world (at the cost of a general delay). (iv) Moreover, the authors report that it is easier to match the brightness and contrast of real and virtual objects and that the user can use the same focus for the real and virtual world (again at the cost of not optimal images).

Together, these advantages of video see-through devices (wider field of view, obscuring possible, less temporal mismatch, better matched brightness and contrast) cause the virtual objects to appear rather solid. This stands in contrast with optical see-through devices, where virtual objects appear rather semi-transparent or ghost-like – characteristics that damage the illusion of reality (Azuma et al., 1997). The illusion of reality and the acceptance of the virtual objects, however, is crucial if the AR system should be able to induce interaction conflicts using virtual objects (as is applied in Section 7.2). Moreover, by using video see-through goggles, we can infer from the field of view of the video camera on the actual field of view of the user and thus determine which objects can be seen by the user (see Section 6.1). Thus, this work uses video see-through devices.

size, weight and wearing comfort Although the size of our HMDs was reduced significantly over the past years (compare Figures 5.2) and so was the weight, our current HMD still weighs 220 g. In our experience, this can result in clumsiness and reduced head movements. Additionally, some participants complained about the weight lying on the root of the nose and the amount of pressure imposed by the method of mounting the

goggles to the head. Some of them stated this to result in eyestrain and headaches.

cost and fragility Both seem not to directly influence the behaviour of the participants but they may indirectly influence the participants' behaviour since they *think* the device to be costly and delicate and thus are especially careful not to do anything that might damage (parts of) the device. In our studies, we noticed caution in touching the devices' cameras or displays and moving the head in such a way that might affect the wires. Moreover, this caution may result in non-perfect configuration of the system (e. g. eye distance not configured correctly) and thus may induce additional problems (e. g. blurred view, discomfort).

displays and cameras Here, several effects occur from the quality of the displays and cameras:

(spatial) resolution and field of view The human field of view (that is the total bin-ocular field of view) measures about 180° horizontally by 150° vertically but varies individually (Dolezal, 1982). But the field of view of the bi-ocular HMD used in this work is much smaller: 42.2° horizontally and 30.4° vertically. According to Dolezal (1982), a narrow field of view tends to make objects appear nearer than they really are and to effectively shrink the environment around the user. Additionally, when using a restricted field of view, moving objects are visible for shorter durations, less context information is available which leads to more and smaller head movements when scanning a scene, (Dolezal, 1982). Likewise, de Vries and Padmos (1997) found that the field of view significantly affected the user's head movements measured by mean total head speed and other measures of motion.

While discussing the effects of HMDs not only the size of the field of view is relevant but also the resolution of the image on the display as well as the combination of field of view and resolution measured in pixels per inch (ppi). When not increasing the resolution, a large field of view is perceived grainy and closer while the same image displayed on a smaller field of view (with higher amount of pixels per inch) is perceived farther away. This may affect the hand-eye coordination (Rolland et al., 1995). Moreover, the resolution of the images also affects the accuracy of the applied tracking methods. Finally, achieving both a wide field of view and high resolution is difficult, given the current technological limitations (Patterson et al., 2006).

brightness and contrast Unnatural configurations of brightness and contrast can give rise to eyestrain and thus to headaches or nausea. Moreover, brightness and contrast of the camera image can influence also the software and induce tracking problems.

curvature The curvature of the display (which is positioned directly in front of the eyes) can cause distortion and thus blurred images especially in the peripheral regions. By this, the user might not notice objects in the peripheral field of view. This means that the user has to increase the number of head movements in order to keep the objects in the centre of the field of view. Additionally, the blur can lead to eyestrain and thus to headaches or even nausea. A distortion can, furthermore, cause registration error. Although it might be technical possible to correct these

distortions by appropriate software, authors like Holloway (1997) argue that this correction might actually lead to more delay and thus to more registration error than the distortion the software tries to correct.

displacement/parallax error This error is caused by the fact that the cameras are mounted away from the eye location and by the used bi-ocular view. This lateral or depth displacement leads to inappropriate pointing or grasping attempts. Rolland et al. (1995) found that subjects could adapt to this form of displacement in AR but had large overshoot in a depth-pointing task after removing the (displaced) HMD. Especially in collaborative tasks it is difficult to ensure that the interaction partner clearly understands what other users are pointing at or referring to. An approach to avoid these problems taken by this work will be discussed in Section 7.1.

latency Because of the several processing steps that are taken of the image (tracking, overlaying virtual objects, etc), there is always a certain amount of latency. Its duration can be influenced by the speed of the computer, its hard disks, the software and the network (since we use XCF, see Section 2.2). This sort of delay is not only a problem because it might annoy the users but it can also reduce the task performance (Ellis et al., 1997), and can create disorientation, nausea, and discomfort (Arthur, 2000). These effects are also known as simulator sickness and motion sickness and seem to occur because of a sensory conflict between proprioception and vision (Barrett and Thornton, 1968). Additionally, a noticeable lag might also hurt the illusion that real and virtual objects coexist.

registration errors The term registration refers to the visual alignment between virtual and real world objects. Holloway (1997) gives a thorough overview of the different causes of registration errors. They found that system delay causes more registration error than all other sources combined and that one millisecond of delay in the worst case causes one millimetre of error. Registration errors might confuse the users and damage the illusion that both worlds (virtual and real) coexist (Azuma et al., 1997).

tracking errors Such errors can be caused by insufficient resolution or brightness/contrast conditions. For the superimposed information these errors might lead to missing objects, blinking objects, objects that are placed at wrong places or to objects that are placed on top of a wrong marker. This might confuse the user and lead to misunderstandings during an interaction. As a result, the user's faith in the correctness of the system and the illusion that the virtual and real world coexist might again be hurt.

interaction Since the HMD blocks the eyes from the interaction partner's view, the users cannot see each other's eyes. Thus, gaze contact is not available for interaction. This inhibits the transfer of a number of social information (e.g liking, attraction, competence, social skills, dominance) or as cues for interaction (e.g. turn-taking, estimating focus of attention) (Kleinke, 1986). A compensation approach, taken by this work will be discussed in Section 7.1.

In spite of all these possible effects of today's head-mounted AR systems on the users' perception and their behaviour, AR is still believed to be a fruitful and promising technique whose benefits outweigh its disadvantages for many applications. Some studies could even measure improvements due to AR in performing an assembly task when comparing AR to non-AR: For example, Tang et al. (2003) observed a lower error rate and lower workload

of participants using instructions displayed in a head-mounted AR system compared to participants using a printed instruction. Baird and Barfield (1999) compared the performance in time and errors of participants using four different types of instructional media: a paper manual, a computer-aided, a video see-through AR display, and an optical see-through AR display. They found that the optical see-through AR display resulted in the lowest task completion times, followed by the opaque AR display, the computer-aided instruction, and the paper instructions respectively. In addition, the error rate was lower in the AR conditions compared to the computer-aided and paper conditions (Baird and Barfield, 1999). Rosenthal et al. (2001) compared AR-guided needle biopsy to standard ultrasound-guided needle biopsy and found a significantly smaller mean deviation from the desired target than with the standard method. Alves Fernandes and Fernández Sánchez (2008) reported user comments that stated AR to help understand forms and volumes of objects in an easy way and to help comprehend the location objects in 3D as well as their relations to others.

To better understand the specific limitations and opportunities of our own setup, we conducted two studies that each compared an HMD condition to a condition without HMD in order to evaluate the effect of the HMD on different behavioural aspects. These studies will be presented in the following sections.

5.2. Influence of HMDs on eye and head movements

This section evaluates the influence of *our* HMD on the users' eye movements and their associated head movements during a visual search task. From the review above, two main aspects that might influence the eye movements can be assumed: the restricted field of view as well as the display with its low resolution and curvature. Both might affect the foveal as well as the peripheral vision. Moreover, head movements and eye movements influence each other mutually. Thus, apart from the resolution, the curvature and field of view, the eye and head movements can also be influenced by size, weight and perceived cost of the HMD. In order to test the effect of the field of view and the resolution of *our* HMD, we conducted a within-subjects study where the participants were asked to search random numbers on a number grid while using an eye-tracking system. To test the effect of the field of view independently from the resolution, there were three experimental viewing conditions under which the participants performed the visual search: unrestricted view, blinders and HMD. We recorded the users' head and eye movements. This study will be briefly presented in the following. Please find further details on the method, the results and implications in Kollenberg et al. (2009) and Kollenberg et al. (2010).

5.2.1. Expectations

Our expectations were that the restricted field of view leads to more head movements with higher amplitude. This should occur in both field-of-view-restricting conditions: the HMD and the blinders condition. The low display resolution and its curvature might lead to peripheral blur. If this has an influence too, the eye and head movements using the HMD should differ from those in the blinders condition: we expect fewer eye movements with lower amplitude.

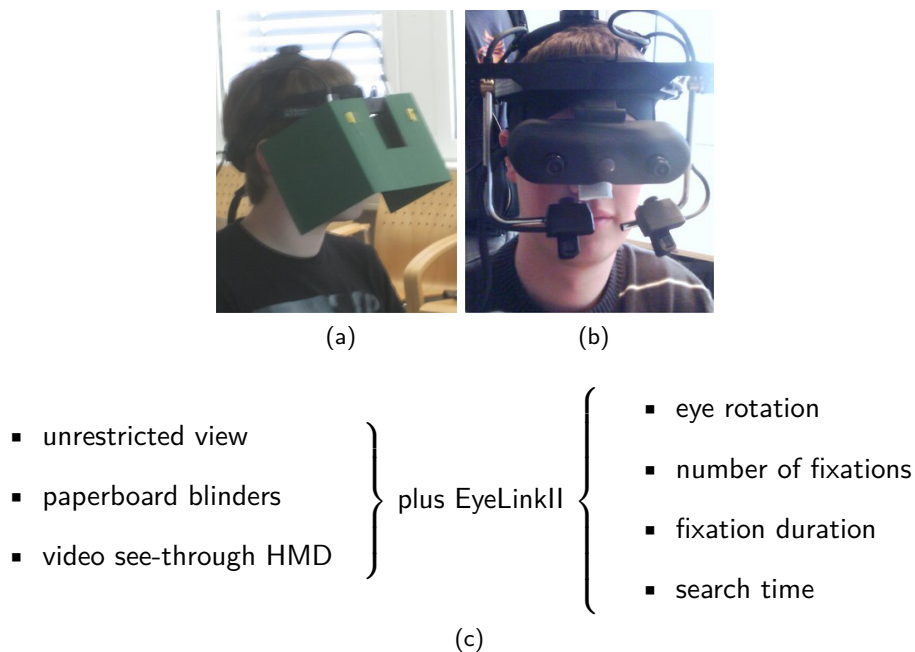


Figure 5.3.: Conditions of the visual search study. (a) Participant wearing blinders and eye-tracker. (b) Participant wearing HMD and eye-tracker. (c) Conditions and measurements for the eye-tracking study.

5.2.2. Method

The independent variable *viewing condition* had three levels: unrestricted view, blinders' view (Figure 5.3a) and HMD view (Figure 5.3b). In order to enable eye-tracking, all three viewing conditions were combined with an eye-tracking system. As dependent variables, we measured the eye rotation, the number of fixations as well as the fixation duration using the eye-tracker. Additionally, the overall search time was measured for each trial. The conditions as well as the measurements are listed in Figure 5.3c.

The stimuli were an 8×6 grid of single-digit numbers (0..9) projected to a 275×205 cm white surface (as described in Section 4.3.5). The participants were seated in a distance of 200cm from the projection surface (see Figure 4.7b on page 65). The target digit was occurring only once and in each trial for all participants at the same grid position. The target number and its position was varied between the trials. All participants completed all viewing conditions. Each condition consisted of 10 trials in random order starting with one practice trial per condition. During the trial, the participants had to detect the target provided as text within the number grid. The sequence of viewing conditions was permuted between the participants. Six participants completed the study.

5.2.3. Results

The *search times* differed significantly between the viewing conditions: in the HMD condition, the participants needed significantly longer time than in the two other conditions: the normal

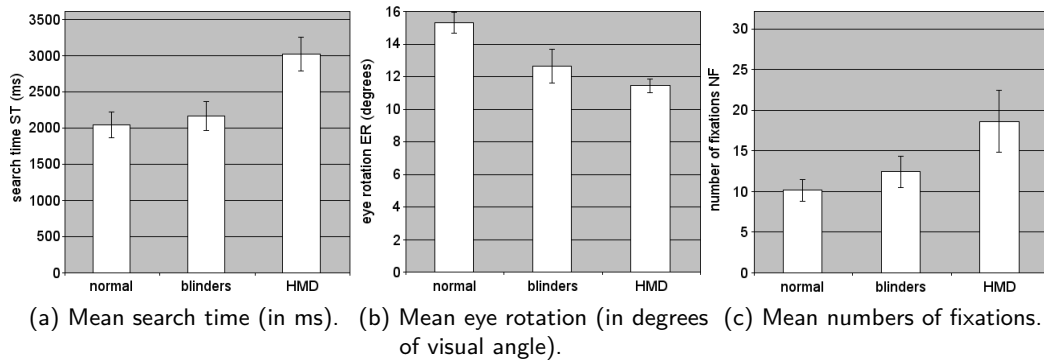


Figure 5.4.: Results from the eye-tracking analysis for the different viewing conditions (normal viewing, blinders, HMD), respectively.

viewing [$T(5) = 4.461; p = 0.007$] and in blinders condition [$T(5) = 3.765; p = 0.013$]. There was no significant difference between the normal and blinders conditions (see also Figure 5.4a). Also the *eye rotation* differed significantly between the viewing conditions: the eye rotations in the HMD condition were significantly smaller than in the normal viewing condition [$T(5) = -9.276; p < 0.001$]. There were no significant effects for the other combinations (see also Figure 5.4b).

For the *number of fixations*, there was a significant difference for normal viewing compared to the HMD condition [$T(5) = -3.202; p = 0.024$]. Other comparisons revealed no significant effects (see also Figure 5.4c).

Likewise, there were no significant effects found for the *fixation duration*. See also Kollenberg et al. (2009) and Kollenberg et al. (2010) for further details to this study.

5.2.4. Discussion

The results show that the search times increase for the HMD condition compared to the normal viewing condition. Additionally, the eye rotation decreases while the number of fixations increase. This means that users of HMDs perform less eye rotation during visual search tasks and thus more head rotation in order to achieve the same shift of gaze direction compared to unrestricted viewing conditions.

Comparing the normal and the blinders' condition, we found neither for search time, nor for eye rotation nor for number of fixations significant differences between the dependent variables although the mean values slightly decreased. Thus, the restriction of the field of view (decoupled from the lower resolution) seems in this study not to have an effect alone but only in combination with the low resolution and the curvature of the HMD. Interestingly, Dolezal (1982) drew another conclusion since he found significant effects for restricted field of view alone. The scenario that was used for the observation could for example explain this difference. In the presented study, we used an artificial visual search paradigm while Dolezal (1982) has used a much more natural everyday setting. On the other hand, it might also be that the effect of resolution and curvature for our HMDs on head and eye movements is even more pronounced than the effect of the field of view but only that the effect of the

field of view is not yet measurable, e. g. because of the small number of participants. Further investigation is needed to learn more about the effect of the curvature and resolution of our HMDs on visual search.

5.3. Influence of HMDs on head movements, speech and task accomplishment

As argued in Section 5.1, head-mounted displays (HMDs) clearly influence the perception and have a considerable weight which might influence several of the user's behavioural signals. As a reminder, today's head-mounted AR systems might lead to:

discomfort A discomfort like eyestrain, headaches or nausea (from simulator sickness) can be induced by the wearing comfort, the bi-ocular presentation or by the video see-through technique since these affect resolution, curvature, brightness and contrast of the HMD. Additionally, a perceivable latency might cause discomfort.

altered body movements Especially head movements can be hindered by size, weight and perceived cost of the HMD as well as by its resolution, curvature and field of view. Since the head and eye movements mutually affect each other, these issues most likely also influence the eye movements.

interactional problems Failing hand-eye coordination and pointing gestures as well as imperfect illusion of coexistence for the virtual and real world may cause interaction problems. These aspects are influenced by the resolution, field of view, brightness and contrast as well as by software issues like registration- or tracking errors and latency. Finally, the lack of eye contact might also give rise to misunderstandings.

In order to improve our understanding of the specific effects of *our* ARbInI system, we investigated in the previous section the influence on eye movements. In this section, we evaluate our system with respect to its influence on head movements, speech and the way in which the participants complete their tasks. In order to analyse the influence of our HMD on the behaviour of the participants, we have again to compare the behaviour of participants using an HMD with a setting where the participants do not wear HMDs. Moreover, we need to quantitatively measure the number and intensity of the performed head movements, the number and duration of utterances as well as the task accomplishment (completion time and error rate).

5.3.1. Hypotheses

Our hypotheses address several features of head movements, the number of utterances as well as performance measures as the task completion time and error rate. For the head movements, we observed two types of head movements so far in our past studies: Firstly, we saw (*communicative*) *head gestures* like head nods, head shakes, head tilts (e. g. to communicate unsureness). These head gestures are used as non-verbal signals during conversation. Moreover, we found vertical and lateral head movements caused by the participants' looking to objects on the table (e. g. in a construction scenario) or for things or people in a room. In this work, we call these latter types of head movements (*searching*) *head movements*.

Searching head movements

The limitation of the field of view caused by the HMD is likely to lead to *increased* head movements. When the participants take a closer look to objects that are spatially distributed, or, when they are searching for objects, there are fewer objects in their field of view than there would be under normal (non-HMD) conditions. Thus, the participants need more head movements to scan the scene. This effect is likely not to be overridden by other effects caused from the weight of the HMD because the participants (in this case) *want* to search for the objects.

Hypothesis 1 (Head movements)

The participants wearing HMDs will increase the number of head movements while searching for objects, compared to those participants that do not wear an HMD.

Communicative head gestures

Looking at the possible reasons for an influence on head gestures as detailed above, our expectation is that head gestures might *decline* while using an HMD. For example, the HMD's weight might cause the participants to feel uncomfortable in moving themselves since they are unsure if the HMD is properly fixed to their head or how far they can move and if the cables are long enough. Another cause might be the restriction of the field of view which causes the object or person in the user's focus to disappear from the field of view (when the users move their head) much sooner than under non-HMD conditions.

Hypothesis 2 (Head gestures)

The participants wearing HMDs will show reduced communicative head gestures (less gestures, lower velocity, shorter gestures) compared to the participants that do not wear an HMD.

Speech

If the head gestures are reduced when using an HMD, a question which might follow from this is if the users tend to substitute their missing head gestures by other signals: for example they might increase their use of vocal back-channelling signals like "mmh", "yes", "no". For example, Kiyokawa et al. (2002) found a tendency to switch from non-verbal communication to speech in their AR-mediated task, if it was difficult for the participants to use non-verbal communication. Apart from head gestures, it is also likely that due to the lack of eye contact, visual back-channelling signals by the listener mostly remain unnoticed. This means, that both the listener and the speaker might compensate for this by *explicitly* giving or demanding verbal feedback instead. Finally, due to the increased focusing on the table, other visual nonverbal signals like eye contact, pointing gestures or body language that transmit e.g. the visual focus of attention are difficult to notice by the interaction partner. Thus, the users might explain this information verbally instead (e.g. "the wind machine which is located in the small room").

Hypothesis 3 (Utterances)

The participants wearing the HMDs will increase their verbal signals measurable by the number and/or length of utterances captured by the speech recognition software.

Performance

The issues of AR seem to hinder the participants severely in how they act. Thus, it is very likely that this also affects their performance in a task in terms of task completion time and error rate. As we have shown in Section 5.2, because of the altered and unfamiliar viewing conditions visual search is more difficult. This might increase the *completion time* for the task. Additionally, the goggles are perceived as uncomfortable. This could lead to a reduced willingness to find the *optimal* solution or impatience and thus to an increased error rate. Moreover, due to the reduced field of view, there are fewer objects in the view so that former errors might be detected less likely along the way while working on other objects.

Hypothesis 4 (Performance)

The participants wearing HMDs will alter the time to complete their tasks compared to those participants that do not wear an HMD. Additionally, they will produce more mistakes than the participants not wearing an HMD.

5.3.2. Method

Concerning the head movements, we expect two phenomena with effects in opposite directions: with HMD more searching head gestures but less communicative head gestures. Since the communicative head movements typically occur in face-to-face conversation while the searching head movements more often occur in an object assembly task, we investigate both phenomena separately in two tasks.

Tasks

In one task, the communicative head gestures are to be investigated. Thus, the participants should communicate as naturally as possible. The topic of their conversation should be irrelevant for the measures but it might be reasonable to ensure that the interaction is distributed more or less uniformly between both participants since we plan to measure both their head movements. Moreover, the other tasks (see below) are more difficult and the wearing of the HMD is straining for many participants so that the first task should be easy. We decided to simply ask the participants to learn to know each other and talk about what they like for about five minutes (see also the scenario description in Section 4.3.7).

The second task aims at investigating the head movements that take place when a user of the system is searching for objects (e. g. on a table). In order to elicit searching, a set of objects has to be available and the task should include repeated search. Easy construction scenarios (e. g. with Lego or building bricks) have been discussed for the second task but the decision

fell in favour of the interactive exhibition design scenario (Section 4.3.4) because this scenario is planned to be used in various other experiments as well and thus offers a comparability with future experiments (Hermann and Pitsch, 2009). Meanwhile, the eye contact between two participants should be minimized since this could provoke communicative head gestures again (which would be an effect in the opposite direction). Thus, the second task should be solved solitarily.

A third task is simply a combination of interaction with object assembly: the participants are asked to discuss their solutions from the previous task and find a joint solution in the interactive exhibition design scenario while merging the objects into one plan.

Each task (or phase) is preceded by an introductory phase in which the experimenter explains the task to the participants and offers to answer questions. These introductory phases are included in the data acquisition since they contain a structured interaction where the experimenter mostly explains and the participants give back-channel feedback or ask questions. Thus, these phases can be used for comparisons to the above tasks.

Apparatus and Stimuli

Since we aim at measuring the influence of AR on interaction (and particularly on head movements) we vary two intensities of load for the participants: in the first group both interacting participants are equipped with video see-through AR goggles while both participants in the control group do not use HMDs. However, not only the influence of the HMD is interesting but also the influence of our whole visual AR approach. Thus, we present the stimuli also in a different way for both conditions: in the control group the pictures are printed on cards with the dimensions 8×6 cm. These cards are vertically affixed to wooden cubes that measure 5.5×5.5×4 cm (see Figure 5.5c). For the HMD group, the pictures of the exhibits are presented using virtual objects (see Figure 5.5a) because participants in our other studies also work with such virtual objects. The virtual objects are depicted on top of ARToolKit markers that are attached to the same wooden cubes as in the non-AR condition (see Figure 5.5b).

Protocol and measurements

Two head-mounted sensors accomplish the measurements of head movements and utterance behaviour: inertial sensors and a microphone. It is necessary to wear these sensors in order to use the automatic classification and annotation features of ARbInI detailed in Section 3 and thus to limit the amount of data that has to be annotated offline for the later analysis. Technically, these two sensors might have an influence on the behaviour of the participants themselves but in previous studies the participants rated the sensors' influence as very low. We consider the benefits from this measurement technique to outweigh this disadvantage. Thus, in both groups (test and control group) both participants wear these sensors.

Five scene cameras capture different views on the study: (a) above the table capturing both participants and the table, (b) frontal view on participant A, (c) frontal view on participant B, (d) whole scene from a sideways' perspective, (e) the laptop screens showing the image that is displayed on the AR goggles. Additionally, ARbInI's system states and the positions of

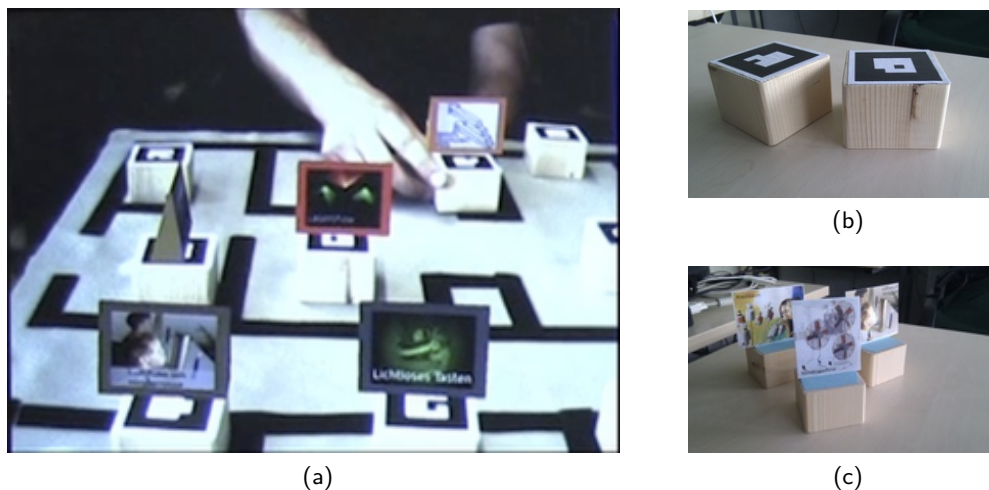


Figure 5.5.: Presentation of stimuli: (a) virtual objects are displayed on top of wooden cubes in the AR condition. (b) wooden cubes with ARToolKit markers without virtual objects. (c) for the non-HMD condition the pictures are printed on cards that are affixed to wooden cubes.

the markers in the field of view are recorded as well as their colouring. All data are equipped with timestamps and stored in the memory database.

Independent Variables

The independent Variables *Intensity of load* (or *Stimuli presentation*) and *Task* arise from the considerations above: *Intensity of load* is a discrete variable with the two values: “with HMD” and “without HMD”. For each trial, two participants share the same condition. The condition is assigned randomly to the participant pairs. *Task* is a discrete variable with the values (explanation below): “smalltalk”, “individual”, “dyadic” (pairwise). Each participant pair is asked to attend all three tasks that are always traversed in the above sequence. Each experimental task was preceded by an introduction phase in which the experimenter explained the following task and the participants could ask questions. Since this is a tutor-listener setting in which typically many head gestures occur, these phases were considered interesting for some of the analyses although the participants did not know that they were to be included in the analysis. For the analyses, the three introduction phases were concatenated to the fourth value of the variable task: “intro”. The study conditions derive from a combination of all values of both independent variables. This two-factorial design results, thus, in 8 conditions that are summarized in Table 5.1.

Dependent Variables

As dependent variables we calculate the time for the completion of the tasks, the participants’ head movements, their speech and their performance:

Completion time All recorded data from all phases are supplied with timestamps and stored to the memory. From this, we can calculate the duration of each study phase for each pair of participants.

		Between subjects	
		Presentation of Stimuli	
Within subjects	Phase/Task	with HMD	without HMD
	intros		
	smalltalk		
	individual		
	dyadic		

Table 5.1.: Experimental conditions for the study.

Head movements The recorded data from the inertial sensors are used to draw conclusions about the amount of head movements in order to allow conclusions about the influence of the HMD condition on the participants' head movements:

- We calculated the overall distance covered by the participants' head movements. For this, we summed the (normalized) absolute values of all three data channels (yaw, pitch and roll) over all time points t :

$$distance = \sum_{t=1}^N |yaw(t)| + |pitch(t)| + |roll(t)|$$

where $yaw(t)$, $pitch(t)$ and $roll(t)$ is measured in [deg/s] and N is the number of time points where a measurement is taken. This measure was computed for each participant and for each phase of the study.

- As a measure for the mean distance we additionally computed the mean value of the measure above:

$$average\ distance = \frac{1}{N} \sum_{t=1}^N |yaw(t)| + |pitch(t)| + |roll(t)|$$

Please note that these measures also include the participants' head gestures. Although we expect contradictory effects in Hypotheses 1 and 2, this is not an issue since the different tasks allow to investigate both hypotheses independently.

Head gestures The head movement data was also analysed for head gestures in several ways:

- To speed up annotation we opted for an automatic annotation using a head gesture classification for the three gestures nod, shake, tilt on the timeseries data from the inertial sensors.¹ The classification used ordered means models and is described in detail in Section 6.3.
- In order to evaluate the classification and to increase our head gesture corpus (see Section 6.3.5), two independent annotators also annotated the head gestures of

¹One of the inertial sensors broke down during the trials so we added WiiMotionPlus sensors to the inertial sensor carriers. At that time, there was no training for the WiiMotionPlus head movement data. Thus, in order to immediately continue with our trials we were forced to rely on offline instead of online classification for the head movements.

nine of the participants manually. The used coding scheme contains *nod*, *shake* and *tilt* gestures as well as the head movement *look* as it happens for example when the participant looks at a third person. Further remarks on the annotation procedure are listed in Section 5.3.2.

- From the classified head gestures and from the manually annotated gestures we computed the number of gestures per class (nod, shake, tilt) per experimental phase.
- We determined for each gesture class the gyroscope channel describing most of the energy (see Section 6.3.4) and calculated on this channel the following measures per gesture: (a) the gesture's overall duration, (b) the number of periods performed during the gesture (c) the period frequency (calculated from the previous two measures) and (d) the maximum and mean velocity (see Figure 5.6 for further explanation).

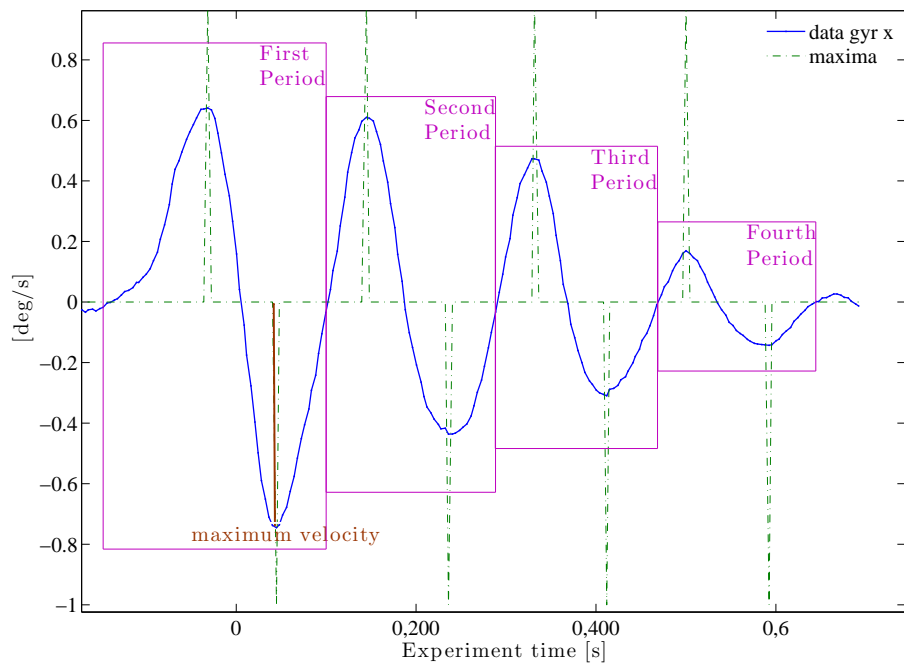


Figure 5.6.: Example nod with measure visualisation: the blue line depicts the data of the x vector from the gyroscope data, the green dotted peaks label the timepoints where extreme values are found in the data (note that their height is always set to 1 for the plot). From these extreme values, we compute the number of periods (magenta boxes). The maximum velocity for this gesture is labelled with a brown line. The duration of this gesture is about 0.65 seconds.

Speech/utterances We analysed the speech data from the participant's microphones with the speech recognition software (see Section 6.2). The resulting phoneme hypotheses

are used to calculate the speech times per participant. Because of performance reasons we recorded in the present study only the resulting hypotheses but not the speech signal itself since the scene cameras additionally recorded the sound.

Performance/Errors The task during the individual and dyadic phase of the study was to design a museum with interactive exhibits (see Section 4.3.4). Some of these exhibits had certain requirements at their surroundings (e. g. dark/silent room). Additionally, there are emissions of some of the exhibits that might affect other exhibits in the room (e. g. one exhibit emits wind which would interfere with an experiment using a candle next to it). The challenge for the participants was to place the exhibits on the museum plan in such a way that the exhibits do not interfere with each other. The final placement of the dyadic task was documented with a screenshot from the scene camera and the objects were annotated by hand. The floor-plan was divided into 7 rooms (see Figure 5.7) and the objects were allocated to these rooms according to the annotations of the final states. From this annotation we rated the performance per participant pair with two methods:

expert error rate For the 16 interactive exhibits that were to be placed during the dyadic task of the study, we created an error matrix rating the interference for each combination of objects (exhibits). This was done by asking three experts (people that designed parts of the study but did not attend the study) to rate each possible set of two objects for their ability to be located in the same room (0: can be located in the same room, 0.5: depends, 1: interfere with each other if in same room). From these ratings, we computed an overall interference matrix. The errors for each object o then could be summed from the respective error values (consulting the interference matrix) of all objects sharing the room with o .

leave-one-out of correctness For each participant pair, we created a matrix rating the correctness for each combination of objects (exhibits). The number of *other* participant pairs that placed the two exhibits in the same room (which means that they *thought* the combination of objects not to interfere) specifies the correctness. Cumulating this correctness for each of the 16 objects in the final state, we calculated the overall correctness for the respective participant pair.

Procedure

The trial consisted of three parts and a questionnaire. First, the participants were equipped with the sensors: microphones, inertial sensors (and according to their respective condition also HMDs). The sensors were explained and the participants were led to adjust the mounting of the sensors so that they could see and move comfortably. When this was finished, the *smalltalk phase* began where the participants were asked to get used to the video recording, the wearable devices and meanwhile get to know each other. After five minutes, the experimenter asked the participants if they felt now comfortable and familiar with the sensors (and the HMD if applicable) or if they needed more time for the acclimatisation.

If the participants agreed to continue, they proceeded with the following two tasks that were using the interactive exhibition design scenario detailed in Section 4.3.4. In this scenario, the

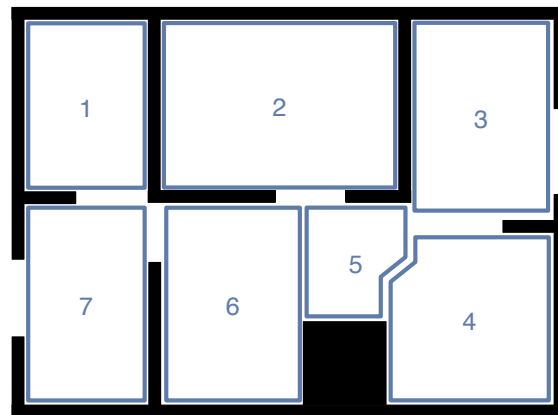


Figure 5.7.: Floor-plan with labels. Objects in room 5 were rated as being in the same room as objects in rooms 4 and 6 while all other rooms were treated solitarily.

participants were asked to place a set of objects representing interactive exhibits in a museum plan in such a way that the exhibits would not interfere with each other. In the *individual phase* the overall 16 objects were divided into two predetermined sets of 8 objects. Each set was assigned to one of the participants. Each participant was then asked to unhurriedly find an arrangement for his or her 8 exhibits alone. The participants could not see their interaction partner and his/her floor-plan during this part of the trial because of a barrier in the middle of the table. Once the participants stated to be finished with this task, the experimenter removed the barrier so that the participants could see each other again.

In a third part, the *dyadic phase*, the participants were asked to discuss their arrangement of the exhibits with their partner and find (again unhurriedly) a joint solution for all 16 exhibits in one of the floor-plans. In the HMD condition, the experimenter additionally explained a feature of AR system: the participants were offered a visual highlighting of the field of view of their interaction partner (see Section 6.1 for details).

Finally, in order to gain subjective results, the participants were asked to complete a *questionnaire* after finishing the third task. After the study the participants were rewarded with €8 per person.

Sample

For this study, we recruited participants from a cross-disciplinary pool. The participants were asked to choose an appointment from a list. Each appointment needed two subscriptions. Thus, the participants themselves created the pairs.

We tested 13 pairs of participants whose age ranged from 18 to 75 years with a mean of 30,6 years. 11 participants stated to be computer scientists or students of computer science. 7 pairs attended the AR condition, their mean age was 30.9 years and 5 of them were female. The remaining 6 pairs attended the non-AR condition, their mean age was 30.3 years and 4 individuals of them were female. Unfortunately, we had to leave out the data from one participant pair because one of the participants chose to stop wearing the HMD after the smalltalk phase because of a sudden feeling of sickness (pair-id two). Since one of the inertial sensors stopped to work in one non-AR trial (pair-id one), we also had to set the

participant aside, who was wearing the malfunctioning sensor. The remaining 23 participants were analysed. There were single tasks where particular sensors stopped working. For these, only the data from the task (but not the whole dataset) was excluded.

Remarks: Head gesture annotations Two annotators tagged the head movements of eleven participants (six from the AR condition and five from the non-AR condition) according to the four gestures *nod*, *shake*, *tilt* and *look*. Table 5.2 shows an overview of the annotated participant pairs and which participant has been annotated by which annotator.

pair-id	with HMD						without HMD					
	4	6	8	9	10	11	1	3	5	7	12	13
A	-	-	-	rl	-	-	rl	-	-	-	-	-
B	-	-	lr	r	lr	-	l	l	lr	-	-	-
A+B	-	-	lr	lr	lr	-	l	lr	lr	-	-	-

Table 5.2.: Head gesture annotation: Participants annotated by the two annotators A and/or B. 'r' is the (from the scene camera's point of view) right and 'l' the left participant.

The task description was to annotate movements like a nod/shake/tilt. Please note, that we include not only confirmative head gestures but also gestures that are used to structure the conversation (e. g. for turn-taking). Thus, we include all movements that *look* like nods, shakes, tilts regardless on their actual communicative meaning. The reason for this is that novice annotators accomplished the head gesture annotations. The annotators were asked to tag only such gestures where the coders would not hesitate about the gesture class. The coders were asked to include in their segments the whole gesture (with their subjective start and end point) with a slight amount of time before and after the gesture if this were possible (sometimes this is not possible since a gesture is preceded or succeeded directly with another gesture or movement so that such surrounding segments would include activity in the sensor stemming from other movements). Please note, that the aim of these instructions was not to annotate *all* occurring gestures during the study but to achieve by means of these annotations a dataset of pure and relatively certain gestures to train the head gesture classification with.

Remarks: Statistical analyses For the statistical evaluations we applied t-tests as well as mixed-between-within analysis of variance depending on the respective research question to answer:

t-tests Our Hypotheses 1 and 2 expect oppositional effects: more head movements during the individual phase when using AR but less head gestures during the smalltalk phase when using AR. Thus, our goal is to learn more about the difference between the HMD and the non-HMD condition per each task. To compare the dependent measure for the condition with HMD with the dependent measure of the condition without HMD, we applied independent two-sample t-tests to each experimental phase. Before applying the t-test, we verified that the samples fulfilled the required preconditions: we applied a two-sample Kolmogorov-Smirnov test to check for underlying normal distributions. Furthermore, a two-sample F-test was applied to check for equality of variances and we

used (depending on the result) the t-test for equal or unequal variances. In the cases where the tests under-reached the α -level = 0.05, we also calculated the effect size η^2 (eta squared) as proposed by Pallant (2005, p. 208). For all these statistical tests, we used the statistics toolbox of MATLAB.

analysis of variance The study used a mixed-between-within design: the stimulus presentation condition is different for each subject (called between-subjects design), while the task conditions are the same for all subjects (called within-subjects design). A mixed-between-within analysis of variance provides results about both main effects (task and stimulus) and interaction effects. For the analyses, we used this kind of test when the research question included the way in which the tasks have an effect on the condition. The assumption of normality was tested with a Kolmogorov-Smirnov-Test (α -level=0.05), covariance equality was tested with Box's test (α -level=0.01 following Pallant (2005, p. 241)) and Levene's test was used to test for equality of error variance (α -level=0.05). All these statistical tests were carried out in SPSS (Version 19.0.1). The program also calculated the estimates of effect size *partial* η^2 (partial eta squared).

Notations and their meanings The following notations used in the results and discussion section:

T(9)=3.09, p=0.013 represents the result of a t-test in the form:

$$T(\langle \text{degrees of freedom} \rangle) = \langle \text{t-value} \rangle, p = \langle \text{p-value} \rangle.$$

F(2,8)=13.07, p=0.003 presents the results of a main or interaction effect of an analysis of variance in the form:

$$F(\langle \text{degrees of freedom} \rangle, \langle \text{error degrees of freedom} \rangle) = \langle \text{F-value} \rangle, p = \langle \text{p-value} \rangle.$$

(partial) eta squared (η^2) Though the p-value indicates if a measured effect can be considered statistically significant, it does not indicate if the measured (significant) effect is a small or great one. This effect size can be calculated using η^2 (eta squared, see Pallant (2005, p. 208)). Eta squared can range from 0 to 1. Cohen (1988, p. 283) proposed the following guidelines to interpret the relative magnitude of the difference measured by the value: 0.01: small effect, 0.06: moderate effect, 0.14: large effect. For example, an effect size of 0.16 calculated for a test comparing our two display conditions in the dyadic task is considered a large effect and means that 16% of the variance in the dyadic task can be explained by the display condition for the respective independent variable (Pallant, 2005, pp. 208).

34:11 reports a duration in $\langle \text{minutes} \rangle : \langle \text{seconds} \rangle$

30 ± 7 denotes: $\langle \text{arithmetic mean (mean)} \rangle \pm \langle \text{standard deviation (std)} \rangle$.

explorative analyses Some differences between the respective tasks are interesting though not covered in the hypotheses explicitly. Hence, we do not evaluate them statistically. Instead, we examine these differences in a simple and explorative way by comparing the means and standard deviations in the respective results table. Similarly, the hypotheses do not incorporate the questions from the questionnaire. Thus, the analyses observe the differences between the display technologies in an explorative way.

The introductory phases preceding each task were concatenated and considered as one task for the analyses.

5.3.3. Results and discussion

This section will present the results of the study. Because of the high number of research questions, the results are alternated with discussions in such a way that a discussion always follows its respective results section. Additional results of the study are also presented in Section 6.3 where the evaluation focus lies on the behaviour of the participants in the non-HMD condition. All recorded data and their automatically added tags have been prepared, filtered and analysed by using the conversion and analysis frameworks described in Sections 3.6 and 3.7.

Completion time per task

The entire study (all 3 tasks plus their introduction) was completed in 20–40 minutes with a mean of 30 ± 7 minutes². How fast did the participants complete the respective tasks? The first task (the smalltalk phase), was terminated by the experimenter after at least five minutes at a time where it seemed appropriate to interrupt the conversation. Thus, this task is not relevant for time comparisons. To compare the participants' completion times for the remaining phases, Table 5.3 gives an overview of the duration of task 2, 3 and the concatenated intros with respect to the experimental condition. The table shows that the participants needed less time in the individual phase compared to the dyadic phase. Additionally, the mean values are much higher in both tasks, the intros and in the overall trial in the AR condition compared to the non-AR condition. A mixed between-within-subjects analysis of variance was conducted to explore the impact of display and task on the duration of the task. There was a statistically significant main effect for task [$F(2, 8) = 13.07, p = 0.003$] with a large effect size (partial $\eta^2 = 0.76$). Additionally, there was a main effect for display [$F(1, 9) = 8.99, p = 0.015$] with a large effect size (partial $\eta^2 = 0.5$). The interaction effect [$F(2, 8) = 0.42, p = 0.67$] did not reach statistical significance. Although we now know that our display groups differ, we do not know for which tasks these differences occur. Applied t-tests comparing the AR condition with the non-AR condition for each task do not find a significant difference for the individual and the dyadic phase in this sample. In the concatenated intro phases however, the participants using HMDs needed significantly more time than the participants without HMD [$T(9) = 3.09, p = 0.013$] with a large effect size ($\eta^2 = 0.49$).

Discussion Our Hypothesis 4 (Performance) expected the participants wearing HMDs to need more time to complete their tasks compared to those participants that do not wear an HMD. Although the t-tests show no significance for the individual and the dyadic phase, the means show a clear difference between the stimulus conditions and the analysis of variance shows a main effect for display. Together, this supports the hypothesis as well as our assumption that the HMD hinders the participants in completing the task in shorter time. Note that the concatenated intros are significantly longer for the HMD group, which shows

²Note, that the time for the completion of the questionnaire was not included in this measurements.

	with HMD		without HMD		
	mean	std	mean	std	
individual phase	5:08	2:33	3:20	1:28	$T(10)=1.41, p=0.189$
dyadic phase	10:25	2:58	7:50	1:58	$T(10)=1.72, p=0.117$
intros	9:44	2:18	6:45	1:46	$T(9)=3.09, p=0.013$
overall trial	34:11	4:08	25:10	5:43	

Table 5.3.: Mean durations of the experimental tasks converted to minutes:seconds. Since the experimenter determined the length of the smalltalk phase, it is not separately listed. Concatenating all introductory phases preceding the three tasks creates the row “intros”. Note that the values for overall trial include all three tasks and the introductory phases between the tasks but not the time used for the questionnaire.

the longer introductions due to more questions and more devices to explain. Additionally, the participants needed less time in the individual phase than in the dyadic phase while the concatenated intros show a mean duration in between. This is also shown by the main effect for task found by the analysis of variance. This effect is not surprising since the participants only had to place half of the objects (8 of 16) into their plan in the individual phase. Moreover, they were allowed to solve the individual task alone while they had to discuss their opinions with their interaction partner during the dyadic phase which explains why this takes more time.

Head movement distances per task

As a measure for the overall amount of movement that is performed, we calculated the head movement distance covered in space summed over all values in all data channels (yaw, pitch, roll) during a task as well as the average distance calculated over the whole task.

Absolute head movement distance covered in space The results are shown in Table 5.4 as well as visualised in Figure 5.8a. Both show that the participants in the AR condition cover less distance in space in the smalltalk and dyadic phase when compared to the non-AR condition whereas they use slightly more head movements in the individual phase. A mixed between-within-subjects analysis of variance was conducted to explore the impact of display and task on the absolute distance. Since the Levene test for equal variances was found significant, we have (for two-way analyses of variance) to set a more stringent alpha level of $\alpha=0.01$ (following Pallant (2005, p. 234)). The test showed an interaction effect [$F(2, 15) = 3.9, p = 0.043$] with a large effect size (partial $\eta^2 = 0.34$) which does not reach significance due to our more stringent α -level but there was a main effect for task [$F(2, 15) = 11.91, p = 0.001$] with a large effect size (partial $\eta^2 = 0.61$). The main effect for display [$F(1, 16) = 3.71, p = 0.153$] did not reach statistical significance. In order to find out if single tasks show significant differences between AR and non-AR, we applied t-tests per task. We can find a significantly lower distance for the smalltalk phase in the AR condition [$T(19) = -2.72, p = 0.013$] with a moderate effect size ($\eta^2 = 0.25$) and for the dyadic phase

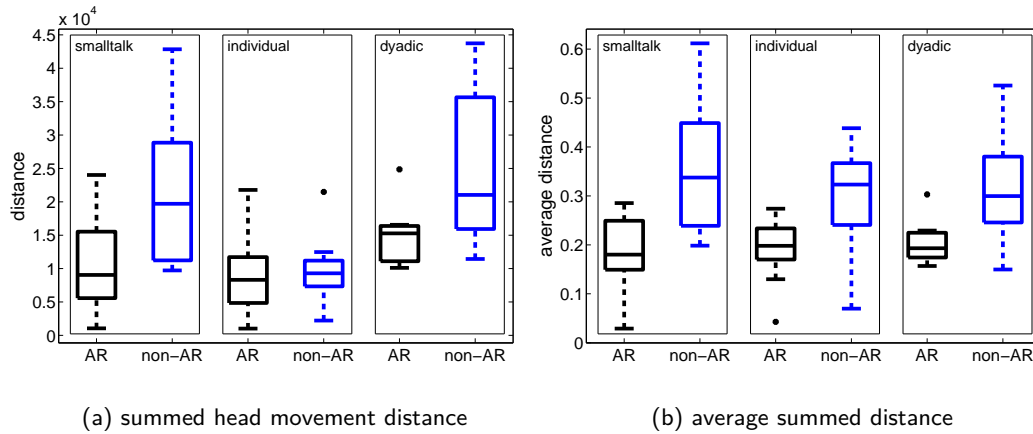


Figure 5.8.: Box plots of the head movement distance covered in space. Each box represents the median (central mark), the 75th percentile (edges of the box) and the most extreme data points (whiskers). Outliers are plotted individually (*). The black boxes depict the participants from the AR condition while the blue boxes depict the participants from the non-AR condition. The outer black boxes frame the phases of the study.

$[T(16) = -2.38, p = 0.03]^3$ with a moderate effect size ($\eta^2 = 0.20$) whereas the distance samples for the individual phase are not significantly different between the two conditions of stimuli presentation.

	with HMD		without HMD		statistics
	mean	std	mean	std	
smalltalk phase	10309	7105	21534	11448	$T(19)=-2.73, p=0.013$
individual phase	9788	6288	9893	4752	$T(20)=-0.04, p=0.965$
dyadic phase	15219	4970	26252	11519	$T(16)=-2.38, p=0.03$

Table 5.4.: Head movement distance in space: $\sum_{t=1}^N |yaw(t)| + |pitch(t)| + |roll(t)|$.

Average head movement distance covered in space Since the duration of the trial differs between the conditions and thus also influences the measured head movement distance covered in space, we also calculated a mean distance. Table 5.5 and Figure 5.8b show that the average distance is (for each experimental phase) lower in the AR condition compared to the non-AR condition. A mixed between-within-subjects analysis of variance was conducted to explore the impact of display and task on the average distance. There was a statistically significant main effect for display [$F(1, 15) = 14.3, p = 0.002$] with a large effect size (partial η^2

³some samples had to be excluded because of recording problems

= 0.49). The main effect for task [$F(2, 14) = 0.19, p = 0.83$] as well as the interaction effect [$F(2, 14) = 0.425, p = 0.66$] did not reach statistical significance. To reveal which tasks lead to the main effect for display, t-tests were applied. The tests show a significantly shorter average distance between the conditions AR versus non-AR: smalltalk phase: [$T(19) = -3.7; p = 0.002$], with a large effect size ($\eta^2 = 0.38$), individual phase: [$T(20) = -2.98; p = 0.007$] with a large effect size ($\eta^2 = 0.29$), and the dyadic phase: [$T(16) = -2.62; p = 0.018$], again with a large effect size ($\eta^2 = 0.24$).

	with HMD		without HMD		statistics
	mean	std	mean	std	
smalltalk phase	0.18	0.09	0.35	0.13	$T(19)=-3.7, p=0.002$
individual phase	0.19	0.06	0.3	0.11	$T(20)=-2.98, p=0.007$
dyadic phase	0.21	0.05	0.32	0.1	$T(16)=-2.62, p=0.018$

Table 5.5.: Mean head movement distance: $\frac{1}{N} \sum_{t=1}^N |yaw(t)| + |pitch(t)| + |roll(t)|$.

Discussion In tasks that encourage interaction (like in the smalltalk phase and in the dyadic phase), the participants seem to cover a greater overall distance if they do not use AR compared to participants using AR. Moreover, the non-AR participants perform their movements in a shorter time so that the average movement distance also is higher than those of the participants in the AR condition. Since head gestures and looking alternately around in the room, at the table or at their interaction partner is not a necessary part of the task, the participants have a *choice* to perform these movements. The willingness to perform those optional movements seems to be reduced under AR conditions to a great extent. Possible reasons can be size, weight and wearing comfort of the HMD as well as lag and blurring of the video stream when moving the head. These problems would additionally increase with the head movement velocity, which also might influence the participants willingness to perform optional movements. In conclusion, the reduced head movements in the smalltalk phase can support our Hypothesis 2 which was that communicative head gestures will be reduced for the HMD condition.

On the contrary, in the individual task, the head movements used for searching objects on the table are necessary for the task. Here, we cannot find a significant difference in the overall movement distance covered in space between the two conditions but a significantly lower mean covered distance for the AR condition in the individual task. The participants seem to need nearly the same amount of movement for the completion of the task (overall distance) but they seem to distribute their movements on a greater amount of time (mean distance). Hypothesis 1 said that the amount of head movements in the individual task should increase induced by the smaller field of view in the AR condition. However, the results show no such increased amount of head movements (neither in the absolute distance nor in the mean distance per task). Why is there no increased amount of movement? A possible explanation can be that the supposed increased head movements are balanced by a reduced willingness of the participants to move their heads at all. This is suggested by the significantly lower distance values in the smalltalk and dyadic phases. Another explanation that there is no

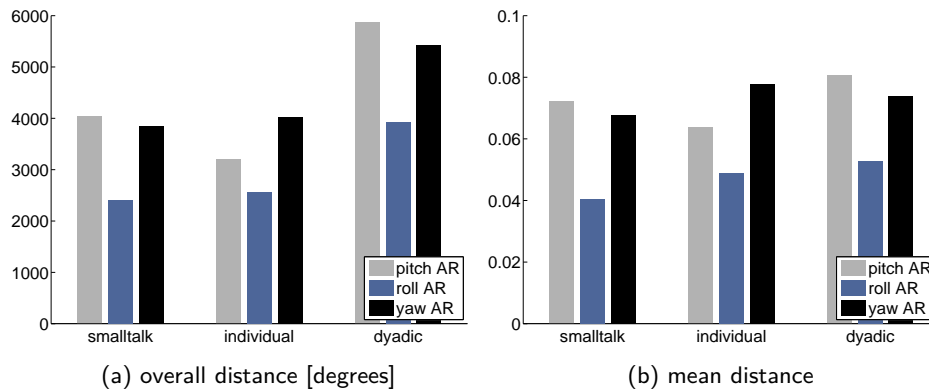


Figure 5.9.: Head movement distance per task per data channel for the AR condition. (a) overall head movement distance (b) mean head movement distance.

difference found between the AR and the non-AR group in the individual phase might be that both groups move their heads in a kind of optimal trajectory where the AR group memorizes the positions of the objects on the table in order to reduce the amount of movement while the non-AR group can use its peripheral view to locate the objects. Another explanation is that the participants using AR find a method to compensate for the restricted field of view. For example, we noticed several participants that leaned back in order to include the whole scene in their field of view. By this compensation, the participants would reduce the necessary amount of head movements to complete the task.

Analysis of single data channels If there is no increased movement in the AR condition for the individual task compared to the smalltalk task, the question arises if this also holds for the single data channels or if the composition of the data channels differs: We could assume that the movement in the individual task would be mostly lateral (measured in the yaw channel) because the objects on the table are distributed more or less laterally. In the smalltalk task, however, we expect (based on observation in a former study with a similar task) mainly vertical movement (measured in the pitch channel) since the communicative gesture that was used most of all (in spontaneous conversation) is the *nod* gesture (see Section 6.3.5). Does the amount of lateral movement increase for the AR condition in the individual phase compared to the smalltalk phase? Although the bar-plots in Figure 5.9 show that the overall as well as the average distance increases slightly between the two tasks (for the lateral movement), neither the overall amount of lateral movement nor the lateral average movement distance increases in a significant way. Likewise, the vertical movement does not decrease in a significant way when comparing the two tasks. Further investigation is surely necessary in order to explain the movement distance phenomena discussed in this section but cannot be provided by this thesis.

Number of head gestures per task

From the annotated head movements, we calculated the overall numbers, the means and standard deviations of annotated gestures per condition. The results are shown per study phase in Table 5.6. Comparing the phases, we can find that a high number of head movements are performed in the smalltalk phase and in the intros for both conditions while the least number of head movements is performed in the individual phase. Moreover, the table shows a difference between the conditions of stimulus presentation: the annotators tagged more head gestures in the condition without HMDs than in the condition with HMD. A mixed between-within-subjects analysis of variance was conducted to explore the impact of display and task on the number of head gestures. Since the Levene test for equal variances was found significant, we have (for two-way analyses of variance) to set a more stringent alpha level of $\alpha=0.01$ (following Pallant (2005, p. 234)). The test showed an interaction effect [$F(3, 7) = 5.75, p = 0.026$] with a large effect size (partial $\eta^2 = 0.71$) which does not reach significance due to our more stringent α -level but there was a statistically significant main effect for task [$F(3, 7) = 9.94, p = 0.006$] with a large effect size (partial $\eta^2 = 0.81$). The main effect for display [$F(1, 9) = 2.4, p = 0.156$] did not reach statistical significance. Although the main effect for display is not significant, it is interesting to investigate the single tasks with t-tests (since the tasks are designed to elicit different and for Hypotheses 1 and 2 oppositional effects). Applied t-tests to all tasks show that this difference is only significant for the dyadic phase: [$T(4.4) = -5.42, p = 0.004$] with a large effect size ($\eta^2 = 0.77$). For these statistics, t-tests with unequal variances were used since the applied two-sample F-test revealed unequal variances.

phase	with HMD			without HMD			Σ	statistics
	abs	mean	std	abs	mean	std		
smalltalk	124	20.67	9.62	180	36	25.53	304	T(4.9)=-1.14,p=0.307
individual	11	1.83	2.03	11	2.2	2.23	22	T(8.2)=-0.25,p=0.805
dyadic	7	1.17	1.07	67	13.4	4.41	74	T(4.4)=-5.42,p=0.004
intros	148	24.67	14.41	134	26.8	12.66	282	T(8.9)=-0.24,p=0.819
overall	290			392			682	

Table 5.6.: Head gesture annotation: Numbers, means and standard deviations of head gestures per phase and per condition. The line called 'intro' gives the gesture occurrences during all three introductions to the tasks.

Discussion The analysis of variance found a significant main effect for task. The least number of head gesture annotations occurred in the individual phase, which is not surprising since the participants were asked not to speak during this task. The occurrences hence were only *look* movements. Most of the head gestures occurred during the parts of the study where natural conversation took place: in the smalltalk phase where the participants were free to talk about a topic they liked for five minutes while getting used to the sensors (and the goggles if applicable) and in the intro phases where the investigator explained the next task to the participants and the participants listened and asked further enquiries. In both cases the

participants had no further task to do than to listen and ask questions, especially, their eyes were free to look at their interaction partner. This seems to encourage confirmative head gestures no matter if they are wearing AR goggles or not. Although the main effect for display was not significant as well as the interaction effect, the t-tests found that during the dyadic phase, there are significantly more head gesture occurrences annotated in the non-HMD condition than in the HMD condition while on the same time there are much less gesture annotations than in the smalltalk phase (not significant). This suggests that the amount of head movements is influenced by the task as well as the stimulus presentation condition: tasks that require the focussing of the eyes on objects (individual and dyadic phase) seem to reduce the number of head movements even if a conversation or discussion takes place (dyadic phase). Additionally, the number of performed head gestures seems to be further reduced by the HMD, an effect which is only significant for the dyadic phase but (although the mean values might suggest otherwise) not for the smalltalk phase. In order to discuss this deeply, it would be useful to design a further experiment comparing a discussion and conversation situation with equal requirements for the participants' visual focus of attention.

Number of gestures per gesture class

Here, we compare the number of gestures per gesture class that were annotated by hand with the gestures that were tagged by the classifier.

Gestures annotated manually Table 5.7 allows us to compare the number of head gestures per gesture and per condition. We can find that the gestures *nod* and *look* were the gestures most annotated while *shake* and *tilt* occurred less often in both conditions. Comparing both conditions, we find higher occurrences for each gesture in the condition without HMD compared to the condition using an HMD. Additionally, we find high standard deviation values. Applied t-tests per gesture show no significant results.

	with HMD			without HMD			Σ	statistics
	overall	mean	std	overall	mean	std		
<i>nod</i>	126	21	10.77	182	36.4	17.33	308	T(9)=-1.63,p=0.138
<i>shake</i>	11	1.83	1.21	33	6.6	7.23	44	T(9)=-1.44,p=0.185
<i>tilt</i>	5	0.83	0.69	8	1.6	1.36	13	T(9)=-1.09,p=0.302
<i>look</i>	148	24.67	19.39	169	33.8	21.08	317	T(9)=-0.68,p=0.516
overall	290			392			682	

Table 5.7.: Head gesture annotation: Number of head gestures per gesture class and per condition. The mean and std values are derived from mean and std values computed over the entire trial per participant.

Discussion The results indicate that the amount of head gestures is slightly but not significantly reduced when the participants use an HMD. The gesture classes occurring most often are the *nod* and the *look* movements, which we also observed in previous studies (see 6.3.5). Additionally, the high standard deviations seem to indicate (as in previous studies) that the

number of head gestures performed by participant is highly individual (see also Section 6.3.5 for an overview of the head gesture corpus built from the annotated head gestures from all studies of this thesis). Thus, the applied t-tests do not find significant results.

Gestures tagged by the automatic classifier Table 5.8 shows the same values as above, except with another tagging method: here the classification module was used (see Section 6.3 for a more detailed overview about this module). Comparing the numbers of gestures tagged by the classification from the trials using an HMD with those using no HMD, we can find a high difference between the two conditions: for every gesture class the classification finds more than twice as many occurrences in the non-AR condition compared to the AR condition. Applying t-tests, this difference is significant for *nod* [$T(12.1) = -3.48; p = 0.004$] with a large effect size ($\eta^2 = 0.37$) and *shake* [$T(10.9) = -2.89; p = 0.015$] with a large effect size ($\eta^2 = 0.28$) whereas it is not significant for the *tilt* gesture. For these statistics, t-tests for unequal variances were applied since the two-sample F-test revealed unequal variances.

	with HMD			without HMD			Σ	statistics
	abs	mean	std	abs	mean	std		
<i>nod</i>	625	52.08	41.08	2046	186	121.37	2671	$T(12.1)=-3.48, p=0.004$
<i>shake</i>	390	32.5	24.63	1447	131.55	111.19	1837	$T(10.9)=-2.89, p=0.015$
<i>tilt</i>	324	27	31.99	752	68.36	70.09	1076	$T(13.7)=-1.79, p=0.095$
overall	1339			4245			5584	

Table 5.8.: Head gesture classification: Number of head gestures per gesture class and condition.

Discussion The results of the classification indicate that there are significantly less head gestures performed under AR conditions than under non-AR conditions. This supports Hypothesis 2 which said that communicative head gestures will be reduced for the HMD condition. In the following, the classification results are discussed with respect to the annotation results.

The comparison of the overall numbers of gestures from the classification with the same values from the annotated gestures show that many more intervals were tagged by the classification than by hand. This is not surprising since all participants' head movements were automatically tagged (24) while manual annotation took only place for 11 participants. But also the mean numbers of gestures per participant are higher compared to the mean numbers of gestures above. This is the case because the classification was planned to tag rather false-positives than to miss actual gestures. Thus, the thresholds for gesture annotations were levelled accordingly low (see Section 6.3 for a discussion about this topic). Apart from that, we can find more *shake classifications* than *annotations* because the classification does not distinguish between *shake* and *look* gestures. Finally, there are much more *tilt* gestures tagged by the classification than have been annotated by hand. On the one hand, the reason could lie in the classification (e. g. that the threshold for classification of a *tilt* gesture needs to be reviewed or that the classification needs to be re-trained since the training set is still to small). On the other hand, it could be that the human annotators missed several *tilt* gestures.

For the latter, possible reasons could be that the annotator did not rate the movement being a gesture since it occurs directly adjacent to another gesture and is less obvious than the adjacent one or that *tilt* gestures generally tend to escape the notice of human annotators more than the other gestures to be annotated (e. g. because they are shorter and less visible than the other gestures). The reliability of the manual and automatic tags will be investigated further in Section 6.3

In conclusion, the actual numbers of gestures differs between the annotations and classifications due to the configuration of the classification to produce rather false positives than false negatives (which leads to much higher mean numbers of gestures in the classifications). Therefore, we will limit the following analyses on the data derived from the annotations. Additionally, we will discuss in Chapter 6.3.3 the reliability of annotations and classifications.

Nevertheless, the results of both approaches agree in their tendencies (more gestures for non-HMD condition, more gestures for *nod* than for *tilt*). This indicates that the wearing of an HMD alters the head movements in such a way that either significantly less gestures occur during the interaction or that a classification/annotation of head movements is highly disturbed by the altered execution of head movements.

Given a correct tagging of head gestures in the interaction, this work provides a general method to analyse rich measures of such annotated and classified patterns. The following sections will analyse duration, the number of periods, the period frequency and the gesture's velocity.

Duration of the gestures per gesture class

Table 5.9 shows the mean durations of the four gesture classes per condition as well as their standard deviations. The duration of an average gesture seems to differ between the conditions: For the condition with HMD the longest gesture is the *look* movement while the shortest is the *tilt* gesture. For the non-HMD condition the *nod* gesture has the longest duration while the *tilt* gesture has again the shortest mean duration. Applied t-tests find no significant differences between the two stimulus presentation conditions for the durations in the gesture classes.

	with HMD		without HMD		statistics
	mean	std	mean	std	
<i>nod</i>	1.6	0.34	2.05	0.35	T(9)=-2.17,p=0.058
<i>shake</i>	1.48	0.81	1.37	0.32	T(9)=0.29,p=0.776
<i>tilt</i>	0.73	0.62	0.98	0.64	T(9)=-0.66,p=0.526
<i>look</i>	1.69	0.9	1.24	0.31	T(9)=1.06,p=0.316

Table 5.9.: Head gesture annotation: Duration (in seconds) of head gestures per condition.

Discussion Our Hypothesis 2 was that the duration of gestures will decrease when using AR compared to the non-AR condition. Although the means seem to support this, the t-tests

found no such significant effects. This indicates that the length of the inspected gestures is not significantly affected by the wearing of an HMD. On the other hand, the non-significant result might also be due to the very small number of annotated subjects.

In our sample the *tilt* gesture seems to have a shorter duration than all other gestures (although the difference is not significant). This could support the theory that human annotators might tend to overlook those kind of gestures as we discussed several paragraphs before which might explain the lower occurrences of the gesture in the annotations done by humans than in the classifications performed by the software. A larger sample would be necessary to investigate these two considerations.

Analysis of head gesture events using the time series data

Having annotated the head gestures in the timeseries data, we can use them to investigate the characteristics and the differences between the single head gesture classes. For this, we reduce the data to the channel representing most of the variance (see Section 6.3.4). The following paragraphs will present the results for the calculated absolute extremum per gesture, the number of periods per gesture as well as the period frequency. Please refer to Figure 5.6 for explanation about the calculation of the measures. Since there are altogether only about 10 *tilt* gestures in all conditions, we omit the measures for this gesture.

Extreme values An extremum is calculated as the highest value between two zero points. A maximal value of all occurring extreme values is calculated per annotation. From this we calculate the mean value and standard deviation occurring over all annotated participants. Table 5.10 lists the results. Comparing the means between the conditions we can find a small but significantly higher mean extremum for *nod* in the non-HMD condition compared to the HMD condition [$T(8) = -2.92; p = 0.019$] with a large effect size ($\eta^2 = 0.52$). Although the means indicate a similar effect for the look movement, this is not significant. For the shake movement, there is no such difference in the means. Comparing the gestures, we can find the highest mean values in the look gesture.

	with HMD		without HMD		statistics
	mean	std	mean	std	
<i>nod</i>	0.51	0.29	0.82	0.42	$T(8)=-2.92, p=0.019$
<i>shake</i>	1.07	0.21	1.08	0.52	$T(9)=-0.03, p=0.977$
<i>look</i>	1.96	0.41	2.56	1.15	$T(9)=-0.82, p=0.433$

Table 5.10.: Maximum rotational velocity during gesture.

Number of periods per gesture class The number of periods per gesture is the number of extreme values found in a gesture divided by two. Table 5.11 displays the mean and standard deviation for the number of periods per gesture. Comparing the means we find a significantly higher number of periods in the non-AR condition compared to the AR condition

for *nod* [$T(8) = -2.89, p = 0.02$] with a large effect size ($\eta^2 = 0.45$) as well as for *shake* gestures [$T(9) = -2.72, p = 0.023$] with a large effect size ($\eta^2 = 0.43$).

	with HMD		without HMD		statistics
	mean	std	mean	std	
<i>nod</i>	2.98	1.54	4.39	1.93	$T(8)=-2.89, p=0.020$
<i>shake</i>	1.9	0.42	3.74	1.51	$T(9)=-2.72, p=0.023$
<i>look</i>	1.56	0.13	1.05	0.19	$T(9)=1.21, p=0.257$

Table 5.11.: Number of periods for the gestures.

Frequency of periods This measure calculates the frequency of periods per gesture. For this, we divided the number of periods by the duration of the respective gesture. Table 5.12 lists the mean period frequencies and the standard deviations. Comparing the means, we find a significantly higher period frequency for the shake gesture during the non-AR condition than during the AR condition [$T(9) = -2.82; p = 0.02$] with a large effect size ($\eta^2 = 0.44$). Although the means indicate this effect also for the nod gesture, this is not significant. We omitted the values for the look movement, since it has no periodical structure (see table above).

	with HMD		without HMD		statistics
	mean	std	mean	std	
<i>nod</i>	3.96	1.44	4.39	1.15	$T(8)=-0.75, p=0.476$
<i>shake</i>	2.96	0.31	5.85	1.86	$T(9)=-2.82, p=0.02$

Table 5.12.: Frequency of periods for the gestures. Here, the look values are omitted since the look movement is not periodical (see discussion).

Discussion The results regarding the *extreme values* indicate that there are, for the nod gestures, significantly higher rotation velocity extreme values in the non-HMD condition compared to the HMD condition. This means that the participants reach a higher overall velocity in their movements when nodding without wearing an HMD. This supports our Hypothesis 2 that expected the movement velocity to decrease when using AR compared to non-AR. Interestingly, this effect cannot be found for the other gestures. A possible explanation for a greater influence of the HMD on the nod movement than on the other movements could be due to the method of mounting the goggles to the head. In the horizontal axis, the HMD is fixated by a bungee cord around the head distributing a possible pressure to a large portion of the head. Thus, when shaking the head, the pressure on the head is not likely to be increased significantly. In the vertical axes, however, the HMD is resting solely on the root of the nose and thereby on a small point. When nodding, the HMD is accelerated with the movement of the head, which might increase the pressure on the root of the nose significantly. Several users complained about the general pressure of the HMD on the nose.

We already knew about this problem prior to the study and glued foam rubber to the goggles at the point that is lying on the nose but this finding seems to show that this modification is not sufficient. A subsequent study should investigate the question how the pressure on the nose could be lessened. The highest extreme values occur during the look movements, which means that this movement reaches the highest velocity. A possible explanation for this could be that movements that are not sinusoidal can be performed faster since the movement is not alternating in opposite directions.

For the *number of periods* we find high period values in the gestures nod and shake and low period values for the look movement which indicates that the former are more sinusoidal movements with more repetitions while the look movement is only seldom occurring as a periodical movement. The reason is, that the look movement is often split into single linear movements (look right) when an additional movement back to the origin is annotated as an extra gesture because there is often some time passing in between. Furthermore, we find higher repetition values for the non-AR condition for the gestures nod and shake compared to the AR condition. This shows that there are less repetitions of the gestures under AR conditions which supports our Hypothesis 2 that the HMD reduces the communicative head gestures. Since the (decaying) periodic repetitions can be seen redundant, the gesture might be cut shorter than usually by the participants wearing HMDs in order to cope with the difficulties burdened by the AR goggles.

For the *period frequency* we found a significantly higher frequency for the shake gesture in the non-AR condition. This means, that also the repetition velocity for the movement decreases during AR conditions which again supports our Hypothesis 2 (Head gestures).

Summarizing the effect of the AR goggles, we find for the nod gesture a reduced maximal velocity as well as a reduced number of periods (and thus repetitions) while the period frequency shows no significant effect. For the shake gesture, the maximal velocity is unaffected by the HMD while the number of repetitions as well as the repetition velocity decreases. Since the look movement is not a gesture but is used to look at a fixed position, the covered distance should stay the same: The duration of the look gestures increases slightly (but not significantly). Thus, it is not surprising that the maximal movement velocity decreases slightly (also not significant) which indicates that the participants use slightly more time for the same movement. However, we should consider the possibility that the non-significant results can be due to the very small number of annotated subjects. A greater sample would allow more reliable conclusions.

Speech

If the AR condition influences the way and amount of communicative head movements, it could also be that the technique influences the amount of speech during a trial. We used the results from the speech recognition and compared the number of utterances, the mean duration and the summed duration of all utterances. The results are listed in Table 5.13. The individual phase was omitted since speech was not allowed during this task. A mixed between-within-subjects analysis of variance was conducted to explore the impact of display and task on the utterance duration. There was a statistically significant main effect for task [$F(2, 19) = 42.97, p < 0.0005$] with a large effect size (partial $\eta^2 = 0.82$). Additionally, there was a main effect for display [$F(1, 20) = 9.13, p = 0.007$] with effect size (partial $\eta^2 = 0.31$).

The interaction effect [$F(2, 19) = 3,49, p = 0.051$] (partial $\eta^2 = 0.27$) did not reach statistical significance though the p-value is near to our α -value. To identify the tasks which lead to the main effects, we applied t-tests: the tests reveal a significantly higher number of utterances in the AR condition than in the non-AR condition for the dyadic phase [$T(20) = 2.66; p = 0.015$] ($\eta^2 = 0.24$) and for the concatenated intro phases [$T(22) = 2.46; p = 0.022$] ($\eta^2 = 0.22$). The smalltalk phase shows no such difference.

phase	with HMD			without HMD			Σ	statistics
	abs	mean	std	abs	mean	std		
smalltalk	954	86.7	38.2	915	76.3	31.6	1869	$T(21)=0.72, p=0.48$
dyadic	1539	139.9	44.4	1029	93.6	37.0	2568	$T(20)=2.66, p=0.015$
intros	672	56.0	35.4	326	27.2	20.0	998	$T(22)=2.46, p=0.022$
overall	3165			2270			5435	

Table 5.13.: Number of utterances in the different study phases and conditions.

Discussion If the participants under AR reduce their amount of communicative head movements, do they compensate these missing head gestures with further verbal remarks? There was a main effect found both for task and for display. Additionally, the t-tests found that the number of utterances increases significantly in the AR condition for the dyadic task. This supports our Hypothesis 3 which expected the amount of utterances to increase under AR conditions compared to non-AR. This does not necessarily mean that the missing head movements are compensated by more utterances. Other reasons are also possible: The increased amount of utterances may coincide with the (although not significantly) longer duration of the trial. Or, the increased number of utterances might result from the missing joint attention caused by the missing eye contact. For example, the participants might need more words to explain the object that they are talking about, to manage turn-taking or to guide their partner's attention. This phenomenon of increased utterances should be regarded and further analysed in future research. Likewise, the number of utterances increases significantly in the AR condition for the intros where the experimenter explained the tasks. This is not surprising, since the experimenter had to explain not only the task (as in the non-AR condition) but also how to work with virtual objects and to explain the highlighting of objects in the partners field of view. However, there is no significant difference found for the smalltalk phase. This suggests that the number of utterances in a pure conversation (without a specific task including work on a table) is *not* very much affected by the use of HMDs. Why did the t-tests find a significant effect for the dyadic task (and the intros) but not for the smalltalk task? While the visual focus of attention is in the dyadic task mainly on the table, it is in the smalltalk phase mainly on the interaction partner's face. Although the wearing of the HMD blocks many nonverbal signals, head gestures can be used as back-channel signal. But for this, the receiver of the signal has to look at the transmitter. This is much more unlikely when wearing an HMD since the field of view is very small. Thus, there might be less to compensate for during the smalltalk phase than during the individual task (as well as in the intros where the participants both look at the objects on the table that are to be explained by the experimenter.

Performance

The task during the individual and dyadic phase of the study was to design a museum with interactive exhibits (see Section 4.3.4). Some of these exhibits had certain requirements towards their surroundings (e. g. dark/silent room). Additionally, there are emissions of exhibits that on the other hand might affect other exhibits in the room (e. g. one exhibit emits wind which would interfere with an experiment using a candle next to it). The challenge for the participants was to place the exhibits on the museum plan in such a way that the exhibits do not interfere with each other. The finish placement of the dyadic task was documented with a screenshot from the scene camera and the objects were annotated by hand.

Expert error rate From this annotation we rated the performance per participant pair in comparison with an expert rating (see Section 5.3.2 for details about this expert error rate). As results, we found 12.2 ± 1.99 misplacements for the HMD condition while the participant pairs in the non-HMD condition did 16 ± 2.93 misplacements. Note that for the error rates, small values indicate a small error rate. Applying a t-test shows that this effect is significant [$T(9) = -2.46, p = 0.036$] with a large effect size ($\eta^2 = 0.27$).

Leave-one-out correctness With the expert error rate, we noticed that there were some of the possible sources of interference which were found by very few of the participants (e. g. the interference of sand with wind or the interference of smoke with the ability to smell). In the expert error rate table, these sources of interference were naturally included. To compare the expert error rate with a performance measure that reflects the sources of interference found by the participants, we designed the “leave-one-out correctness” as a performance measure that is computed from the results of all participants (see again Section 5.3.2). With this, we calculated the correctness based on all other participant pairs.

Here, we found 37 ± 13.13 correct ratings for the HMD condition while the participant pairs in the non-HMD conditions had 47 ± 16.15 correct ratings (note that now small values indicate few correct placements). Applying a t-test shows no significant effect [$T(9) = -1.11, p = 0.296$].

Discussion Hypothesis 4 (Performance) was that the participants using the HMDs would make more mistakes than the participants not using HMDs. However, our results show the opposite: the HMD-group makes significantly less mistakes than the non-HMD group for the expert error rate. This result is surprising since the HMD group has no known advantage over the non-HMD group. The stimuli are the same except they are presented differently. A possible reason for this might be that they use more time for the completion of their tasks (see completion time) and thus might perform the task more thoroughly and have more time to think about the objects.

When interpreting the two performance measures, the results show that for the expert error rate, we found less errors in the HMD condition compared to non-HMD while for the leave-one-out correctness we found less correct placements in the HMD condition compared to non-HMD (though not significant). This means we cannot confirm the error rate part of Hypothesis 2 (Head gestures).

Apart from this, the participants were not informed about when exactly two exhibits would interfere. With this complex task, there are multiple possible solutions. For our calculations, we defined rooms in the floor-plan. It is possible for example, that the rooms 4,5,6 were seen as only one room by the participants while we defined 2 rooms with room 5 expanding both rooms 4 and 6. Additionally, big rooms might have been interpreted big enough for two interfering objects. Some participants also invented mobile partition walls that they used to partition rooms. Finally, instead of defining all exhibits in one room to possibly interfere with each other it might be that the participants thought only adjacent objects to be possible to interfere. A specific definition of the concept of interference should be included in future introductions to the task.

Coordination and joint attention

Section 5.1 discussed possible effects of the HMD on the interaction: since the field of view is restricted, the participants can either look at their interaction partner or at the task (e. g. objects) on the table. This means that head movements, facial expressions and eye contact are not available as communicational cues in those intervals where they focus on the table. In Schnier et al. (2011b), we investigated the users' coordination and their methods to establish joint orientation under these conditions using conversation analysis. We found a development in the method how the participants referred to objects during the task. First, the reference was given only verbally, then the verbal reference was combined with a gestural pointing to the object. Subsequently, the explicit perspective of the listener was given ("from your perspective in the upper right corner"). In the later cases, the participants seemed to develop a new method: they lifted the object to focus from the floor-plan for some seconds which allows their interaction partner to orient towards this object more easily. This method is further improved during the task by adding hand movements to the lifting movement. The authors propose to implement a technical method that facilitates the establishment of joint orientation (see Schnier et al. (2011b,a) for further details).

Questionnaire

The following paragraphs will present and discuss the results from the questionnaire that the participants were asked to fill in after finishing all tasks. We used two questionnaires: one for the participants of the AR condition and another that omitted questions regarding the AR technique and the HMD for the participants of the non-AR condition. The original questionnaires in German and their translations can be found in the Appendix A.1. For the answers to the questions, we used a rating scale ranging from negative to positive with four possible choices and not explicitly named intermediate steps. Since the questions of the questionnaire were not included in the hypotheses, we do not analyse them using inferential statistics but analyse the results using plots. For each question, we provide a bar plot showing the percentages of participants in the respective group that chose this answer. To facilitate the comparisons between the respective groups, the text additionally gives percentages summed over answers: one sum for each side of the scale.

Simplicity of the task “How simple did you find the exhibition-design-task?”. The participants could choose from a four-choices scale ranging from “very difficult” to “very simple” with no explicitly named intermediate steps. The answers of the participants are presented in Figure 5.10a divided into the two groups (AR, non-AR). Although the participants of the AR condition rated the task less simple than the non-AR group, still most of the participants (71% in the AR condition and 91% from the non-AR condition) rated the task as simple. The questionnaire included free space for individual comments of the participants. Many participants wrote positive comments regarding the museum design task. Comments included that they liked the scenario and task and found it easy to understand though being an interesting and complex task or that they had fun during the study. Some participants proposed possible applications for the AR system. Some complained about challenges in the task and would have liked to know more about the interactive experiments previous to the task.

Comfort of system For the question “How do you rate the wearing comfort of the system?” the results are presented in Figure 5.10b. While most of the participants in the non-AR condition rated the system as comfortable, only 36% of the AR-participants stated the system to be comfortable. Thus, we can find a difference between the two conditions in the perceived comfort which seems to be induced by the AR goggles. In the comments several participants from the AR group complained about the AR goggles to be strenuous to use (reasons mentioned were: field of view, image resolution, lag, weight and pressure on the head/nose). Four participants experienced slight forms of nausea or headache. From the non-AR participants however, nobody experienced nausea or complained about the trial being strenuous. Only some participants complained about (as they wrote) “minor problems” with the inertial sensor (being not optimally mounted to the head) or the cables. Some of the AR participants also complained about the influence of the AR goggles on the interaction. The missing eye contact with their interaction partner was rated as unnatural and difficult as well as the perceived distance to the objects.

Naturalness of the participants’ head movements “How natural do you rate your head movements during the phases of the experiment?” Figure 5.11 shows the results: in the left bar plot for the AR questionnaire and in the right bar plot for the non-AR questionnaire. The participants rated their head movements in the smalltalk task similarly. In each condition 50% of the participants rated their head movements as natural. For the subsequent tasks, the participants wearing AR goggles rated their head movements generally less natural than the participants from the non-AR condition. Precisely, in the individual phase, 91% of the non-AR participants rated their head movements as natural while only 57% of the non-AR condition chose the left side of the scale. In the dyadic phase 92% chose the left side of the scale while only 28% of the participants wearing an HMD chose this side of the scale. Another interesting finding from these results is that the non-AR group rated their head movements during the smalltalk task less natural than in the other two tasks: in the non-AR condition 50% of the users rated their head movements as natural while for the other tasks 91% (individual) and 92% (dyadic) rated their head movements natural.

Why could the head movements be perceived as less natural during the smalltalk phase? Possible reasons for this are that they had no challenging task to do during this phase (except

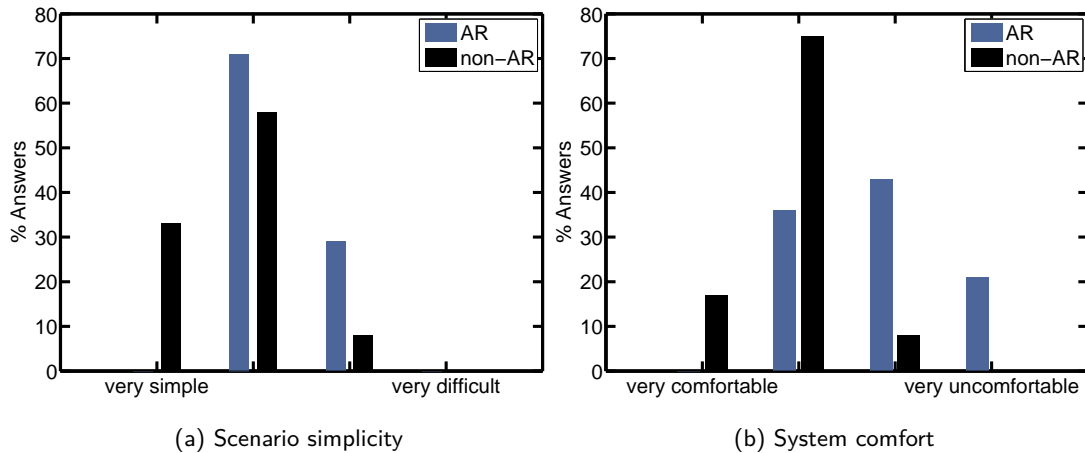


Figure 5.10.: Questionnaire: (a) How simple did you find the exhibition-design-task? (b) How do you rate the wearing comfort of the system?

small-talk) and thus might have noticed the system in this phase more than in the other ones. Moreover, the task was designed (and explained) to the participants to be for the smalltalk with the system. This means, the effect could be caused by the actual smalltalk and by the expectation that they needed smalltalk.

Interference from the system components “How much did the devices mounted to your head interfere with your actions?” “How much did the video cameras stationed in the room influence your interaction with your partner?” Both questions had a four-choices scale ranging from “very much” to “very little” with no explicit intermediate labels. A fifth choice was labelled “not at all”. Figure 5.12 shows the answers of the participants to these questions. Again, the results from the AR condition are shown in the left bar plot while the participants’ answers from the non-AR condition are presented in the right bar plot.

The plots show that *all* participants in the non-AR conditions rated the disturbance from both, sensors and the cameras, as little. Actually, more than 50% even chose the box “not at all”. In the AR condition, the microphone was also rated as disturbing very little or not at all. The same holds for the video cameras except for one participant that rated the cameras as very much influencing. The inertial sensors were also perceived by one of the participants as interfering while the remaining participants chose the right part of the scale or even the box “not at all”. Different results can be found for the AR goggles: 79% of the participants rated the goggles as interfering with their actions whereas the remainder rated the goggles either as little interfering or “not at all” interfering. In summary, the cameras, the microphone as well as the inertial sensors were perceived by nearly all participants as little influencing or interfering with their actions while the AR goggles were perceived by most of the participants as interfering with their actions.

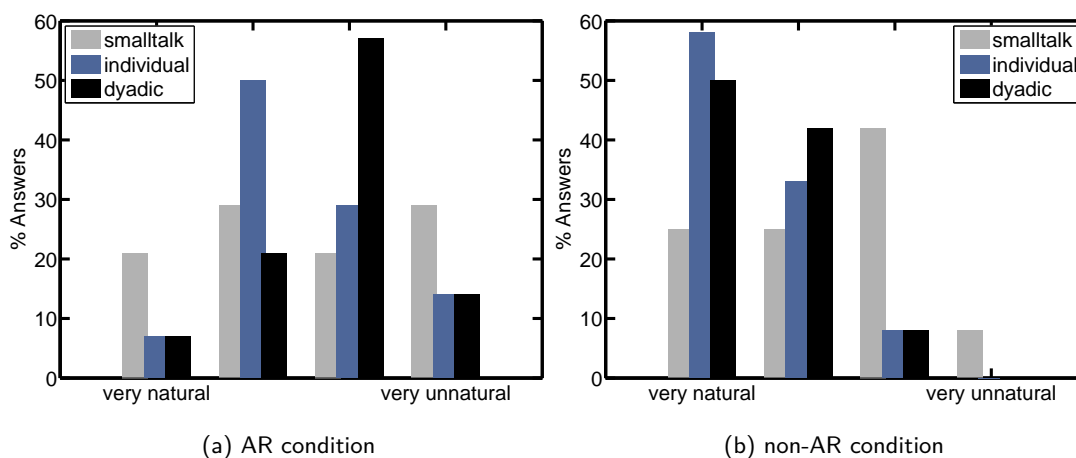


Figure 5.11.: Questionnaire: How natural do you rate your head movements during the experimental phases? (a) Answers of the participants in the AR condition (b) Answers of the participants in the non-AR condition.

5.3.4. Summary

This study investigated how exactly the users' head movements are affected by the HMD and how speech and task completion are influenced. Confirming our Hypothesis 2, we found fewer head gestures when using AR, especially in tasks that demand visual attention (as the dyadic phase did). Additionally, we found slower velocity, fewer periods and lower period frequency for specific gestures (nods or head shakes). Together, the results from the head gesture analysis suggest a reduced willingness of the HMD users to perform head movements if not necessary. The duration of the gestures, however, seems not to be affected significantly by the stimulus presentation method. Other than expected by Hypothesis 1, we found not more head movements when searching for objects using AR (in the individual task). We discussed possible reasons for this, for example a reduced willingness of the participants to move their heads and compensation methods that reduce the number of head movements (leaning back to include the whole scene in the field of view). On the other hand, we found a higher number of utterances in the AR versus the non-AR condition when the visual focus of attention was on the table (as it is in the dyadic task) and thus could confirm our Hypothesis 3. This suggests that some disadvantages of the HMD are compensated by increased use of verbal language (e.g. confirmations, back-channelling, requests). In the study, the subjects using AR needed significantly more time to complete their tasks than those not using AR (as proposed by Hypothesis 4). The expected reduced performance for the AR participants, however, was not supported by the data. One participant pair was also analysed by conversation analysts who reported changes in how the participants establish joint attention under AR.

The questionnaire results showed a high acceptance of the scenario used in the individual and dyadic phases of the study. Most of the sensors (and the video cameras) seem not to hinder the participants in their actions. But the AR goggles were rated especially uncomfortable and were rated interfering with the participants' actions. Moreover, the AR goggles seem to

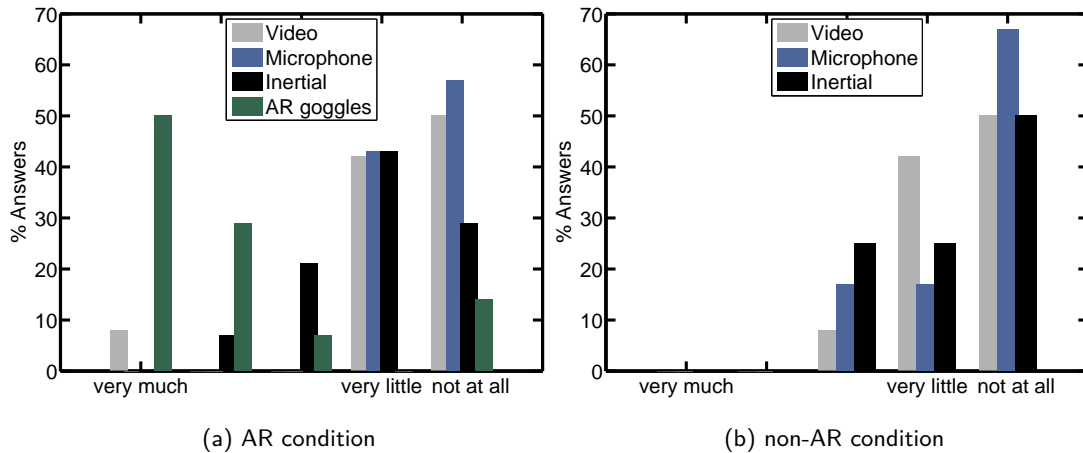


Figure 5.12.: Questionnaire: How much did the devices mounted to your head interfere with your actions? How much did the video cameras stationed in the room influence your interaction with your partner?

affect the perceived naturalness of the head movements, which coincides with the findings from the head gesture analyses. Thus, the participants seem to be aware of the AR goggles during the entire trial and they seem to believe them to actively disturb the interaction and their movements.

5.4. Summary and Discussion

In this chapter, Section 5.1 presented a detailed overview of the known drawbacks of Augmented Reality through HMDs and their possible effects on the behaviour of their users. The ways in which head movements can be influenced were summarized. These were the size, weight and perceived cost of the HMD, as well as by its resolution, curvature and field of view. Since the head and eye movements influence each other mutually, the reasons that alter the users' head movements will also usually affect their eye movements. Discomforts like eyestrain, headaches or nausea (from simulator sickness) could occur from the method of mounting the goggles to the head, from the bi-ocular presentation or from the video see-through technique, since these affect resolution, curvature, brightness and contrast of the HMD. Additionally, a perceivable latency might cause discomfort. Finally, there are several aspects that might give rise to interactional problems with one or more interaction partners: the users' hand-eye coordination, pointing, the task performance and the imperfection of the illusion of the coexistence of the virtual and real world. These interaction aspects are influenced by the resolution, field of view, brightness and contrast as well as by software issues like registration or tracking errors and latency. Finally, the lack of eye contact might also lead to interactional problems.

Subsequently, two studies have been presented that investigate the influence of *our* HMDs on the behaviour of their users. The first study (Section 5.2) used eye trackers to investigate visual search under three different restrictive conditions: HMD, field-of-view-restricted and

unrestricted. Confirming our hypotheses, we found more head movements and less eye movements during field-of-view-restricted visual search than during unrestricted visual search. This discovery that the participants using AR show reduced eye movements and simultaneously increased head movements will be used for an approach to track the visual focus of attention in Section 7.1.

Unlike Dolezal (1982), we found no significant effects for the restricted field of view alone. We discussed that possible explanations might be the differing scenarios for the observation or that this might indicate an even more pronounced effect of curvature and distortion of the HMDs compared to the effect of the field of view. This would be good news for future generations of video see-through HMDs as the display quality might be easier to improve with existing technology than the field of view without simultaneously increasing size and weight.

The second study (Section 5.3) investigated how exactly the users' head movements are affected by the HMD and how speech and task completion are influenced. We found a significant reduction in the use of head gestures when participants used a head-mounted display. Generally, participants using the HMD reduced the duration and the speed of the respective gestures while repeating sinusoidal gestures less often. The general decrease of head movements is particularly noticeable in the reduced head movements for nod and shake (periods or velocity). Together, the results from the head gesture analysis suggest a reduced willingness of the HMD users to perform head movements if not necessary. Otherwise we could not confirm our hypothesis about the participants needing more head movements to complete the same tasks. We discussed that this might be due to compensation strategies. This would indicate that the effects on the head movement is less pronounced than expected since the participants seem to find effective compensation strategies for some of the drawbacks. Additionally, we found a higher number of utterances, which suggests that some disadvantages of the HMD are compensated by increased use of verbal language (e. g. confirmations, back-channelling or requests). The subjects using AR needed significantly more time to complete their tasks than those not using AR. However, the expected reduced performance for the AR participants was not supported by the data. Surprisingly, the results investigating the error rate indicated for one measure even *lower* error rates in the HMD condition compared to the non-HMD condition. This is very interesting – even if this result is caused by the (not significantly) longer duration of the task. Using conversation analysis, we could report changes in how the participants establish joint attention under AR.

In the questionnaire, the AR goggles were rated uncomfortable to wear (pressure on the nose/head), and the users complained particularly about the resolution and the small field of view. Additionally, the HMDs were said to interfere with the participants' actions, and affected the perceived naturalness of head movements. The participants seemed to be aware of the AR goggles during the entire trial (unlike the other sensors and cameras).

Combining the results from the measurements with the results from the questionnaire and user comments from both studies, the AR goggles seem to affect the interaction in terms of eye movements, head movements, speech, and task accomplishment. While on the one hand, head gestures and eye movements decrease, on the other, the number of utterances increases, along with the task completion time. Although from the studies there is no evidence as to which of the reviewed AR issues causes these altered behaviour when using our HMDs, we can conclude that some issues seem particularly noticeable: the wearing comfort, low resolution, small field of view and lack of eye contact.

How can we reduce these influences on the interaction? Section 5.1 argued that for some of the issues with HMDs there might not be a *complete* solution in the next few years (e. g. latency, resolution and field of view). Other issues are likely to be *improved* in connection with technical developments in the future. They particularly have to be considered in the design or selection of HMD hardware (e. g. range of the field of view as well as resolution and curvature of the HMD) as already proposed by several researchers (e. g. Arthur (2000); Patterson et al. (2006); Papagiannakis et al. (2008)). Furthermore, there is an issue that seems to affect the behaviour of the users (at least of our HMDs) that is to our knowledge only touched upon in the literature: the wearing comfort of the devices. Here is room for simple yet potentially effective improvements such as choosing another mounting method or padding the area where the goggles rest on the root of the nose. The effectiveness of such improvements should then be evaluated with a questionnaire, for example that of Knight and Baber (2005) who developed rating scales for the comfort of wearable devices (including HMDs) by brainstorming 92 wearable comfort terms and letting subjects cluster these into groups. Here, however, the scales were created only out of the clusters while terms that did not fit into clusters were ignored. If such scales are used, it should be thus considered (a) whether the wearable comfort terms can be translated to other languages, (b) whether the terms that could not be clustered are really negligible and (c) most importantly, whether the users should not only rate the comfort measures, but also if this value is still tolerable for the user and how important they judge this measure to be for this (AR) application. Finally, another issue occurred during this chapter that should be monitored in following studies: a possible influence on user behaviour by the perceived cost and perceived fragility of the used devices. An easy method to get first insights into this is to ask the participants of following studies in questionnaires subsequent to the studies how fragile and how costly they rate the wearable devices that they used during the study and if they think this to have affected the way in which they performed the task given.

While the transferability of findings acquired from research using ARbInI to normal collaboration without HMDs must be viewed as critical under the light of the insights presented in this chapter, the following two chapters will present advantages of ARbInI with and without using HMDs.

Most of the disturbances identified in this chapter seem to be induced by the AR goggles. In contrast to this, we found that most of the other sensors (and the high number of video cameras) seem not to hinder the participants in their actions. Thus, since the AR feature is optional in ARbInI, Chapter 6 presents features of ARbInI that can be used without the AR feature. For example, the AR goggles can be substituted with less obtrusive head cameras (e. g. spy cameras) without using the AR paradigm. The following chapter presents a method to use such video to infer to the objects in the field of view.

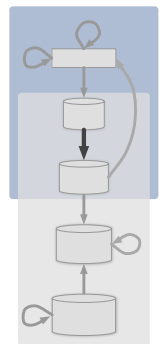
Beyond the features using video, ARbInI implements a large variety of other valuable methods, for example automatic annotation methods that allow for the analysis of parts of the corpus even during its creation. These save the researcher much laborious work. Thus, results found in research using ARbInI can be used as (easily achieved) hypotheses for non-mediated interaction. Additionally, automatic transformation methods that are implemented in ARbInI (conversion & analysis) are presented in Chapter 6. With these, the system can be used for comfortable exploratory, qualitative and quantitative research. Finally, the system offers a wide range of methods to systematically investigate AR-mediated interaction.

The automatic classification methods that are developed in Chapter 6 will be used in Chapter 7

to actively modify the interaction using the experimentation system. For this, approaches to use the features provided by multimodal AR will be presented. Firstly, Section 7.1 will present an approach to enhance the interaction and to compensate for the missing eye contact that is criticized by several users: a display of the focus of attention. For this, we can benefit from the result of the study in Section 5.2, where we found that the participants using AR reduce their eye movements and increase their head movements likewise when performing visual search. From the position of an object in the field of view we can, thus, infer on the user's visual focus of attention. This attention focus display is evaluated in a study and the results are discussed. Secondly, Section 7.2 will present an approach to actively disturb the interaction by providing false information, and discuss the possibilities of this method for analysing misunderstandings and repairs during interaction.

6. Computer-aided investigation of interaction

The previous chapters proposed the **Augmented-Reality-enabled Interception Interface (ARbInI)** as a tool to facilitate the investigation of interaction. They presented the hardware and software components as well as the features of the ARbInI system. The previous chapter analysed side-effects of head-mounted AR on the behaviour of the users and found that the users' eye and head movements as well as head gestures, speech and performance can be affected by the use of head-mounted AR. This chapter now focusses on techniques implemented in ARbInI that aid the researcher in analysing interaction. Thereby, it particularly discusses the features of ARbInI that do not necessarily rely on the use of HMDs. The approaches used in this thesis to automatically tag human behaviour are presented: (a) a method to track objects in the user's field of view, (b) the integration of an automatic speech recognition and (c) an approach to an automatic recognition of head gestures. Moreover, (d) the speech recognition will be combined with the head gesture recognition thus analysing the mutual timing of both signals. Each of these is combined with the conversion and analysis methods detailed in Sections 3.6 and 3.7 thereby analysing exemplary data and demonstrating how these methods can be used to gain knowledge about nonverbal behaviour and about the participants' behaviour in the respective scenarios. Finally, an outlook will be given that shows further possibilities of the presented techniques.



6.1. Automatic tracking of objects in the field of view

In Section 1.6, the importance of the visual focus of attention for interaction was introduced. Thus, to gain a detailed analysis of interaction (a process this chapter intends to facilitate) the consideration of the visual focus of attention is of special importance. Usually, researchers use eye-tracking to analyse these cues deeply. However, a permanent integration of an eye-tracking system into ARbInI is not intended since this would be costly. Moreover, the users rate the combination of an eye-tracker with our HMDs uncomfortable. However, the *exact* tracking of the focus of attention is not necessary for every research question. Already an approximate monitoring may be sufficient to facilitate the analysis of many questions (e. g. automatically annotating if the participant is looking at objects or at the interaction partner or estimating which part of the table is currently in the focus). Thus, we can use an easier approach by using the head-mounted camera that records the interaction from the participant's point of view to estimate the focus of attention. This idea is not new: for example Yoshida and Smith (2008) estimated the focus of attention of toddlers from a camera fixed to the heads of infants and Stiefelhagen (2002) estimated the focus of attention of participants in meetings from their head pose. Both have shown that it is possible to infer the focus of attention from the head pose.

When using head-mounted displays instead of head-mounted cameras, the estimation of the focus of attention is even easier since the field of view through the goggles is smaller than without. A preceding project to this work¹ found that their participants often did not notice hints that were displayed in the peripheral view of the display and reported the impression that their users focused mainly on the middle of their screen (Hanheide, 2006).

Together, the hypothesis arises that it is possible to (roughly) determine the focus of attention from the middle of the field of view, which may already be sufficient for several research questions. In order to verify this hypothesis, we conducted a case-study where we combined our HMDs with an eye-tracker. The results suggest that the participants in fact focus more on the middle of the screen although the peripheral areas are still used (but less often) (see Section 5.2 for details). This means, by using ARbInI's integrated camera and an estimation of the region of interest (the middle of the field of view), we can rely on a less accurate tracking of attention at the benefit of an unchanged load for the participants (assuming we block the peripheral perception by blinders). Furthermore, ARbInI also integrates a marker tracking (see Section 2.2) that shows virtual objects on top of the real-world anchors. Using this tracking, we can automatically determine which markers are at which positions in the field of view. By integrating an automatic analysis of the scene and a mapping of the markers to the respective objects that are in the field of view, we gain an easy estimation of the objects in the visual focus of attention.

All tracked positions of all objects that are in the user's field of view can easily be logged to the memory, which allows for integrating them into the multimodal data corpus. Instead of transferring these data into textual annotations (as for example the head gesture classifications, see Section 3.6), the data is automatically converted into a video. This video can then be shown in parallel to the video recordings from the head camera. It shows the object-ids at their positions in the field of view on a black background. Such visualization can be very helpful in the analysis since the researcher sometimes might have difficulties to identify the objects in the video stream from the head camera (e. g. because of the current view angle of the camera) and can then refer to the object-id from the automatically generated video. Figure 6.1 shows a screenshot from such a video.

Furthermore, this automatic tracking of the positions of objects in the field of view can be provided to the interaction partner. Since the AR goggles block the view on the interaction partner's eyes, it is usually difficult to follow his or her eye gaze direction. The object position data can be used to compensate for this: an audiovisual display of the partner's visual attention focus was introduced in Section 3 and will be evaluated in Section 7.1.

6.1.1. Outlook

Apart from the discussed implemented features (providing the data as a video and/or as a display of the partner's attention focus), there are also possibilities for analysis of such logged data that have not been implemented yet. For example, the data could be used to calculate an overall viewing time per object in order to find out if some objects are particularly interesting or challenging in a task.

By attaching an additional marker to the interaction partner's head or AR goggles, we could

¹The Vampire project "Visual Active Memory Processes and Interactive Retrieval" (IST-2001-34401)



Figure 6.1.: Tracked object IDs in the field of view.

easily log attempts of the participants to look at the other's face/eyes. An interesting question to answer is whether such attempts decrease during the study because the participants learn that most of the face is masked by the goggles and is, thus, not useable as an interaction cue. When using head-mounted cameras instead, we can easily keep track of mutual gaze. For example, this can be used to investigate the timing of mutual gaze in correlation with head gestures. In conclusion, the tracking of the objects in the field of view can be used in ARbInI at nearly no cost but offers rich possibilities to help the researcher with the analysis of the interaction.

6.2. Automatic annotation of speech times

As an additional cue, ARbInI allows with its integrated head-mounted microphones for an automatic analysis of the speech signals. The data from the microphone is directly processed by a speech recognition software by Fink (1999). According to Rudnicky et al. (1994) and Basapur et al. (2007), the speech recognition of natural and spontaneous speech nowadays is still far from being comparable with the human speech understanding. More specifically, Munteanu et al. (2006) report that although state-of-the-art speech recognition systems can achieve less than 3% error rate in perfect conditions, the error rate increases when the acoustic conditions are not optimal (e. g. 2 persons turn-taking) and can even reach 40-45%. Since we cannot guarantee optimal acoustic conditions in our studies (e. g. noise of computers, noise from outside), we do not try to reconstruct the speech from the phoneme hypotheses but instead use the speech time annotation: which participant is speaking when? This annotation can be used for an analysis of the utterance duration, the overall number of utterances and the speech percentage.

6.2.1. Analysis of the speech data of this work

The speech data was analysed for speech times in the interactive exhibition design study (see Section 5.3 for an introduction to the study and further results). Table 6.1 shows the average

values for utterance duration, number of utterances, the sum of utterances and the average pause duration for the different phases/tasks of this study. Since speech was not allowed during the individual phase, this phase is omitted from the evaluation. Each of the study's three tasks was preceded by an introduction phase, where the experimenter explained the task to the participants and offered to ask questions. The three introduction phases were concatenated for the evaluation and called 'intros'. Comparing the three different phases, the table shows a noticeable difference of the intro phases to the smalltalk and dyadic (pairwise) phases: there seem to be less and shorter utterances during this phase which is also supported by the overall speech percentage which is smaller than 10% while it is about 30% for the other phases. Between the smalltalk phase and the dyadic phase, the mean utterance duration does not differ much while the mean number of utterances is higher for the dyadic phase. However, the mean speech percentage remains nearly constant.

	Exhibition design study		
	smalltalk	dyadic phase	intros
mean number of utterances	76.25	93.55	27.17
mean utterance duration	1.67	1.62	0.96
mean speech percentage	32.97	32.09	7.02

Table 6.1.: Utterance analysis per task of the exhibition planning scenario. The left column gives the results from the smalltalk phase, the middle column from the dyadic phase and the right column summed about all intros.

6.2.2. Discussion

The results regarding the comparison of the smalltalk and the dyadic phase indicate that there is no difference in the proportion of utterances between these two tasks: Although the mean number of utterances per participant indicates more utterances, the mean speech percentage shows no difference. This means that the higher number of utterances is simply caused by the longer duration of the dyadic task (see last row of Table 6.9). Concerning the introduction phases, the results show a reduced amount of speech of the participants as well as shorter utterance duration. This is likely to be caused by the nature of these phases: the experimenter explains the next task and gives hints on how to reach the goal and what to keep in mind. During this, the participants both mostly take the listener's part. The recorded utterances thus are likely to be listener feedback like "I see" or short questions like "what is this box?". In contrast to this, the smalltalk phase and the dyadic phase *encourage* a conversation or discussion and the participants share the parts of the listener and the speaker. Examples for such utterances are "Have you ever been to an interactive museum?" or "There is an object called wind machine here. I think this might disturb the exhibits with candles". In such a dyadic interaction, there are naturally more and longer utterances compared to two participants listening to a third person. This explains the longer duration of the utterances, the increased number of utterances and the overall speech proportion compared to the introduction phases.

6.2.3. Outlook

While the results show how speech recognition in combination with the conversion & analysis methods can be used to gain knowledge about the duration and numbers of utterances in specific interaction situations there are many more prospects for this technique. For example, the speech times could be analysed for their timing thereby offering possibilities to investigate the overlaps between speaking turns. Furthermore, the recordings of the speech can then be used for an automatic pitch analysis that can be combined with the analysis of the turn-taking. Finally, these two cues could then be combined with the analysis of the focus of attention (especially mutual gaze) as it is described in the previous section and with the head movement analysis (described in the subsequent section). In conclusion, a variety of analyses is possible with such a corpus.

6.3. Automatic tagging of head gestures

Section 1.6 introduced the importance of head gestures for interaction and especially for successful turn-taking. For an investigation of head gestures, typically video data is annotated manually. This laborious process can be dramatically facilitated by a combination of machine-learning methods with a tracking technique. Possible techniques are detailed by Vatavu et al. (2005) who claim that vision-based approaches (e.g. Morency et al. (2002)) have the advantage that the measurements can be done unobtrusively but are dependent on constant lighting conditions and on a full view at the interlocutor's face. Moreover, high processing power is needed. However, in our multimodal Augmented Reality collaboration analysis scenario, a full view on the face is not available since some of the participants are wearing HMDs. Thus, we use a sensor-based approach.

We use motion sensors mounted on the participant's head, which grant lighting independence and almost unrestricted mobility (in a table-centred scenario). Normally, this approach has the disadvantage of being more obtrusive, but since the participants already wear heavier HMDs, the lightweight sensor should not cause them further hindrance. Additionally, an important practical advantage in this context is the easy applicability of our system without any calibration.

To examine the possibilities of such a sensor-based head gesture recognition system, we accomplished a comprehensive study with 10 participants in comparatively natural setups. Thereby, we abdicated any sensor preparations and adjustments. See Section 6.3.5 for details about the data acquisition.

6.3.1. Training and classification

The training data were acquired in the animal guessing scenario (see Section 4.3.6) and annotated manually (see Section 6.3.5 for details on the annotation and an introduction to the resulting data corpus). All data were recorded at 100Hz and reduced to 33Hz and normalized to zero-mean and unit variance. For the recognition only the data from the gyroscopes with 3 DOF rate-of-turn were used (see Section 2.1 for details on the hardware). For the recognition we used a the so-called ordered means models (OMMs), an approach to

machine learning of time-series and sequences. OMMs are inspired by and similar to hidden Markov models (HMMs) (Rabiner, 1989). Großekathöfer and Lingner (2005) could show that OMMs provide a high level of robustness in terms of fragmented or insufficient data while needing less computational power than HMMs and that they, nevertheless, achieve similar generalization results as HMMs. The following sections analyse the recorded data offline and online.

6.3.2. Offline evaluation

To estimate the accuracy of our approach for head gesture recognition, we performed two different evaluations with regard to separability/robustness and transferability (see also Wöhler et al. (2010); Wöhler (2009)):

Evaluation 1 (*Robustness*)

The first evaluation tested whether the classifier is suitable and robust for this kind of data. For this, we randomly partitioned the available data from all 10 participants into equally sized training and test sets.

Evaluation 2 (*Transferability*)

In a second evaluation, we investigated the classifiers' transferability to new participants. Thus, we used data captured from 9 participants as training data while the data from the remaining participant was used as test data (test participant). We accomplished this evaluation for each participant.

Thereby, to analyse the mutual influence of head movement classes on the performance, we repeated both evaluations four times, each time with a different set of head gesture classes. Since *nod* and *shake* are the most frequently occurring head gestures during question-answer situations, every set included these two gestures: (a) *nod*, *shake* (b) *nod*, *shake*, *tilt* (c) *nod*, *shake*, *look* (d) *nod*, *shake*, *look*, *tilt*. Please find the exact parameters used for the evaluations in Wöhler et al. (2010).

Results

The results of Evaluation 1 reveal classification success rates between 86.36% and 97.48% (see Table 6.2, middle column). The best rate was achieved when *nod* and *shake* were used as trained head gestures, whereas the lowest rate occurred when all four classes were used. For the case of three trained classes, there are two different results depending on the third added class: the addition of *tilt* results in a classification rate of 95.32%, while the addition of *look* results in a classification rate of 87.90%. Please note that random classification leads to 25% accuracy with 4 classes, whereas 2 classes reach 50% by chance.

The classification rates from the second evaluation, which are weighted averages over all participants, range from 75.95% to 98.40% (Table 6.2 last column). Here again, the best rate was achieved with two classes (*nod* and *shake*), and the lowest rate occurred when all four gesture classes were used. In order to examine the mutual influence of all four gesture

Trained head gestures	Classification rates	
	Evaluation 1	Evaluation 2
(a) <i>nod, shake</i>	97.48%	98.40%
(b) <i>nod, shake, tilt</i>	95.32%	94.82%
(c) <i>nod, shake, look</i>	87.90%	79.49%
(d) <i>nod, shake, look, tilt</i>	86.36%	75.95%

Table 6.2.: Classification rates from Evaluation 1 and 2.

real gesture	classified gesture				performance
	<i>nod</i>	<i>shake</i>	<i>tilt</i>	<i>look</i>	
<i>nod</i>	405	6	20	16	90.60%
<i>shake</i>	10	260	7	28	85.25%
<i>tilt</i>	2	4	36	2	81.82%
<i>look</i>	12	132	27	139	44.84%

Table 6.3.: Confusion matrix from Evaluation 2 with all 4 gesture classes accumulated over all 10 participants. Samples on the diagonal are classified correctly.

classes, we generated an overall confusion matrix (see Table 6.3). This matrix is the sum of the confusion matrices of all 10 runs. Samples on the diagonal are classified correctly. While most samples for *nod*, *shake* and *tilt* are classified correctly, the number of correct classified samples per class is considerably lower for the *look* class. More precisely, 132 of 310 *look* samples have been mistakenly accounted to the *shake* class.

Figure 6.2 shows the classification rates for the head gesture class sets per participant. The rate ranged from 61.76% to 100%, where the class set (a) with *nod* and *shake* achieved the best performance results again. Similarly to the first evaluation, the system reached very high accuracy with class set (b) for almost all participants. The classification rates decrease for both class sets that include the *look* class. The figure also seems to show Participant 3 provides the most difficult data for classification.

Discussion The first evaluation tested whether the classifier is suitable and robust for this kind of problem. We found that all four head gesture classes are easily separable although the classification rate slightly decreases for the four classes case (Table 6.2, middle column).

With the second evaluation, we investigated the transferability to new participants. We observed that all four classes are still easily separable. As a further result, we found good transferability to new participants for class sets (a) and (b). For the class sets that include the *look* class (c and d), we cannot conclude stable transferability. However, with more than 75% performance, these classifiers still provide good hypotheses (Table 6.2, right column).

Overall, especially the *look* class seems to have a negative effect on the classification rate. This finding is further supported by Figure 6.2 and especially by the confusion matrix in Table 6.3. About half of the *look* gestures were classified as *shake* gestures. A likely reason for this might be the similarity of both movements. We assume the classifiers to assign

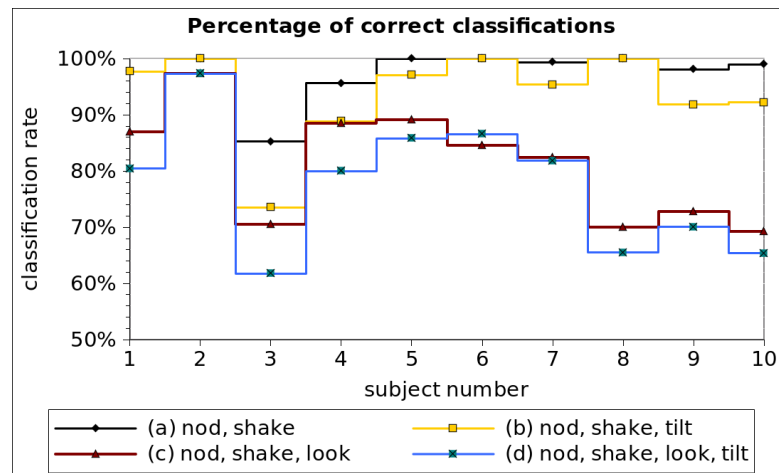


Figure 6.2.: Head gesture classification rates for all ten participants from Evaluation 2. Note that the connection of the dots does not indicate interim values.

look movements as fragmented *shake* gestures. This means that the classification has to distinguish singular lateral movements from repeated lateral movements. A reason why this is difficult is that the annotated *look* movements are not in every case preceded and succeeded by a non-movement interval. A further influence might be that our dataset is biased in the number of examples per class (much more *nod* and *look* samples than *shake* and *look* samples). This is the result of our comparatively natural acquisition scenario. Although we could have asked participants to perform the four gestures repeatedly we assume such resulting gestures to be much more artificial. We claim that our data acquisition method is superior since the nativeness of the recorded gestures should be an advantage for online recognition scenarios.

6.3.3. Automatic/Online classification

To expand the proposed system to online classification some extensions have to be applied. First of all, to process a continuous head motion data stream from the sensor we partition the data via a sliding window approach into fragments. Additionally, we establish a rejection scheme in case no head gesture is performed: based on the posteriori probabilities we define thresholds by which, if under-run, classification is rejected. Please find details on this in (Wöhler, 2009) and (Wöhler et al., 2010). In preliminary tests we achieved promising results². Thus, we used the classification in a subsequent study with the interactive-exhibition planning scenario (see Section 4.3.4 for a description of the scenario and Section 5.3 for the hypotheses and other results of the study). The following section develops a method to evaluate the classified head gestures of this study and discusses the results.

²cf. <http://www.techfak.uni-bielefeld.de/ags/ami/research/hgr/index2.shtml>

Comparison of classification and annotation (or two annotators)

In order to evaluate the classification, the gestures annotated manually have to be compared with the gestures tagged by the classification. Although it would be possible to simply compare the false positives and the false negatives with the correctly classified tags, we believe that this is not the best evaluation for this problem. As a reminder, the purpose of the classification is to provide a reliable head gesture annotation. For this, all interaction data (including the classification data) is transferred into an ELAN corpus (see Section 3.6) in order to allow for a multimodal analysis of the data. Since the classification results are not expected to be perfect, the plan is to correct the classification by hand. This means, that an appropriate evaluation of the classification should actually describe the *amount of work* that would have to be done by a possible human corrector instead of simply reporting false-positives and false-negatives.

Figure 6.3 visualizes the comparisons to be performed: The two blue tiers³ (more precisely their annotations) are to be compared. Thereby, the upper blue tier represents the annotations (that are the gestures that have been tagged by a human annotator which we treat as the reference) while the lower blue tier represents the classification (these are the gestures tagged by the automatic recognition that are to be evaluated). The grey lines illustrate the segments that are overlapping (grey line in the middle) and non-overlapping or extending (the grey line at the top for the annotation tier and the grey line at the bottom for the classification tier).

There is a great difference in the amount of work that a potential annotator would have to do in correcting the different occurring error cases: The greatest amount of work falls on *finding the segments* in the data where gestures take place. For this, an annotator has to watch the whole video, observe a gesture, stop the video, highlight the proper segment and tag it with the correct annotation value and start the video again. In comparison to this, other possible corrections are much less work. To *correct existing tags*, the amount of work is much less. Here the annotator can jump from classification to classification and only has to play the video during this short intervals which saves a lot of time. There are three different cases what can be wrong with such an existing tag:

- The segment is false (the classification tagged a gesture where there is actually no gesture (annotated)). This is marked as 'false-positive' in Figure 6.3. This segment has to be deleted which can be done in ELAN using the context menu clicking 'delete'.
- The tag value is wrong ('nod' instead of 'shake'). This is marked as 'non-match' in the Figure. This tag description has to be corrected which can be done using the context menu clicking 'rename annotation' and entering the new value.
- The tag does not fully overlap (the classification segment begins before/after the gesture annotation starts and/or ends before/after the gesture starts). This is marked as 'extend' in the Figure. Here, the annotator has to find the correct position or length of the segment by playing a short interval preceding and succeeding the annotation before the segment can be corrected.

As the descriptions of the amount of work for the corrector shows, the correction is easier for

³Reminder: the word 'tier' describes a set of annotations that share the same characteristics. In this case, the annotations are grouped by their origin: annotation versus classification (see Section 1.6.4 for more information).

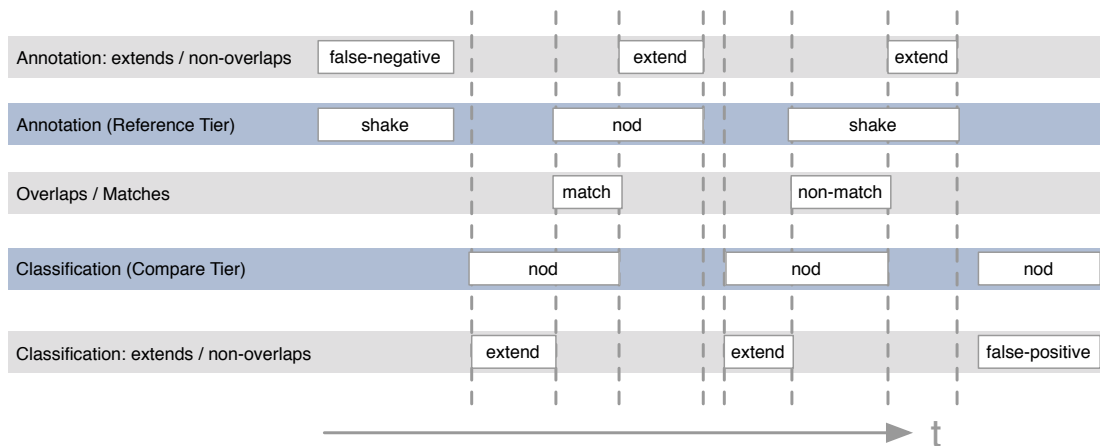


Figure 6.3.: Schema: possible overlap cases for the tags. The vertical blue lines comprising time-intervals of annotations are called tiers and represent the annotations tagged by hand (upper blue tier) and the tags set by the classifier (lower blue tier). When comparing the annotations of such tiers on a timeline t , the overlapping intervals (*overlap*, middle grey line) are called match or non-match respectively according to the matching of their values. The overlaps are often followed and/or preceded by extensions. Annotations in the reference tier that do not overlap with a corresponding annotation in the compare tier are called false-negative while annotations in the compare tier that do not overlap with a corresponding annotation in the reference tier are called false-positives.

deleting and value correcting than for the time correction where the corrector has to find out first which is the correct position or duration of the segment.

To calculate the amount of work for the corrector we take three steps. In the first two steps we ignore the classified gesture class (match/non-match) concentrating only on the correctness of the *segmentation*. First, we distinguish between annotations that would remain to add or to delete by the corrector.

$$\begin{aligned} \text{to add} &= \frac{\# \text{false-negatives}}{\# \text{annotations}} \\ \text{to delete} &= \frac{\# \text{false-positives}}{\# \text{classifications}} \end{aligned}$$

Additionally, we calculate a special measure that describes the cases in which an annotation overlaps with a real gesture but has to be shifted in time or needs a length correction. Here, we group the amount of 'extend' (that is the length correction) into three arbitrary classes:

$$\begin{aligned} \text{small length correction} &= \frac{\# \text{extends}[0.25 - 0.5s]}{\# \text{overlaps}} \\ \text{medium length correction} &= \frac{\# \text{extends}[0.5 - 1s]}{\# \text{overlaps}} \\ \text{large length correction} &= \frac{\# \text{extends} > 1s}{\# \text{overlaps}} \end{aligned}$$

	annotation vs. classification		annotation vs. annotation	
	with HMD	without HMD	with HMD	without HMD
% to correct tag	38	29	2	0
% to add	61	29	30	14
% to delete	75	76	0	12
needs length correction				
% small	13	17	15	12
% medium	33	35	17	7
% large	34	45	12	5

Table 6.4.: Reliability of classification and annotation. Average values over all annotated participants for classification vs. annotation. For the annotation versus annotation columns each column refers to one participant.

In a third step, we calculate the *correctness of the annotation values* of all those annotations that overlap ('nod' vs. 'nod' is correct ('match') whereas 'nod' vs. 'shake' is wrong ('non-match')). Here, the corrector only has to reset the annotation tag, which is the least amount of work of all the discussed corrections.

$$\text{to correct tag} = \frac{\#non-matches}{\#overlaps}$$

where *#overlaps* is the number of annotations that overlap (while ignoring the annotation values).

Results

The method described above can be used to compare the classification of head gestures with the annotations. The aim is to evaluate the amount of work a possible corrector would have to invest to gain a correct annotation of the head gestures. Moreover, we can also use the same method to compare the annotations of two different annotators. This latter comparison is particularly interesting, to review the significance of the former comparison: if two annotators disagree to a great extent about the annotations, the comparison of the classification with the annotation cannot reach satisfactory results. The results of both comparisons will be described in this section.

Classification versus annotation The left half of Table 6.4 shows the results for the comparison of the classification with the annotation both for the HMD and the non-HMD condition. The annotation value has to be corrected for 29% (HMD) and 38% (non-HMD) respectively of the overlapping annotations and from the classifications, the correction annotator has to delete 75% and 76% respectively. More importantly, the classification missed, in average, 61% and 29% respectively of the gestures. Concerning the correctness of the lengths and positions of the segments, the annotator has to correct between 7% and 33% of the gestures in their length (by either extension or shrinkage).

Annotation versus annotation The right half of Table 6.4 shows the respective results for the comparison annotator A and B. Please note that since only two participants were annotated by both annotators (see Table 5.2 in Section 5.3.2), the data is calculated from one participant with and one without HMD. In this case, the annotation value differed for 2% versus 0% of the overlapping annotations while from annotator B tagged 0% or 12% gestures that annotator A did not annotate and, conversely, annotator A annotated 30% or 14% gestures that annotator B did not annotate. Concerning the correctness of the length and positions of the segments, the two annotators do not agree about the exact beginning and end of the gesture for 4-12% of the gestures.

Discussion

While for the *classification versus annotation* the corrections and deletions are a significant amount of work, they still seem to be manageable. But the classification missed in average 29-61% of the gestures that were annotated by the annotator. This means that correction annotators still would have to go through the whole corpus file if they wanted to make sure that *all* relevant gestures are tagged. If the aim is that all gestures are to be annotated, the classification tier cannot be used as a good pre-annotation at this stage of the head gesture classification. If, on the other hand, the aim is to annotate *some* of the gestures (e. g. in order to provide them to the co-participant) it might be possible to adjust the parameters to reduce the amount of false-positives.

Although the amount of missed gestures is certainly lower when comparing *annotation versus annotation*, there are still up to 30% gestures, that the other annotator did not rate a gesture or missed. Additionally, the annotators disagree about the annotated value of one of the gestures in the HMD condition which leads to the 2% value for 'to correct tag': this is a gesture that seems to start with a nod and is superimposed by a tilt movement. One of the annotators annotated both gestures while the other one tagged only the nod movement but over the whole duration of both movements. Nevertheless, the question arises why the inter-coder-agreement seems to be so low between the two human annotators. One possible reason might be that we asked untrained coders to rate the head gestures. Additionally, the task instructions for the annotators (see Section 6.3.5) do neither ensure that every gesture is annotated (since the annotator might rate some as unsure and as a precaution leave it un-annotated while another annotator would not rate the same movement as unsure and annotate it) nor do these instructions create exact segments for the gestures. Firstly, because of the inclusion of preceding and succeeding data which might have a different length for different annotators and might also be truncated by other movements. Secondly, since gestures often fade out annotators might choose different finish points for the gestures.

In the results, we also found a pronounced difference between AR and non-AR. For the comparison of both annotators, these values are based on one participant per condition. To find out if this difference occurs by chance or if there is an effect, we would need more data that is annotated by more than one annotator. For the classification versus the annotation, however, this difference is particularly striking for the percentage of annotations to add. The higher number in the HMD condition could be explained by our finding that head movements are affected by the HMD (e. g. less amplitude, slower movement, less repetitions as we described in Section 5.3.3) and, thus, the classification of those movements is affected in

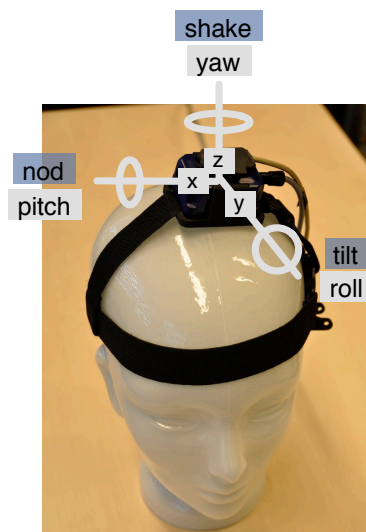


Figure 6.4.: Position of the MT9 sensor on the head and resulting gyroscope axes.

such way that fewer gestures are segmented correctly. Moreover, because of a lack of training data from an HMD setup prior to the study, the classification has been trained exclusively with gestures from a non-HMD scenario. Thus, the classification should be retrained and repeated with the new annotated gestures. For the annotations to delete, this difference is highly influenced by the significantly higher number of classified gestures for the non-HMD condition than for the HMD condition (see Table 5.8). Since there are more classifications in the non-HMD condition a correcting annotator would have to delete more classifications.

6.3.4. Analysis of relevant axes for detailed analysis of gestures

During the analysis of head gestures in past studies we noticed that for our four gesture classes *nod*, *shake*, *tilt* and *look* the movement happens mainly in *one* of the channels of the gyroscope. For the *nod* movement, this principal channel is the pitch channel, for *tilt* it is the roll channel whereas the yaw channel shows the greatest variance during *shake* and *look* movements (Figure 6.4 shows how the sensor is attached to the head). In order to verify this subjective observation, we computed for each of the gesture classes a principal component analysis (PCA) and plotted the data for each gesture and for each experimental condition before and subsequent to the axis transformation.

Results

The results are summarized in Table 6.5. The Table shows that 70-84% of the variance in the data can be explained (in average) by the single chosen data channel while the axis transformation can improve this to 81-94% of the variance for the first axis. Additionally, Figure 6.5 shows boxplots for the variance explained by the chosen data channels respectively for the AR and the non-AR condition. The figures show for all gestures annotated by hand the percentage of variance of all three data channels that can be explained by the most important data channel.

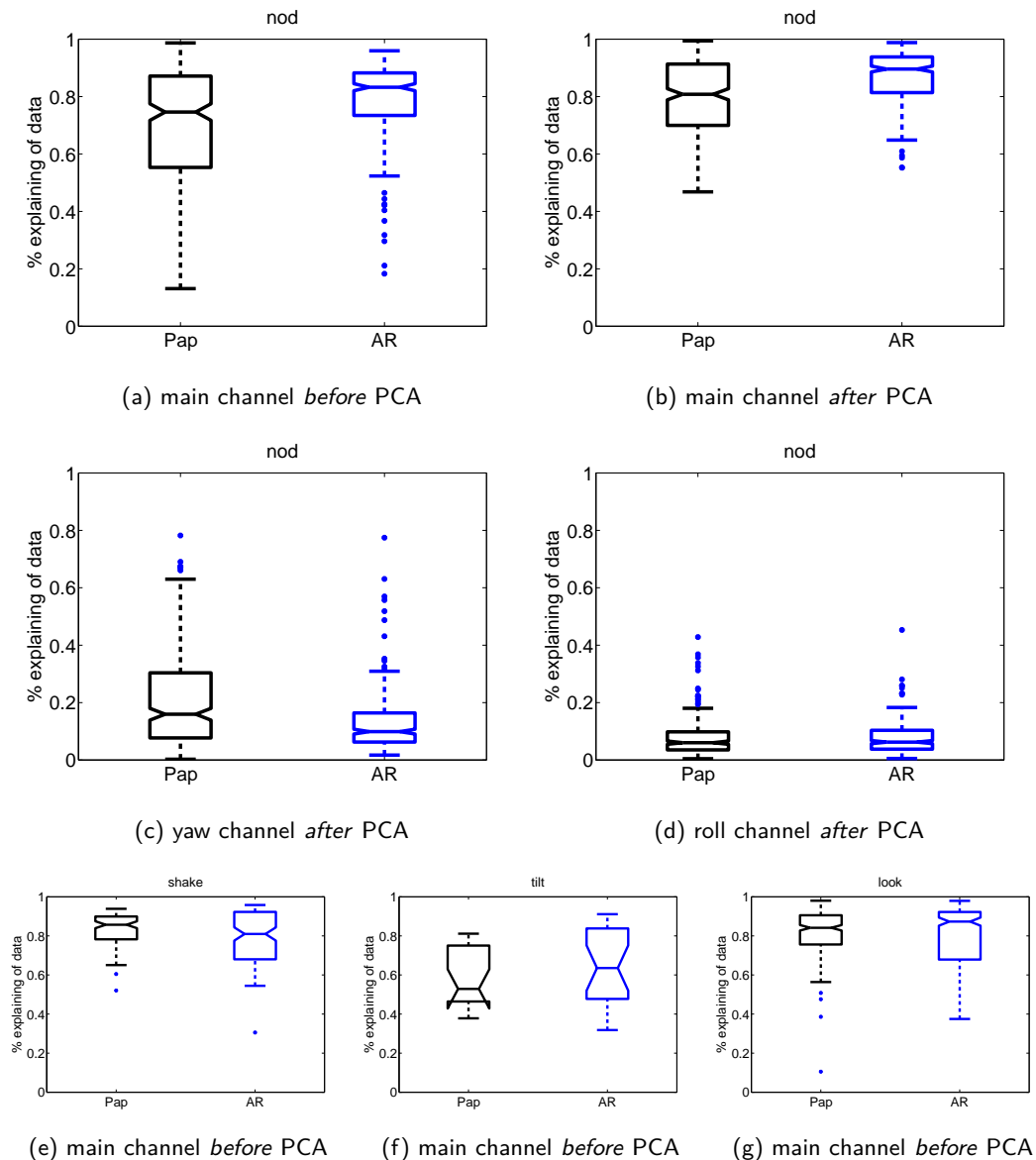


Figure 6.5.: Variance box plots for the four movement classes. Blue boxes represent the HMD condition (AR) while black boxes depict the non-HMD condition (Pap). Figure (a) to (d) show detailed plots for the *nod* movement. Figure (a) displays the percentage of the overall variance (summed over all three data channels) that is covered by the chosen data channel. Figure (b) shows the percentage of variance covered by the principal axis *after* a principal component analysis has transformed the axes. The percentage of variance covered by the (c) yaw and (d) roll data channel. (e)-(g) show the same plot as in (a) for the *shake*, *tilt* and *look* movements. General hints to box plots in this work: Each box represents the median (central mark), the 25th percentile (notch), the 75th percentile (edges of box) and the most extreme data points (whiskers). Outliers are plotted individually (+).

			before PCA		after PCA	
			mean	std	mean	std
	rotation around					
<i>nod</i>	pitch	x	76	6	83	5
<i>shake</i>	yaw	z	79	11	89	9
<i>tilt</i>	roll	y	70	16	81	12
<i>look</i>	yaw	z	84	3	94	2

Table 6.5.: Analysis of most relevant axes per head movement class: Percentage of variance that is explained by the respective axis.

Discussion Although the axis transformation can improve the percentage of variance that can be explained by the most important axis, the chosen data channels still explain most of the gesture taking place: for comparison, Figures 6.5 (e) and (g) show the box plots for the yaw and roll channel for the gesture *nod*. This means that the head movements take place mainly in the identified principal channel. Thus, for analyses investigating the characteristics of head gestures (see Section 5.3.3), we are able to reduce the data to the chosen column according to the gesture type in order to analyse the variability within this channel thereby focussing on the sinusoidal structure of the gestures.

6.3.5. Head gesture corpus

As part of the development of training and evaluation data for the head gesture classification, we annotated head gestures in dyadic or triadic conversations in three studies with four different scenarios and two different sensors. In sum, we collected 1910 gestures. The following sections will first explain how the data were annotated by hand and which gesture classes were annotated. An overview of all parts in the corpus will be given subsequently. In the studies, all participants were German native speakers and briefly informed about the purpose of the study and which kind of data was recorded.

Remarks: Annotation by hand

We synchronized the sensor data with the scene camera video and annotated the head movements in ELAN (Hellwig and Uytvanck, 2004). Annotators tagged the head movements of the data according to the three head gestures occurring most frequently: *nod*, *shake* and *tilt* (occurred as a gesture of uncertainty). Please note, that we include not only confirmative head gestures but also gestures that are used to structure the conversation (e.g. for turn-taking). Thus, we include all movements that *look* like nods, shakes, tilts regardless of their actual communicative meaning. The reason for this is that we intended to use the head gestures as training data for a classification and the classification would not distinguish between the communicative meaning of the gesture but simply learn a typical movement pattern.

Additionally, we asked the annotators also to tag *look*-movements. It describes a sideways movement of the head as it occurs in an object-choice task when looking from one to another

object on a table or when exploring the room. The reason to add this movement is that we noticed it to occur frequently in our studies. Although the *look* movement is similar to a head shake, it should not be classified as a shake gesture since the communicative meanings of both movements are different. Thus, it is important to discriminate explicitly between these two movements to avoid confusion for the analyst.

The annotators were asked to tag only such gestures where they would not hesitate about the gesture class to include the whole gesture in their segments. For this they were to select a subjective start and end point with a slight amount of time before and after the gesture if this were possible (sometimes this is not possible since a gesture is preceded or succeeded directly with another gesture or movement so that such surrounding segments would include activity in the sensor stemming from other movements).

Summarizing, the aim of these instructions was not to annotate *all* gestures occurring during the study but to achieve a dataset of relatively certain gestures to train the head-gesture classification with by means of these annotations.

Animal guessing task

The head movement data was recorded using the Xsens MT9 inertial sensor (see Section 2.1). With this sensor attached to the top of the participants' heads, we recorded the head movements of 10 participants during dyadic or triadic conversation (see Figure 4.8). The participants were asked to talk about whatever they liked or to play an animal guessing game (see Section 4.3.6 for a description of the task). Since we noticed that our participants tended to avoid head shakes (Dierker et al., 2009a) we asked some of the participant pairs to answer the questions in the animal guessing game without speaking in order to gain a more balanced training set. The conversation was video-taped by a scene camera and stopped when it became stagnant. One annotator annotated all data.

The head motion data was then sliced into the annotated segments. Table 6.6 gives an overview of the resulting samples per participant and the speaking/ non-speaking condition. Initially, these data were used for the training of the classification which was then used in the exhibition design study presented in Section 5.3.

Interactive exhibition design scenario

The classification was trained with the above data and used in a subsequent study, the exhibition design study. The study was split in three tasks: a smalltalk task (smalltalk), an individual exhibition design task (individual work with objects on a table without interaction partner) and a dyadic exhibition design task (dyadic interaction including objects on a table) (see Section 4 for details). The head movement data was again recorded using the Xsens MT9 inertial sensor (see section 2.1). With this sensor attached to the top of the participants' heads, we recorded the head movements of 26 participants during dyadic (or triadic with the experimenter) conversation.

In order to evaluate the reliability of the classification, it was again necessary to annotate the head gestures in this study by hand. Two annotators manually tagged the data of five participants of the non-HMD condition according to the same rules as above. The head motion data was then sliced into the annotated segments. Table 6.7 shows the resulting samples.

S	Number of Events for				length	duration	speech allowed
	Nod	Shake	Tilt	Look			
1	27	17	0	2	0.98s	11 m	yes
2	22	14	0	2	1.1s	11 m	yes
3	33	1	0	0	1.39s	15 m	yes
4	22	1	4	3	1.09s	15 m	yes
5	35	67	2	37	1.4s	33 m	no
6	27	38	6	26	1.2s	33 m	no
7	77	81	15	47	1.46s	27 m	no
8	15	16	1	49	1.04s	27 m	no
9	122	37	13	65	1.42s	43 m	no
10	67	33	3	79	1.17s	43 m	no
Σ	447	305	44	310			

Table 6.6.: Head gesture corpus: animal guessing scenario. Number of head movement samples per participant. S: participant number, length: medium length of event (in seconds), duration: overall duration of the measurement (in minutes). The participants 5–10 were asked to answer the yes-no questions of their partners without verbal utterances.

Artificial head gestures

Additionally, we asked 7 participants to provide "artificial" head movements. The participants were asked to wear a sensor setup and follow the instructions of a computer program that asked them to nod, shake or tilt their head repeatedly while recording their head movement data using the gyroscopes. The participants were able to abort the saving of a sample if they thought the sample to be erroneous (e. g. because of doing a false gesture). We used the Wii MotionPlus sensor to acquire the data. The data were sliced into the samples by the computer program. Table 6.8 shows the resulting samples.

Discussion on the different parts of the head gesture corpus

As described, we used several different scenarios to acquire the head gesture data corpus. The easiest scenario (with the least amount of work for the data acquisition) is the artificial one, where the wearers of the sensors were told to nod, shake or tilt their heads repeatedly as during conversations. The advantages for this method are that the data acquisition is very fast since the laborious annotation process is fully skipped because the acquisition software can cut the training samples already during the recording. However, this method also has several disadvantages: First, the recorded gesture is always preceded by a pause where the participant reacts to the request. Second, the gestures are often incomplete in the end since the length of the gesture is predetermined by the software and not by the user. This might lead to improper training data since natural gestures usually fade out as we discussed previously. Finally and most importantly, the gestures are likely to be unnatural since the

Participant	Number of Events for				length	duration
	Nod	Shake	Tilt	Look		
1	39	3	2	5	1.73	20
2	6	4	1	36	1.31	20
3	37	2	0	58	1.35	20
4	40	3	4	15	0.99	23
5	60	21	1	55	1.67	23
Σ	182	33	8	169		

Table 6.7.: Interactive exhibition design scenario: Number of head movement samples per participant. Length: medium length of event (in seconds), duration: overall duration of the measurement (in minutes).

Participant	Number of Events for			length
	Nod	Shake	Tilt	
1	20	20	20	3.46
2	20	20	20	3.49
3	20	20	20	3.46
4	20	20	20	3.47
5	20	15	20	3.48
6	19	20	18	3.43
7	20	20	20	3.52
Σ	139	135	138	

Table 6.8.: Artificial head gestures: Number of head movement samples per participant. Length: medium length of event (in seconds), duration: overall duration of the measurement (in minutes).

participants are fully aware of the reason for their gestures and that they are recorded. This might cause the participants to perform untypical strong/slight or short/long gestures. This might affect the training of the classification. Additionally, the artificial corpus might mislead the researcher or the classification to assume that all gestures are equally likely. That this is not the case in natural interaction is shown in Table 6.6 particularly in the cases where speech was allowed and Table 6.8. Speech seems to be accompanied in several cases by head nods while only very seldom being accompanied by head shakes or head tilts, a phenomenon which is (for example for the shake gesture) also described by Hadar et al. (1985) or Noller and Callan (1989). With these considerations in mind, this part of the head gesture corpus was not yet used as training data for the classifier and also will be excluded from the analysis of the corpus data in the following section. In the future this data acquisition scenario could be improved by asking the participants to nod in different durations or intensities.

In contrast to this, the head gesture samples from the animal guessing and exhibition design

scenario are recorded in a much more natural recording situation. Although the participants are told that their head movements will be recorded, they have to complete a (more or less challenging) task that is much more likely to distract them from the recording situation and their movements. We believe that this method is superior to the artificial scenario since the resulting head gestures should be such gestures as they occur during interaction. By this, we gain a head gesture corpus with typical head movements from interaction.

6.3.6. Analysis of the corpus data

The two parts of the head movement corpus described above that contain natural head gestures will be analysed in this section according to the head movement distances. Table 6.9 shows the measured head movement distances (absolute and average, see Section 5.3.2 for the calculation) for the two studies presented in this thesis that used continuous head movement recording. The studies used different scenarios: the animal guessing scenario (see Section 4.3.6) and the interactive exhibition planning scenario which consists of three phases (see Section 4.3.4). Comparing the distances, we find much higher distance values for the animal guessing scenario compared to all three phases of the exhibition design scenario. Note, that the overall duration of the recording also differs to a great extent. Comparing the average distances, we again find a noticeable difference between the animal guessing scenario and the three exhibition design scenario tasks. However, the distances of the three phases of the latter scenario do not differ to the same extent.

	animal guessing scenario	exhibition planning scenario		
		smalltalk	individual	dyadic
overall duration [min]	24.84	5.92	3.33	7.84
overall distance [deg/s]	198145	215341	108822	288771
average distance	1.22	3.54	3.32	3.5

Table 6.9.: Head movement distances per scenario. The left column gives the results from the animal guessing scenario, the right column from the smalltalk phase and the dyadic phase in the exhibition design study.

Discussion The much shorter overall distance in the animal guessing scenario can be explained to a great extent by the much longer duration of the task. However, the average distance value also shows a noticeably shorter distance than for the other three scenarios in the exhibition planning study. Meanwhile, these average distance values do not differ to a great extent when compared to each other. Why is there such a great difference between the average distances in the animal guessing scenario to all three tasks of the exhibition planning scenario? The reason can be that the task, which was to think of an animal and to answer yes-no-questions from the interaction partner trying to guess it. Firstly, this task might lead to longer phases of thinking (while keeping the head still) between the questions. Secondly, we know that speakers normally do not continually look at the listener but instead only at specific times during speech production (see Section 1.6.3). Maybe, because of the forbidden verbal answers, the speakers are not that free to look away from the listener because they might miss the nonverbal answer of the listener.

Thirdly, talking people use their head movements not only to give back-channel feedback but also to structure their utterances (e. g. stressing words, see Section 1.6.3). According to McClave (2000), head movements accompanying the speech have semantic and structural functions. Maybe speakers move their heads less when asking yes-no-questions compared to smalltalk. If the speaker produces exclusively questions which are moreover of a short duration (e. g. "Is it a mammal?"), it might be possible that these questions are accompanied by only very few head movements. For example, a single head movement might accompany the word "mammal" that would convey the information that the question is ending here and that an answer is expected. Since the structure of these questions is not complex, not many other structural head movements might be needed. To our knowledge there is no literature yet that compares the speaker's and listener's head movements during yes-no-questions with the head movements during spontaneous speech. The relationship of questions and head movements can be especially analysed in a follow-up study.

Additionally, the previous section on automatic speech recognition found shorter durations and fewer occurrences of utterances for the introductory phases compared to the other phases. In this section, we found shorter average head movement distances for the animal guessing scenario compared to the other scenarios. Unfortunately, there is no speech data in the animal guessing scenario and no head gesture data in the introductory phases. An interesting question would be to investigate both phenomena (utterance durations, head movement distances) in all available scenarios in order to further investigate the influence of the scenario on these interaction signals.

6.3.7. Outlook

Despite the promising results from the offline analysis, the results for the online analysis are not yet satisfactory. The number of classifications that have to be corrected is certainly too great to rely on the classifications at this stage.

Nevertheless, this work provides a general method to include automatic classifications of head gestures and a general approach to analyse rich measures of such annotated and classified patterns. Modifications that should be considered in the future are re-adjustment of the parameters for the HMD condition or a re-training of the classifier with head gestures from the same scenario since both the HMD and the scenario seem to affect the characteristics of head movements. Moreover, the training data should be redundantly annotated (by trained annotators) to make sure that the head gestures used for training are not superimposed by other movements and preceded or succeeded by a short interval in which no movements takes place. Finally, an idea is to investigate the probability of head gestures combination with the speech recognition. As Section 1.6.3 describes, there are specific phases in utterances where a back-channel signal is particularly likely. These probabilities could be used to improve the head gesture recognition. The following section will investigate such correlations.

6.4. Timing of speech and head gesture data

The previous sections presented the prospects for computer-aided investigation of separate behavioural cues as focus of attention, speech and head gestures. Apart from these separate

investigations, it is also possible to analyse the interplay of these cues. Joining the annotated speech times with the annotated head gestures, we are able to investigate complex multimodal phenomena as the timing between utterances and head nods.

6.4.1. Method

The aim of this section is to show how an analysis of the timing can be accomplished. For this, we will traverse all analysis steps exemplarily. Subsequently, we will give an outlook that demonstrates which kinds of results can be achieved. Speech and head gesture data were taken from the interactive exhibition design study (see Section 5.3). The tags from the classifiers for both signal cues were corrected manually for a 5-minute interval in the intro phase. Using the analysis toolbox, the ELAN file (.eaf) can be imported⁴ with all its annotations and all linked Files (media and timeseries files), thereby preserving the configurations (e.g. data offsets from synchronization). As a general overview of the data, they can be plotted⁵ in MATLAB, as shown in Figure 6.6a. Since the research question is to calculate the timing of head gestures in comparison to the speech of the interaction partner, only those phases of the study have to be included in the analysis, where conversation took place. The exhibition design study consisted of more than one scenario. Thus, we want to reduce⁶ the data to a special phase/task of the study. The reduced/sliced data can be plotted again in order to compare it with the non-sliced data, which is shown in Figure 6.6b.

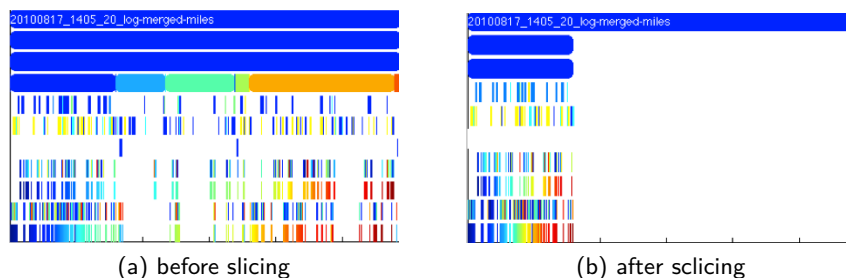


Figure 6.6.: Data file sliced according to one phase of the study and plotted in MATLAB.

Once the data is reduced to the correct phase of the study, we begin analysing the timing of the data. For this, we have to find for each head gesture annotation of participant A the preceding or overlapping speech annotation of participant B (and likewise for each head gesture annotation of participant B the preceding or overlapping speech annotation of participant A). The distance between the starting point of the head gesture and the stopping point of the utterance has to be calculated⁷. The possible distance cases are illustrated in Figure 6.7. The annotations of the two blue tiers are to be compared. Thereby the upper blue tier represents the utterance tags of one participant while the lower blue tier represents the tagged nods of the other participant. The grey line in the middle shows the distances calculated from both tiers. If nods overlap with the utterance, the distance value is negative

⁴elanReadFile.m

⁵elanPlot.m

⁶elanSlice.m or elanTimeseriesSlice.m

⁷elanCorrelateTiers

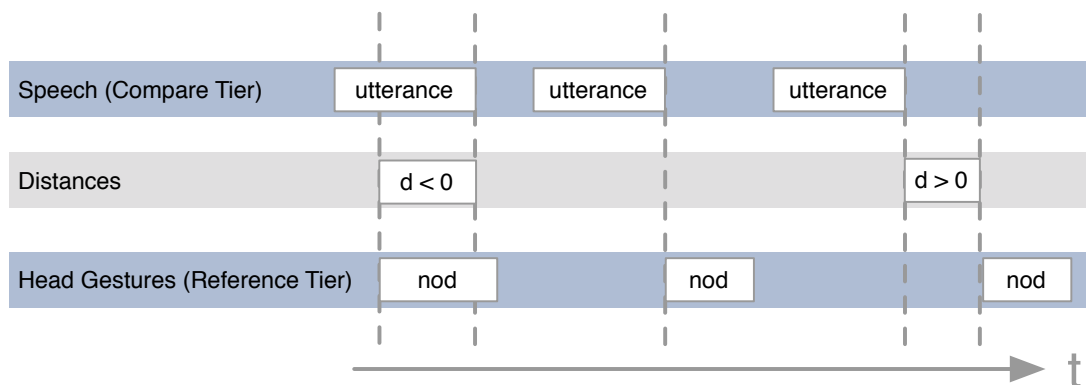


Figure 6.7.: Schema: distance calculations for annotations

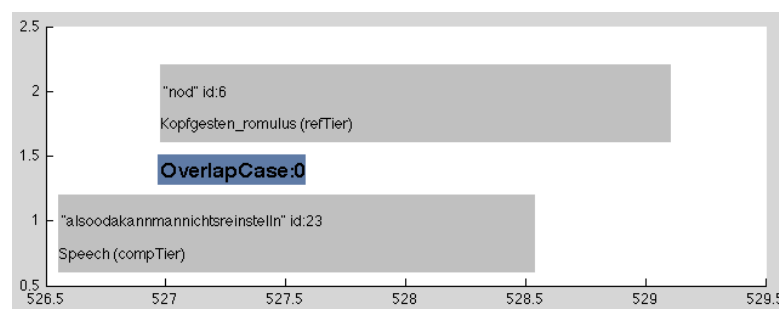


Figure 6.8.: Screenshot of MATLAB plotting overlap between two annotations. The overlap case specifies the kind of extension (0 means the annotation in the lower tier extends the annotation in the upper tier).

(left example) and if there is a short interval between the end of the utterance and the start of the nod, the calculated distance is positive.

According to these distance calculations, the function returns a list of distance values. For an intuitive supervision of these calculations, the script plots each distance comparison.

The analysis provides us with all measured distances for each participant. These can then be analysed using descriptive or inferential statistics in order to describe the resulting data.

6.4.2. Outlook

For an outlook on possible results that can be achieved using this technique, we plotted a histogram showing the distribution of the measured distances (see Figure 6.9). The histogram shows that the distances are distributed both in the negative as well as in the positive part of the scale. Additionally, the highest number of occurrences lies in the negative scale (as the tooltip shows in bin 16 ranging from $[-0.516, -0.346]$ with the bin centre of -0.431 seconds). Moreover, for this data sample, there seems to be an interval from $[-3$ to $-1.8]$ seconds where no nods occur. There is a similar but shorter interval around zero where again no nod occurs in this sample.

The histogram might indicate that most of the nods occur shortly before the utterance ends.

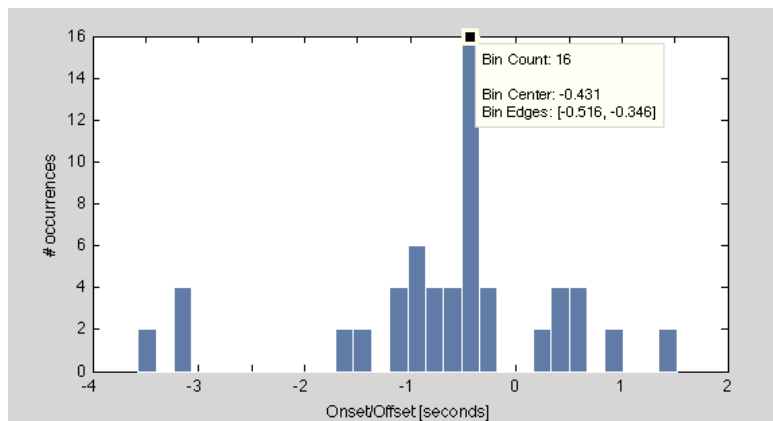


Figure 6.9.: Histogram of the measured distances between speech and head gestures.

Additionally, there are intervals where no nods occur, particularly in the interval around zero and the interval from $[-3$ to $-1.9]$ seconds. This means, nods occur in this sample very often in the finishing of utterances while most of them occur before the speaker finished, none of them exactly at the end of the utterance, and some slightly afterwards. However, since the data is computed from a very small sample, these preliminary hypotheses should be reviewed in a larger data corpus. Since this would require the correction of the speech and head movement data of a larger sample, it is outside of the scope of this thesis.

Nevertheless, this example demonstrates how this method of combining multimodal cues can be used for analysis: the timing between arbitrary behavioural cues (either tagged automatically or manually) can be analysed. Moreover, since ARbInI provides various methods to analyse time series data these can be combined with the timing analyses.

6.5. Summary

This Chapter addressed the possibilities of computer-aided investigation of interaction. The sections above discussed three different approaches for this: automatic speech annotation, automatic tracking of the objects in the field of view, and an automatic classification of head gestures and evaluation of head movements. Additionally, we combined two of the methods and computed the average distance of utterances of one participant to the head nods of the other participant. This section will discuss the results of these methods.

Concerning the automatic tracking of the objects in the field of view (see Section 6.1 we argued that it is possible to roughly determine the focus of attention from the field of view when using head-mounted displays. We presented a method to track the objects in the field of view and automatically transfer the resulting object positions to a video. The simple videos showing object ids on black background are helpful for the researcher. However, a more detailed analysis of the resulting data was beyond the scope of this thesis. In further analyses, the method can be used for automatically calculating durations in order to analyse the amount of time each object was in the field of view for one or both participants. This may allow further conclusions about how the interaction partners establish joint attention. Besides, the automatic tracking can be extended easily to monitor glances at the interaction

partner and automatically tag intervals of mutual gaze.

The automatic speech recognition in Section 6.2 was used to tag the speech times during one study. Investigating the speech times of three scenarios of this study, we found notably differences between the introductory phases of the study and the smalltalk and dyadic discussion phase in the utterance duration and numbers. We discussed possible reasons for this concerning the characteristics of the introductory phases. Apart from the conducted analyses, the method can also be used to analyse the timing and overlaps between speaking turns or combined with other cues that can be automatically analysed as voice pitch, focus of attention and head gestures.

For the automatic tagging of head gestures in Section 6.3, we investigated the approach both offline and online and developed a special method for calculating the amount of work a human corrector would have to do. While the offline analyses showed promising results, the online performance is not yet satisfactory. We discussed modifications concerning the training data set as well as improvements for the classifier. An improved method for head gesture recognition holds further prospects: since the participants' eyes are focused on the table and the field of view is reduced because of the head-mounted displays, users are likely to miss head movement signals of their interaction partners. An interesting idea is to provide the tracked head gestures as a sonification to the partners. With such a technique, the drawbacks through the reduced field of view might be compensated at least partially.

Furthermore, the section presented a head gesture data corpus consisting of about 1900 head gestures and continually head movements recorded from different non-HMD scenarios. Although we recorded an additional corpus containing head gestures during HMD-mediated interaction in the interactive exhibition design study in Section 5.3, these were not included here since the focus of this chapter lies on non-HMD interaction. Two of the non-HMD corpus parts were analysed for the overall and average movement distance. The results suggest that the animal guessing game elicits other head movement behaviour than the other scenarios under investigation. We discussed possible reasons for this concerning the characteristics of the animal guessing scenario.

Section 6.4 combined the data from speech recognition with the data from the automatic tagging of head gestures. The results seem to suggest for this specific scenario that the participants mostly begin with their nods shortly before the speaking participant ends the utterance. However, since this is analysed only for one participant pair over a short interval, this should be further investigated in future. Nevertheless, the method can be used for various different research questions investigating the timing of different interaction cues.

7. Actively influencing interaction with ARbInI

The previous chapters introduced ARbInI and examined its influence on the user's behaviour as well as the possibilities for the analysis of interaction both with and without the use of Augmented Reality (AR) features. The aim of this chapter is to actively influence the interaction, particularly to enhance or disturb the interaction. Why can we learn something about interaction by influencing the interaction? By disturbing or enhancing the interaction, we introduce situations that are new to the users. Thereby, we can analyse the interaction in such unfamiliar situations in a systematic way. Moreover, we can compare the interaction signals that the participants use in such unfamiliar situations with those that they use in well-known situations.

Firstly, Section 7.1 will examine how human-human interaction can be influenced with ARbInI in a positive way. More specifically, we aim at supporting the interaction with the help of features that are provided by multimodal AR.

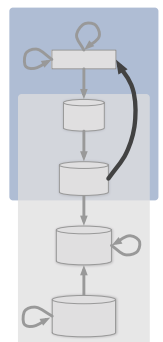
Secondly, Section 7.2 will discuss how we can actively disturb the interaction with the help of features that are introduced by the close coupling of the system and the users.

7.1. Guiding attention

This section presents an approach to enhance the interaction by supporting the accomplishment of the task with a multimodal attention focus display. This display provides information about the partner's focus of attention to each user by mutually coupling the two setups of the ARbInI system as described in Section 3.5. The following study evaluates the effectiveness of the system in a simple search task by means of reaction time and error rates.

As Section 5.1 argued, AR systems on the one hand augment the real world but on the other hand they also reduce the user's perception of this world (e. g. by reducing the wearer's field of view and masking the eyes of the interaction partner). In order to oppose this, we decided to re-enhance the interaction by giving the participants the audiovisual augmentations explained in Section 3.5. The study was also published in (Dierker et al., 2009b)

We assume that these audiovisual augmentations using the AR goggles will only enhance interaction when compared to a setting also using AR goggles for the simple reason that currently these devices are too disadvantageous in terms of lag, resolution, field of view, weight, shutter time, monoscopic vision and dynamic range (see Section 5.1). Compared to a setting where both participants wear no AR devices we therefore do expect the general problems of AR to outweigh the benefits of the augmentations to enhance interaction. However, our point is that we are able to at least partially compensate some of the disadvantages of AR devices (mainly the reduced field of view and the hampered head movements) with this technique.



Our hypotheses are:

Hypothesis 5 (Reaction time)

The participants will exhibit a shorter reaction time in the condition with both augmentations compared to the condition without augmentations.

Hypothesis 6 (Error rate)

The participants will show a lower error rate in the condition using the audiovisual augmentations compared to the condition without them

7.1.1. Method

The following section will give an overview of the experimental setup for the study. Beginning with a description of the tested conditions and the measurements the section will proceed with information about the sample, the stimuli and the procedure of the experiment.

To test if the augmentations actually improve the interaction using AR via HMDs, we chose to compare two experimental conditions: the “highlighting on” condition where both the visual and the auditory augmentations were provided and the “highlighting off” condition where neither visual nor auditory augmentation was given. The reason not to distinguish between the visual and the auditory augmentations in the experimental conditions was to test every pair of participants in all conditions while still not overburdening the participants with a very long experiment (most people get tired wearing the AR goggles over a longer period of time Arthur (2000)).

As dependent variables we consult objective as well as subjective criteria to measure the effect of the “highlighting on” condition. Objective dependent variables are reaction time (the time needed to finish the task successfully) and error rate (the percentage of successfully finished tasks). The reaction time is measured using button presses on Wii Remotes¹. In each phase of the trial the participants have to press the button of the Wii Remote to continue with the trial. Error rates are calculated from offline annotations of the scene camera data.

Additionally, we determine the subjective user experience in a questionnaire with multiple-choice answers (see the German questionnaire and the translations in the Appendix A.2). We asked the participants two questions to estimate their subjective usage of the augmentations. The first was “How much did you use the visual augmentations?”, the second was “How much did you use the auditory augmentations?”. There were four checkboxes for the answers and the scale ranged from “very much” to “not at all” with the intermediate steps not named explicitly. Moreover the participants answered the questions “How helpful did you find the visual augmentations?” and likewise, “How helpful did you find the auditory augmentations?”. Here, the scale ranged from “very helpful” to “not helpful at all” in four steps (intermediate steps were not named explicitly). There was a fifth possible answer “disturbing” for these two questions.

¹see www.nintendo.com/T1/textemdashcontrollers#remote

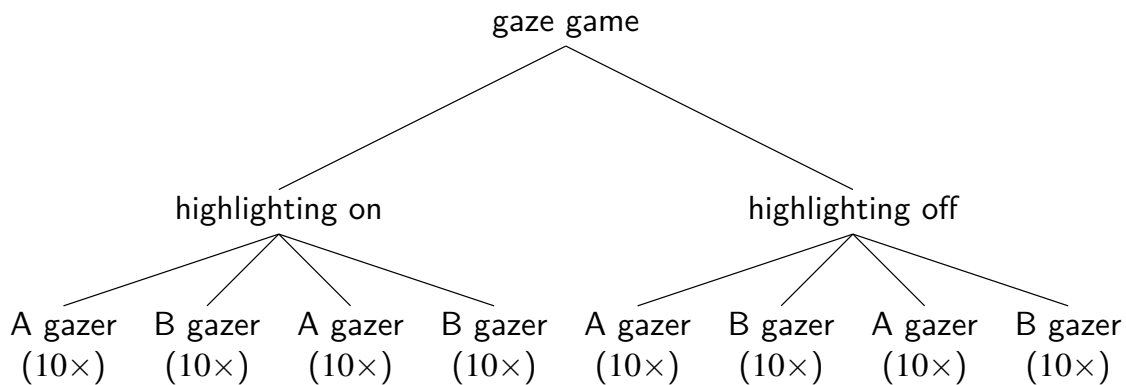


Figure 7.1.: Process flow of the gaze game. The tree is traversed left to right, every leaf being an action the participants had to take. The leaf nodes also indicate whether participant A or participant B takes the role of the gazer and that ten cycles are to be completed before switching roles. The order of “highlighting on” and “highlighting off” was interchanged for half of the participant pairs.

7.1.2. Scenario: the “gaze game”

The gaze game scenario is described in detail in Section 4.3.3. In each game cycle, the gazer looks at a certain virtual object and the searcher has to guess the object. After ten cycles the roles of gazer and searcher are switched. Additionally, after two blocks of ten cycles per person, the highlighting condition is changed, so that each pair of participants plays 40 cycles with augmentations and 40 cycles without. Specified by the order of their arrival, the pairs of participants began alternating with augmentations and without augmentations respectively. Figure 7.1 shows the process of the game.

7.1.3. Procedure for the study and sample

The trials for the study consisted of three parts: In the first part, the participants were equipped with the devices and were given time to explore and adapt to the AR system. Letting them sort and group the objects regarding several criteria supported this. The participants were asked whether they felt comfortable with the system before proceeding. The visual and auditory augmentations were switched on during this phase for all participants to give everyone the opportunity to get used to the full system. The second part was the gaze game task we explained above. The participants wore the AR goggles and headphones the whole time. Additionally, their eyes were laterally shielded to prevent them from bypassing the goggles as some participants did in a preliminary study. During the whole task, time measurements of Wii Remote button presses were taken. Finally, subsequent to the gaze game, the participants were asked to put down the equipment and to fill in a questionnaire.

For this study, we tested 13 pairs of participants. Unfortunately, we had to leave out the data from two pairs of participants later. One pair, because they misunderstood the “highlighting off” condition in the gaze game task by guessing instead of looking at their partner and estimating the partner’s gaze direction. The other pair could not finish the task due to time

limitations. The remaining 22 participants were 11 male and 11 female. Their age ranged between 18 and 28 years and the mean age was 21.8 years. The majority of the sample were students. The participants were asked to schedule their appointment using an online tool. Thus, we had no influence on the composition of the participant pairs. Nine out of eleven pairs had different gender. The two remaining pairs had the same gender. Seven pairs did know each other beforehand. Six of the pairs of participants began with augmentations, five of them without. Each participant played only in a single pair. The duration of the gaze game task varied from 16 to 26 minutes with an average of 21 minutes. All participants were paid a fee for their participation.

7.1.4. Results

From the gaze game part of the study, we measured the reaction time and computed error rates. We annotated all scene camera videos offline according to the participants' success in the task. Thereby, we left out all trials that could not be rated (because of disturbances or technical problems) for computing both the reaction times and the error rates. The performance results are presented in this section as well as the questionnaire results.

Reaction time

Our Hypothesis 5 said that the participants will show shorter reaction times in the condition with highlightings compared to the condition without highlighting. In the gaze game task, the main goal to achieve was to get the right object not to be quick by all means. Thus, we exclusively considered the successful cycles for the reaction time because in the unsuccessful ones the searchers pressed the button before they had actually found the right object. Time measurements were taken from logged timestamps of Wii button presses: we computed the difference of the button press constituting Event *a* and the one being Event *b* (see Figure 4.4b on page 61). The mean search time² for the “highlighting on” condition is 2.56 ± 0.98 seconds whereas the mean search time for the “highlighting off” condition is 4.16 ± 1.97 seconds. The comparison of the means for reaction time with paired two-sample t-test³ showed statistical significance [$T(10) = -2.5; p = 0.03$].

As visualization, a boxplot is shown in Figure 7.2 for both conditions. The boxes span between the lower and the upper quartile, the horizontal lines are the medians whereas the thick black bars depict the mean values. Whiskers show the minimal and maximal search times. For each participant pair the conditions are given in the correct order (first condition on the left). The “highlighting on” condition is represented with red boxes whereas grey boxes represent the “highlighting off” condition. The last two boxes show the overall search times for both conditions (yellow (left) for the “highlighting on” condition and blue (right) for the “highlighting off” condition). For this plot only the cycles with correct outcome were considered. The means for the overall search times visualize our previous results: the mean for the “highlighting on” condition is lower than for the “highlighting off” condition. Considering the means for the participant pairs, we can see that except for participant pair 6, all means of

²notation: arithmetic mean \pm standard deviation

³Homogeneity of variances was shown by f-test [$F(10, 10) = 0.25; p = 0.04$], both samples have underlying normal populations.

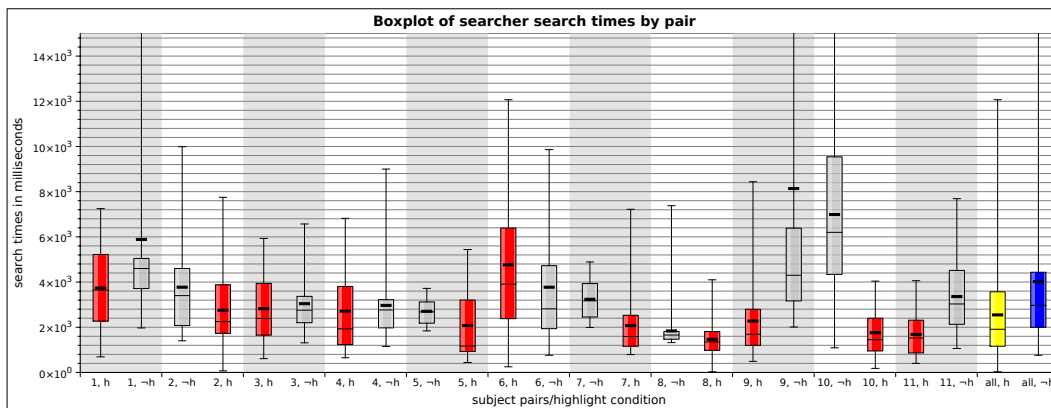


Figure 7.2.: Boxplot of the searcher's search times. Boxes show the interquartile range and the medians (horizontal lines), thick black bars depict the mean values. Whiskers show minimal and maximal search times. For each participant pair, the conditions are given in the correct order (first condition left). Red (darker) boxes represent the “highlighting on” condition, grey (brighter) boxes represent the “highlighting off” condition. The last two boxes show the overall search times for both conditions (yellow/brighter for the “highlighting on” condition and blue/darker for the “highlighting off” condition). Only the successful cycles were considered.

the participant pairs are lower for the “highlighting on” condition than for the “highlighting off” condition.

Error rates

Our Hypothesis 6 expected the participants to show a lower error rate in the condition using the audiovisual highlighting compared to the condition without. For the “highlighting on” condition, the participants made an average of $2.72 \pm 3.41\%$ errors whereas for the “highlighting off” condition the average error rate was $36.86 \pm 15.88\%$. Strong significance is found by a paired two-tailed t-test⁴ from the error rates [$T(10) = -6.39; p < 0.01$]. A comparison of both graphs in Figure 7.3 visualizes this result.

Questionnaire

For the question “How much did you use the visual/auditory augmentations?” the questionnaire results are shown in Figure 7.4a. We found that most participants rated their usage of the visual augmentation very high while they rated their usage of the auditory augmentation very low. Thus, there is a clear difference between the rated the usage of the visual augmentations compared to the auditory augmentations.

For the question “How helpful did you find the visual/auditory augmentations?” the answers are presented in Figure 7.4b. Most participants found the visual augmentations helpful.

⁴Homogeneity of variances was shown by f-test [$F(10, 10) = 0.05; p < 0.01$], both samples have underlying normal populations.

However, 16 participants rated the helpfulness of the auditory augmentations on the negative half of the scale and 4 participants found them even disturbing. Thus, there is a clear difference between the valuation of helpfulness for the visual versus the auditory augmentations.

7.1.5. Discussion

Our hypotheses were that the participants have a *lower error rate* in the “highlighting on” condition compared to the “highlighting off” condition and similarly that they exhibit a *shorter reaction time* in the “highlighting on” condition.

For the reaction time, we found a significant difference between both highlighting conditions and thereby support our Hypothesis 5. Moreover, only for one participant pair, the mean in the reaction time for “highlighting on” condition is lower compared to the “highlighting off” condition. We suppose that this is due to a lack of faith of this participant pair in the reliability of the highlighting which resulted in multiple verifications of the assumed view direction each time before communicating the decision. Nevertheless, we can conclude that the audiovisual augmentations cause a lower search time for the searcher in general.

Moreover, the measured error rates support our Hypothesis 6 because they are significantly lower for the “highlighting on” condition. This means that in sum, both augmentations (visual and auditory) together induce a faster and less error-prone task completion.

We propose the following four changes we made to the task between the pre-study and the study presented here as possible explanations for the difference in the results of the pre-study and the presented study: Firstly, the gazers are no longer allowed to choose an object by themselves because we present a random object to them. Secondly, the positions of the objects on the table are shuffled before each cycle so that the searcher cannot learn the objects by heart. Thirdly, we increased the number of possible objects from five to six. Fourthly, we prevent the participants from looking past the goggles by blinders. To explain that the task became more difficult yet the error rate of the “highlighting on” condition was not affected, we suggest the improved visual augmentation. While the highlighting was simply yellow for all objects in the partners’ field of view in the pre-study, the highlighting in this study is a colour gradient from the centre of the partners’ field of view to the outer region. We consider a combination of both randomization techniques and the increased number of possible search objects to be the crucial factor for the different error rates for the “highlighting off” conditions of both studies. Nevertheless, there is no proof for this hypothesis yet. This effect might as well have been caused by the influence of the blinders which should be investigated in a subsequent study.

Concerning the user experience measured by the questionnaire, we found a pronounced difference between the augmentation modalities. The participants found the visual highlighting much more helpful than the auditory highlighting. Some even rated the sonification distracting. Similarly, they rated their usage of the visual augmentations much higher than their usage of the auditory augmentations. We suggest three possible reasons for this: Firstly, there is much less information conveyed by the auditory augmentations than by the visual ones. While the visual augmentations make clear which object is being looked at and how centrally, the auditory augmentations’ main potential function is that of an activity monitor. Even

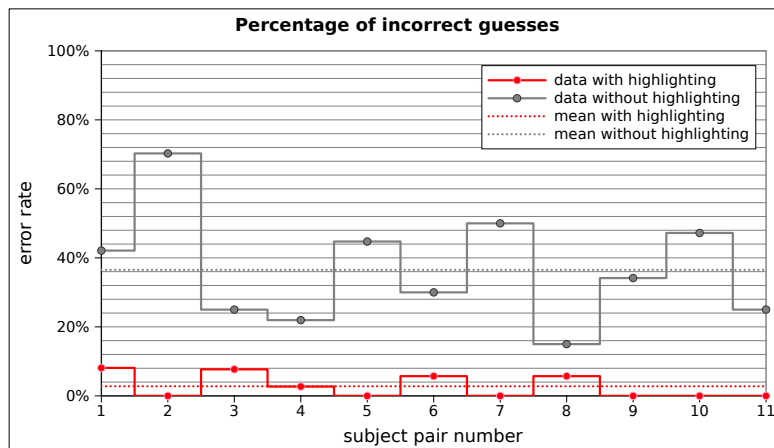
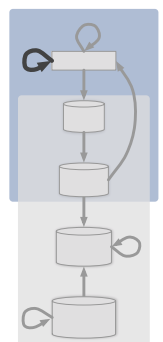


Figure 7.3.: Error rates for both tested conditions. The red dots (on the lower curve) belong to the “highlighting on” condition whereas the grey dots (on the upper curve) belong to the condition where the highlighting was switched off. Note that the connection of the dots does not indicate interim values. Dotted horizontal lines specify the average values for both conditions.

this was impeded by the fact that even moderately fast head movements caused the image to be blurred to the point where the markers were unrecognizable. More sophisticated sonifications might therefore still be promising (see Section 3.5.1.2 and Mertes (2008); Mertes et al. (2009)). Secondly, the visual modality is more commonly used for joining attention in everyday interaction than the auditory one. Therefore, there might be a training effect which could be shown by longitudinal studies. Thirdly, auditory cues could work more on a subconscious level than visual ones. Separating the highlighting modalities consecutive studies should investigate this.

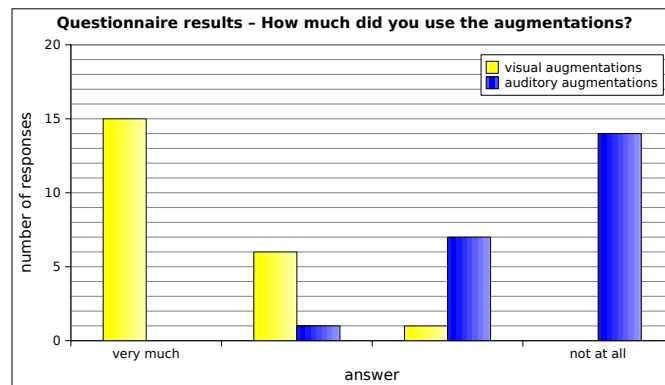
7.2. Disturbing interaction

This section discusses the possibilities of the ARbInI system to disturb the interaction of its users. The first section describes the general approaches of other researchers to investigate and provoke disturbed interaction followed by a description of the methods that were used in this work. These methods use multimodal Augmented Reality (AR) features of the ARbInI system to introduce misunderstandings between the two users. Two methods have been introduced in Section 3.2: (a) by changing static characteristics of the virtual objects (b) by affecting the interaction behaviour of the virtual objects. Both of these methods can be introduced with auditory or visual AR features as will be discussed in the following. After this, we present initial observations on pre-tests exploiting these approaches.

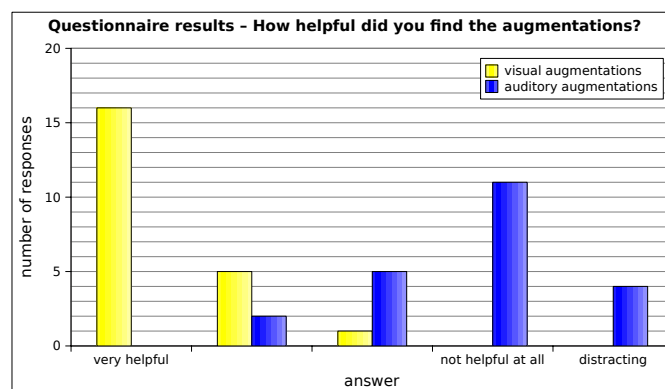


7.2.1. Review of literature

Why is it interesting to disturb the interaction? Everyday interaction does not always run smoothly (Chomsky, 1965). Often, speech is aborted or alienated, sentences are re-formulated



(a)



(b)

Figure 7.4.: Answers from questionnaire concerning usage and helpfulness of the highlighting.

midway or ungrammatical fragments occur. In fact, most smalltalk includes a noticeable amount of word- or grammatical errors (Schegloff et al., 1977; Goodwin, 1980). Interaction partners also experience both unintended and intended (e. g. jokes) misunderstandings. They might disagree on topics as well as on the meaning of words. Even lies and deceit occur in some interaction situations Ekman et al. (1991); DePaulo et al. (1983). Thus, it is not only interesting to analyse interaction during perfect conditions but also when misunderstandings occur.

Misunderstandings can occur on the phonic, lexical, syntactic, semantic, pragmatic, as well as situational level (Zaefferer, 1977). In research of misunderstandings, the sources of misunderstandings have been investigated (e. g. Schegloff (1987)) as well as the way in which the participants cope with misunderstandings (Bazzanella and Damiano, 1999). According to Verdonik (2010), there are still many open questions. In the following two sections we will shed a closer look on *lexical misunderstandings* and *semantic misunderstandings*, their respective repairs and how the experimenter can provoke such misunderstandings.

Inducing lexical repairs with auditory AR

A short overview of repair mechanisms was given in the introduction to the topic in Section 1.6. There, we introduced verbal as well as nonverbal repair initiation strategies (this means

strategies that signal a following/meanwhile repair). Particularly, we noted the frequent use of head shakes and pauses as initiation signals.

Apart from human-human interaction, repairs are also especially interesting in the field of human-machine interaction. The design and development of virtual agents and robots has to take repairs into account since the machine has to distinguish the repairable⁵ sequence from the repaired sequence and thus has to find intended message in the human speech. Since this is a complex task, it is not surprising that speech repairs decrease the accuracy of speech recognition to a noticeable extent (Shriberg and Stolcke, 1996; Young, 1996). There have been approaches to increase the speech recognition accuracy by including a repair detection based on signals that co-occur with the speech repairs. The general idea is to combine the speech recognition with such nonverbal signals that are available *without* word recognition. The researchers use for example signals as pauses (Bear et al., 1992), filled pauses (e.g. 'uh') (Goto et al., 1999), prosodic features (Liu et al., 2006) or head movements (Chen et al., 2002).

Despite these approaches, the nonverbal signals that warn the interaction partner of a repair are still not fully understood. For their research, the researchers need training corpora that show different types of repairs. Some of the above approaches used existing speech corpora for their training. The problem is that such corpora might not be available in every language (which is important since nonverbal signals are cultural-specific as Section 1.6 discussed). Moreover, those corpora mostly differ in their interaction situation (e.g. in terms of topic, scenario or noise). This means, that there are no training corpora readily available that can be used to train the speech recognition of an arbitrary agent or robot in an arbitrary scenario. Others of the above approaches manually created training corpora for their specific interaction situation, which is a laborious process. Furthermore, the researcher does not know if and how many repairs will occur in the recorded interaction and, thus, has to literally wait for repairs to occur. Thus, it would be easier to create corpora if the experimenter were able to provoke such misunderstandings at arbitrary points of time during the interaction. Using the `C5MIXER` of ARbInI (see Section 3.2), it is possible to filter and distort the participants' utterances. This can be used to induce the lexical misunderstandings sought for. Then a researcher no longer has to *wait* for spontaneous misunderstandings to occur but instead can provoke misunderstandings at any point in time and on arbitrary topics.

Inducing semantic misunderstandings with visual AR

So far this section considered repairs for abortions and word or grammatical repairs. However, misunderstandings occur also for example in reference, meaning or expectations (e.g. Schegloff (1987); Taylor (1992); Drew (1997)). However, we believe that such misunderstandings cannot be induced by the auditory technique mentioned above. A successful semantic disturbance would require not only distortions but also replacements of words or utterances. This is not only technically challenging but it would be even more difficult to make sure that the interaction partner who hears the replaced utterance does not notice that it has been replaced. Otherwise, if the replacement would be noticed, we believe the interaction partner would very likely react as if a lexical misunderstanding would have taken place as

⁵In accordance with others (e.g. Drew (1997)), we use the term "repairable" instead of erroneous to describe the word or utterance that will be repaired since not in each case there is an objective error.

we do for example on the telephone whenever a noise causes a word or an utterance to be incomprehensible. Thus, a semantic repair would not be initiated. In traditional research, the solution is to choose appropriate tasks that make misunderstandings likely to occur – an approach that is discussed by Healey and Thirlwell (2002) who state that this influences the recorded repairs very much. Moreover, the researcher still has to wait for repairs to occur. Another method to provoke misunderstandings is to employ a confederate/confidant. This is a person who seems to be another participant of the experiment but actually is a co-experimenter who acts as an actor/actress. Section 3.2 already discussed the problems and benefits of using confederates in interaction research. It was argued that this approach increases the risk of influencing not only the interaction situation but also the characteristics of the repair that is to be investigated. However, if we want to learn about repairs, such influences might bias not only the experiment but also the results and conclusions.

Another approach to induce misunderstandings is to use the help of video-mediated AR. With its virtual objects, the ARbInI experimentation system is particularly suited to disturb (misguide or deceive) the participants. As Section 3.2 described, static characteristics and interaction behaviour of the virtual objects can be varied during the experiment. This can be used to induce misunderstandings in reference, meaning or expectations. This again means that the researcher can actively disturb the interaction in order to induce such misunderstandings in any situation.

Analysing repairs

The resulting repairs of both approaches can subsequently be analysed, e.g. for the initiation signals. Several researchers emphasized the importance of nonverbal signals for signalling repairs. Here, ARbInI offers rich possibilities: the system allows analysing speech times, intonation, visual attention and head movements as well as their respective timing.

7.2.2. Method

Above, we proposed visual disturbance methods for the investigation of semantic misunderstandings. Of the two discussed methods we chose the disturbance method that alters the interaction behaviour of the virtual objects and applied it in pre-tests. These pre-tests took place during the study investigating the enhancement method that was described in Section 7.1. In the study, the highlighting feature that was introduced before and that the participants learned to employ was disturbed. Overall, the experiment consisted of 80 cycles, 40 of them without highlighting (and thus without disturbance) and 40 with highlighting. In order not to disturb too many cycles, the method was applied in random cycles during the experiment. In total, 27 cycles were disturbed. 0 to 6 disturbances were applied to the individual pairs, which resulted in average disturbance of $2,45 \pm 1,86^6$ cycles per participant pair. Of the 11 participant pairs, 9 encountered disturbances, 2 did not encounter disturbances. If applied, the disturbance was always applied to the participant that chose the seat to the left of the experimenter (B), never to the participant sitting right (A). This means that this disturbance only was applied to the cycles where A was gazer and B was searcher (see

⁶<mean> ± <standard deviation>

Figure 7.1 and Sections 4.3.3 and 7.1.2). An advantage of this is that this enables us to investigate mistrust of both participants towards the system and towards their interaction partners.

7.2.3. Initial observations

For an overview of the observations, we computed the overall experiment duration as well as the mean duration of the search phase during both disturbed and non-disturbed cycles. These values were computed for all participant pairs. Additionally, we chose one participant pair (No. 9) with the highest number of disturbances (in this case 6) for a closer investigation. For reference, we computed the above values also for this participant pair. The Results are shown in Table 7.1.

	all participant pairs	participant pair 9
mean duration for search phase		
- in non-disturbed cycles [seconds]	2.56 ± 0.98	2.27 ± 1.59
- in disturbed cycles [seconds]	8.26 ± 5.75	14.32 ± 9.05
disturbed cycles	$2,45 \pm 1,86$	6
overall duration of experiment [min]	21 ± 3	26

Table 7.1.: Results of disturbances. The left column gives the results from all participants as average values per participant pair, the right column the results for the participant pair that is analysed below in more detail.

The mean values show that the duration of the search phase in the disturbed cycles is much longer than in the non-disturbed cycles. This is also true for participant pair 9 where the difference between the disturbed cycles and the non-disturbed cycles is even more pronounced. In order to gain more detailed insight in what is different between the game cycles that are disturbed and the ones that are not disturbed, we will present two dialogues of both conditions. The following sequence shows a typical utterance sequence as it occurs in the cycles of the gaze game *without* disturbance.

```
01 A [focuses on the car and presses her button]
02 B: [presses button] Auto?
      Car?
03 A: m-hm
```

As the sequence shows, the utterances are short and the search cycle is finished very fast. However, if the disturbance technique is applied, the dialogue can differ from the above one. The chosen participant pair experienced 6 disturbances, 2 in the first ten cycles and 4 in the third ten cycles (please refer again to Figure 7.1 for the process of the gaze game). The following dialogue demonstrates this:

```
01 A [focuses on the car and presses the button]
02 B: Boh, ich kann hier grad gar nix einsch"atzen
      Wow, i cannot judge anything
```

[laughs; leans to the side trying to follow the gaze direction]

03 B: Also jetzt ist es ehrlich gesagt nur—ich tippe mal auf
irgendwas
Well to tell the truth, that is only—i guess something
[brings the chair backwards]

04 A: irgendwas ist richtig
Something is correct
[laughs]

05 B: Du guckst, ne?
You are looking, are you?

06 A: ja ...
yes ...
[shrugs shoulders]

07 B: Das Haus.
The house.

08 B: ja genau.
Yes, exactly.

09 [B presses button]

In this dialogue, participant B shows her unsureness in her utterances when the highlighting is disturbed: In 02 and 03 she tells her partner that she has problems to judge and that she will have to guess. Participant A answers with a joke in 04. In Line 05, participant B asks for explicit confirmation that A is really looking at the object to search, which is confirmed by A in 06. In 07 B finally says that she believes her partner being focusing the house, which is confirmed by B in 08.

Participant B's unsureness is not only shown in the utterances but also in her behaviour. While participant A focuses on the house, participant B first looks for highlighted objects on the table (Figure 7.5a), then she leans to the side trying to follow the gaze direction (Figure 7.5b), then she brings her chair backwards and tries again to follow the gaze direction from this different angle (Figure 7.5c). Finally, she brings the chair again forward and chooses the house (Figure 7.5d). Thereby, the participant pair seems to forget about the time constraints and takes instead significant time to joke and repeatedly verify the viewing direction. B's delayed button press in 09 furthermore shows this, which she should already have pressed before utterance 07.

In sum, in this example, participant B seems to compensate the disturbed highlighting by trying to follow the partner's gaze manually (as in the non-highlighting condition). For this, she alters her viewing angle on her interaction partner. During this phase, both joke and seem to ignore the time constraints. In other cases, we also noticed unsureness in the participants and similar compensation strategies as the above mentioned multiple visual verifications and multiple questions.

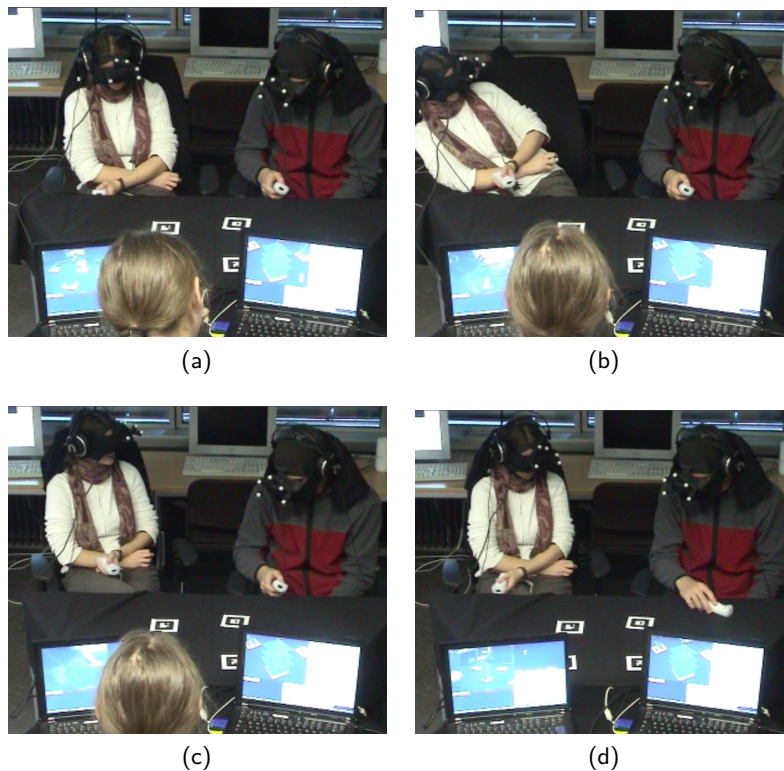


Figure 7.5.: Movements of participant B (left) during a disturbed cycle.

7.2.4. Discussion

In the investigation of the implemented auditory and visual disturbance methods that aim at investigating lexical and semantic misunderstandings respectively, we focused on one of the methods of which we chose the visual techniques. While ARbInI offers the same features for the investigation of repairs for both techniques, the reason to choose the visual disturbance methods lies in the inducing technique. Although ARbInI is particularly suited to induce acoustic disturbances, the technique for inducing the disturbances could also be used in other auditory-mediated communication as for example on the telephone. The visual disturbance method that aims at investigating semantic disturbances, however, is only possible using AR in this way. This is because the video see-through technique promises particularly intense illusion of reality (see Section 5.1). Additionally, the participants share the same interaction space with virtual objects that are anchored at real world objects. Together, we expect these characteristics to support the participants in believing that they see objects with equal characteristics and that the interaction behaviour stays consistent during the experiment.

The initial observations seem to support this theory. We were able to induce disturbances in 27 cases. Analysing the examples, we could show unsureness and both verbal and nonverbal compensation strategies in the disturbed participant. Since the disturbances did only affect one of the two participants, the participant pairs often suspected their interaction partner to perform the task in a non-correct way. However, one participant (whose partner was disturbed) assumed a technical reason and tried to find a compensation strategy. Unfortunately, a more

detailed and particularly a quantitative analysis of the whole data corpus was beyond the scope of this work. This will be topic of future research. Furthermore, the disturbance method has so far only been exploited in a study that did not record head gestures and speech. Here is room for many multimodal investigations that focus on the role of these behavioural signals in disturbances. Additionally, this method of controlled inducing of disturbances may also have some risks. Some of the side-effects have been investigated in Chapter 5. With further studies it has to be investigated whether the induced disturbances elicit such behaviour in the participants that is in fact similar to the behaviour in handling misunderstandings.

With the ongoing developments in mediated communication, the proposed approach is also promising for the investigation of problems and the participants' attempts for handling them in mediated communication: For example, we noticed during the experiments of this thesis, that some participants in the beginning asked if they will see the same objects as their interaction partner (which was confirmed in the studies without disturbance). Interestingly, this happened *before* the task began. In the progress of the task, however, some of them had difficulties with accepting tracking errors and understanding that these errors are only perceptible for themselves and not for their partners although exactly these errors were explained in the introduction by the experimenter.

Also during the experiments that induced disturbances, the participants showed a huge faith in the correctness of the augmentations as described above. Apart from one participant, all others seemed to suspect their own or their partner's behaviour to be the cause of the disturbance. Frohlich et al. (1994) also describe this phenomenon and report that their users often treat the machine as faultless.

Together with the promising preliminary results, we can conclude that this method is useful to induce arbitrary disturbances in human-human interaction while ARbInI is suited to analyse the resulting repair strategies. Possible research questions for the future are:

- how do participants realize that there is a problem or misunderstanding?
- how do they let their partner know that there seems to be a problem?
- how do participants cope with disturbances: what are their compensation strategies? Which solutions or circumventions, explicit clarifications or changes of situation models do they explore?
- especially in mediated communication: (a) how many disturbances can occur until the faith of the participants in the correctness of the system dissolves? (b) if the faith is destroyed: do they develop new compensation strategies with which they treat the now-known attempt of the system to disturb?

8. Conclusion

The present work has developed contributions to research processes investigating nonverbal behaviour in human-human interaction with respect to four goals:

1. to develop new methods in hardware and software that facilitate the entire research process, thereby providing: (a) a closed-loop interface improving experiment control, data acquisition and automatic tagging (b) a research process infrastructure enhancing data conversion and analysis and (c) appropriate scenarios benefiting from the features of the closed-loop interface
2. to evaluate the closed-loop interface in terms of its influence on the interaction signals under observation
3. to exploit the research infrastructure in investigating interaction with the help of the aforementioned scenarios
4. to enable the active disturbance or support of the interaction mediated by the interface in order to investigate how users cope with problems in interaction and how they benefit from supportive information

The following sections will summarize the main results for each goal.

Contributions to the entire research process In Chapter 2, this thesis presented a closed-loop interface to investigate human-human interaction focusing on nonverbal behaviour. The interface is called the Augmented-Reality-enabled Interception Interface (ARbInI). ARbInI's aim is to facilitate experiment control, data acquisition and data tagging with the aid of modular hardware and software components.

Two participants are equipped with two identical wearable setups of ARbInI. Using head-mounted devices, ARbInI de-couples the users' sensory perceptions from the outside world, intercepts the interaction signals using sensors and re-couples them by means of multimodal displays. In contrast to other techniques for mediated communication such as telephone or video-conferencing, the users are still able to interact in a shared space.

By means of the interception, the interface is able to exploit multimodal Augmented Reality techniques to *control* the experiment. Thereby the interface prevents the experimenter from influencing the participants, which ensures comparable experimental conditions. Additionally, the control functions reduce the amount of work for the experimenter.

While intercepting, the interface also *records* the interaction signals from both participants' point of view thereby using multimodal wearable sensors. Each sensor is fitted for the recording of one specific signal type in order to enable an efficient analysis of the data. This allows for automatic *tagging* of those data thereby reducing the researcher's work of manual tagging by a great extent. For example, an automatic tagging of the speech times and a tracking of the focus of attention was integrated. Additionally, a head gesture classification approach was presented and evaluated in detail. A comprehensive head gesture corpus was accumulated

from natural interaction via manual tagging in order to be used as training data. However, it has to be acknowledged that the recognition rate for the head gesture tracking is not yet satisfactory. Possible modifications to improve the recognition performance were discussed in Section 6.3.7.

Apart from the closed-loop interface, this thesis also provided an infrastructure for the automatic preparation and analysis of the recorded (and pre-tagged) data (see Chapter 3). The methods of this infrastructure go hand-in-hand with ARbInI and automatically *convert* (transform and synchronize) the data for a joint display. This synchronized multimodal corpus can then easily be presented in an existing tool that allows browsing such data. Moreover, a toolkit offers to automatize (or script) the filtering and statistical *analysis* of the process. Furthermore, it enables the analysis of complex relationships between multimodal data types.

As a means of investigating interaction mediated by ARbInI, Chapter 4 presented and discussed a set of scenarios that exploit ARbInI's features and elicit interesting user behaviour. While 4 scenarios are used in this thesis to answer specific questions or to gain a head gestures as training corpus for the classifier, the remaining 4 scenarios focus on encouraging rich nonverbal behaviour by using AR features.

Evaluation of the closed-loop interface Using the scenarios developed, empirical studies investigated the effects of ARbInI's head-mounted displays on interaction in terms of eye movements, head movements, speech times and task accomplishment in Chapter 5. The first study investigated visual search and indicated that compared to non-HMD conditions, participants used more head movements and less eye movements if they used an HMD. In the second study, we found a general decrease in the use, movement velocity and duration of head movements. Particularly, we measured reduced velocity and repetitions in sinusoidal head gestures such as nods and shakes. These decreased head movements seem to be contradictory to the increased head movements in the first study. We suggested that this is due to a compensation strategy of the participants to increase the number of objects in their field of view, which besides reduced the amount of necessary head movements. Additionally, we found an increased number of utterances and higher task completion times in the AR condition. Generally, we concluded a significant influence of our HMDs on the investigated nonverbal behaviour. Thus, while results from AR-mediated interaction have to be questioned carefully before transferring them to non-mediated interaction, the approach becomes with nowadays' rapid technical improvements more and more attractive as a means for creating hypotheses and investigating interaction. Moreover, ARbInI offers to investigate *mediated* interaction under controlled conditions. With its modular approach, several other sets of equipment that do not include HMDs are also possible and have been used for investigating interaction.

Investigating interaction This thesis illustrated that ARbInI, as well as the associated analysis infrastructure, is particularly suited to investigate nonverbal behaviour like head movements, focus of attention and speech timing. In Chapter 6, the data preparation and analysis methods are used for an analysis of the interaction corpora that were generated in this thesis. Thereby, the prospects for research on timing and structure of interaction signals were illustrated. Furthermore, average values for characteristics (duration, appearance, repetitions) for nonverbal signals were evaluated and discussed with respect to the scenario.

Actively influencing the interaction The ARbInI interface can couple its users mutually so that information from one participant can be provided to the other one (Chapter 7). In order to investigate the possibilities of ARbInI to *enhance* interaction, a method to highlight the partner's focus of attention in the field of view was integrated. An empirical study showed that the users could benefit from these enhancements, which led to significantly shorter reaction times and error rates in a simple search task.

In order to investigate misunderstandings and problems in interaction, the possibilities of ARbInI to actively *disturb* interaction were discussed. A method to misguide the users' attention has been implemented and adopted. Although the analysis of the effects of such disturbances has barely been touched in this work, the preliminary results show the potentials for this technique in the analysis of such disturbed interaction situations.

Future perspectives

The achieved research results and the outlined research prospects suggest that the infrastructure presented in this thesis opens up new paths for multimodal interaction analysis. Nevertheless, many of those paths could only be touched very briefly in this work and demand further investigation or improvements.

For example, while the recording of the head movements was shown to be sufficient for the analyses, the visual recordings of the objects' positions in the field of view could be enhanced by one very simple alteration: an additional marker at the head of the participant to automatically detect situations where one participant looks at the other and especially mutual eye contact between the participants is established. Together with a recording of the pitch, and the head movement recording, this would further assist to investigate the timing and structure of turn-taking in interaction.

Concerning the automatic tagging methods, the head gesture recognition should be improved by using trained coders or redundant coding (more than one coder) for the creation of the training corpus. Additionally, the head gesture recognition could be combined with speech information (e. g. speech pauses or automatic intonation analysis for questions) and mutual gaze (elicitations of back-channels by the speaker looking at the listener) for improvements. Further, ideas for automatic tagging are to include a hand gesture recognition, e. g. (Heidemann et al., 2004) or using an acoustic packaging approach (Schillingmann et al., 2009) in order to structure other actions (e. g. gestures, body movements) into action sequences. This would further reduce the amount of manual work for the researcher and enhance the resulting corpora thereby offering more structured multimodal information for the analysis of complex relationships between the signal types.

For the data processing, this thesis provided a modular conversion approach that synchronizes and transforms all data into rich and multimodal ELAN file bundles and provides a configuration file for each bundle. Future projects could easily use this conversion infrastructure since it is applicable to different scenarios and research questions and has already been used in human-robot interaction. Nevertheless, there are some alterations that would improve the analysis of interaction further. Since annotations are temporal data and temporal data is particularly suited to be perceived via the auditory display, the conversion would benefit from an interface that allows perceiving the annotations not only visually but also acoustically.

Another idea is to implement a *replay* function that allows to re-experience the experiment with all system states as a video-audio stream from the participant's point of view. This would allow the researcher to put himself/herself in the position of the participant experiencing the same audiovisual stimuli.

For the analysis toolbox, an integration of sound functionality would be reasonable: it would be nice if the sound (loudness, pitch) could be analysed in parallel with the timeseries data and the annotations. This would allow facilitating the analysis of the correlation of pitch with head movements and mutual eye contact to a large extent. Additionally, there have been already requests for an .eaf file export function by external users.

In the questionnaire and in their comments, the participants rated the interactive exhibition design scenario particularly interesting and the task was shown to be suited for the investigation of nonverbal behaviour both under AR and under non-AR conditions. Improvements for the AR conditions could augment the current characteristics for the respective room configuration in terms of the different methods of interference (e. g. sound, lighting, smell). This should increase the task performance in the AR condition compared to a non-AR condition where the same information could be available on a sheet of paper.

There are also future research questions that are particularly interesting to be considered with the help of the infrastructure presented in this thesis. For example, the correlation of head gestures with speech and eye gaze can be investigated as well as the synchrony of head gestures with action events (e. g. gestures, manipulations) of participants.

Furthermore, the prospects of Conversation Analysis (Sacks et al., 1974) have only started to be used in spite of its promising techniques (see Section 5.3). These techniques are particularly suited to be combined with the quantitative analysis infrastructure provided in this thesis thereby bridging the gap between quantitative and qualitative analyses. For example, the ongoing effort is to analyse the detailed ways in which participants establish, sustain and manipulate joint attention and to investigate the sequential organisation of their actions as proposed in Hermann and Pitsch (2009).

Concerning the features of the closed-loop interface ARbInI, a wise advancement for the future would be to use the sonification not as an activity monitor but for entirely different information: the room loudness in the exhibition design scenario. In our studies, it was evident that it was difficult for the participants to estimate the room loudness in order to decide whether a further exhibit would overburden the room or not. With such a sonification, however, the soundscape of a room could be represented and the participants could more easily come to their decision. An additional idea would be to map the tracked head gestures onto a sonification thereby providing the visual focus of attention visually and the performed head gestures by sound. For the visual channel however, it would most likely be even more helpful for the users to highlight the space on the table that is in the partner's field of view. This was already proposed and implemented by Mertes (2008) but has not been used in studies so far.

The ongoing system improvements aim at enhancing the speed and reliability of the marker tracking in ARbInI (Neumann, 2011). Additionally, Section 5.4 identified several issues for the hardware that should be corrected (e. g. the load relief on the root of the nose). Such modifications would further improve the illusion of reality for the users of the system and thus especially allow for the investigation of misunderstandings and misguidances. This topic has hardly been touched in this thesis, notwithstanding its potentials in interaction research.

A. Appendix

A.1. Interactive exhibition design study

A.1.1. Questionnaire AR group

German	Translation	Answers
Bitte beschreiben Sie kurz in eigenen Worten Ihren Eindruck beim Benutzen des Systems.	Please describe shortly in your own words your impression during the use of the system	(open question)
Kannten Sie Ihren Interaktionspartner schon vor dem Experiment?	Did you know your interaction partner prior to the experiment?	Yes – No
Trugen Sie im Experiment eine (normale) Brille unter dem Augmented-Reality-Gerät?	Did you wear (normal) glasses in combination with the Augmented Reality goggles during the experiment?	Yes – No
Hatten Sie vor diesem Versuch schon einmal ein Augmented-Reality-System benutzt?	Have you ever used an Augmented Reality system prior to the experiment?	Yes – No
Wie beurteilen Sie den Tragekomfort des Systems?	How do you rate the wearing comfort of the system?	Very comfortable... very uncomfortable
Wie angenehm fanden Sie die Nutzung des Systems?	How do you rate the convenience of the system?	Very convenient ... very inconvenient
Falls Sie das Tragen oder die Nutzung des Systems unkomfortabel fanden: Wie hat sich der fehlende Komfort geäußert?	If you found the usage of the system uncomfortable: how did this lack of comfort show?	(open question)
Wie schnell haben Sie sich an das System gewöhnt?	How fast did you get used to the system?	very fast ... very slow not at all
Wie stark hat Sie das Vorhandensein der im Raum befindlichen Video-Kameras in der Interaktion mit Ihrem Partner beeinflusst?	How much did the video cameras stationed in the room influence your interaction with your partner?	very much ... very little not at all

Wie stark hat die an Ihrem Kopf angebrachte Technik Sie bei Ihren Handlungen beeinträchtigt? (die Augmented-Reality-Brille, das Mikrophon, der blaue Sensor am Kopf)	How much did the devices mounted to your head interfere with your actions? (the Augmented Reality goggles, the microphone, the blue sensor on your head)	Very much, very little not at all
--	--	-------------------------------------

Page 2: Questions concerning the task “museum planning”

Wie leicht fanden Sie die gestellte Aufgabe?	How simple did you find the task?	very simple... very difficult
Bei der Museumsplanung konnten sich die dargestellten Experimente gegenseitig auf verschiedene Arten stören. Welche „Störungs-Arten“ konnten Sie selbst identifizieren als Sie allein das Museum planten?	During the museum design, the experiment were able to inhibit each other mutually. Which ways of disturbance did you identify during the individual phase?	(open question)
Welche weiteren „Störungs-Arten“ sind Ihnen durch die Zusammenarbeit mit Ihrem Partner noch aufgefallen?	Which additional ways of disturbance did you identify together with your partner?	(open question)
Wie natürlich schätzen Sie Ihre Kopfbewegungen in den Versuchsteilen ein? (im Eingewöhnungs-Gespräch am Anfang, in der Einzelaufgabe, in der Gemeinschaftsaufgabe)	How natural do you rate your head movements during the phases of the experiment (in the smalltalk phase, in the individual phase, during the group phase)?	very natural ... very unnatural
Wie hilfreich fanden Sie die (farblichen) visuellen Hervorhebungen?	How helpful did you find the (coloured) visual highlighting?	very helpful ... not helpful
Wie stark haben Sie die visuellen Hervorhebungen genutzt?	How much did you use these visual highlightings?	very much... very few not at all
Was hat Ihnen besonders gut gefallen? Was hat Ihnen besonders schlecht gefallen? Was würden Sie ändern?	What did you like especially? What did you dislike especially? What would you change?	(open question)
Welche (anderen) Anwendungen könnten Sie sich für ein solches System (z.B. im Alltag) vorstellen	Which (other) applications can you imagine for such a system (e. g. during everyday life)?	(open question)

Thank you for your participation!

Table A.1.: Translations for the exhibition design study questionnaire – AR group.

Fragebogen zur Museumsplanung (AR)

Geschlecht _____ Alter _____ Fachbereich/Beruf _____

Bitte beschreiben Sie kurz in eigenen Worten Ihren Eindruck beim Benutzen des Systems.

Kannten Sie Ihren Interaktionspartner schon vor dem Experiment?	Ja <input type="radio"/>	Nein <input type="radio"/>			
Trugen Sie im Experiment eine (normale) Brille unter dem Augmented-Reality-Gerät?	Ja <input type="radio"/>	Nein <input type="radio"/>			
Hatten Sie vor diesem Versuch schon einmal ein Augmented-Reality-System benutzt?	Ja <input type="radio"/>	Nein <input type="radio"/>			
Wie beurteilen Sie den Tragekomfort des Systems?	Sehr komfortabel <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr unkomfortabel <input type="radio"/>	
Wie angenehm fanden Sie die Nutzung des Systems?	Sehr angenehm <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr unangenehm <input type="radio"/>	
Falls Sie das Tragen oder die Nutzung des Systems unkomfortabel fanden: Wie hat sich der fehlende Komfort geäußert?					
Wie schnell haben Sie sich an das System gewöhnt?	Sehr schnell <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr langsam <input type="radio"/>	Gar nicht <input type="radio"/>
Wie stark hat Sie das Vorhandensein der im Raum befindlichen Video-Kameras in der Interaktion mit Ihrem Partner beeinflusst?	Sehr stark <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr wenig <input type="radio"/>	Gar nicht <input type="radio"/>
Wie stark hat die an Ihrem Kopf angebrachte Technik Sie bei Ihren Handlungen beeinträchtigt?	Sehr stark <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr wenig <input type="radio"/>	Gar nicht <input type="radio"/>
- die Augmented-Reality-Brille	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
- das Mikrophon	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
- der blaue Sensor am Kopf	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fragen zur Aufgabe „Museumsplanung“

	Sehr leicht			Sehr schwer	
Wie leicht fanden Sie die gestellte Aufgabe?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Bei der Museumsplanung konnten sich die dargestellten Experimente gegenseitig auf verschiedene Arten stören. Welche „Störungs-Arten“ konnten Sie selbst identifizieren als Sie allein das Museum planten?					
Welche weiteren „Störungs-Arten“ sind Ihnen durch die Zusammenarbeit mit Ihrem Partner noch aufgefallen?					
Wie natürlich schätzen Sie Ihre Kopfbewegungen in den Versuchsteilen ein?	Sehr natürlich			Sehr unnatürlich	
- im Eingewöhnungs-Gespräch am Anfang	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
- in der Einzelaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
- in der Gemeinschaftsaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Wie hilfreich fanden Sie die (farblichen) visuellen Hervorhebungen?	Sehr hilfreich			Gar nicht hilfreich	Störend
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Wie stark haben Sie die visuellen Hervorhebungen genutzt?	Sehr stark			Sehr wenig	Gar nicht
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Was hat Ihnen besonders gut gefallen (+)?

Was hat Ihnen besonders schlecht gefallen (-)?

Was würden Sie ändern (ä)?

Welche (anderen) Anwendungen könnten Sie sich für ein solches System (z.B. im Alltag) vorstellen?

Vielen Dank für Ihre Teilnahme!

A.1.2. Questionnaire non-AR group

German	Translation	Answers
Bitte beschreiben Sie kurz in eigenen Worten Ihren Eindruck beim Benutzen des Systems.	Please describe shortly in your own words your impression during the use of the system	(open question)
Kannten Sie Ihren Interaktionspartner schon vor dem Experiment?	Did you know your interaction partner prior to the experiment?	Yes – No
Trugen Sie im Experiment eine Brille?	Did you wear glasses during the experiment?	Yes – No
Hatten Sie vor diesem Versuch schon einmal (oder mehrfach) an einer Studie teilgenommen, in der Ihre Kopfbewegungen gemessen wurden?	Have you ever participated in a study where your head movements were recorded prior to this experiment?	Yes – No
Wie beurteilen Sie den Tragekomfort der Sensoren?	How do you rate the wearing comfort of the sensors?	Very comfortable... very uncomfortable
Falls Sie das Tragen oder die Nutzung des Systems unkomfortabel fanden: Wie hat sich der fehlende Komfort geäußert?	If you found the usage of the system uncomfortable: how did this lack of comfort show?	(open question)
Wie schnell haben Sie sich an das System gewöhnt?	How fast did you get used to the system?	very fast ... very slow not at all
Wie stark hat Sie das Vorhandensein der im Raum befindlichen Video-Kameras Ihr Verhalten während des Experiments beeinflusst?	How much did the video cameras stationed in the room influence your interaction with your partner?	very much ... very little not at all
Wie stark hat die an Ihrem Kopf angebrachte Technik Sie bei Ihren Handlungen beeinträchtigt? (der blaue Sensor am Kopf, das Mikrophon)	How much did the devices mounted to your head interfere with your actions? (the blue sensor on your head, the microphone)	Very much, very little not at all

Page 2: Questions concerning the task “museum planning”

Wie leicht fanden Sie die gestellte Aufgabe?	How simple did you find the task?	very simple... very difficult
Bei der Museumsplanung konnten sich die dargestellten Experimente gegenseitig auf verschiedene Arten stören. Welche „Störungs-Arten“ konnten Sie selbst identifizieren als Sie allein das Museum planten?	During the museum design, the experiment were able to inhibit each other mutually. Which ways of disturbance did you identify during the individual phase?	(open question)

Welche weiteren „Störungs-Arten“ sind Ihnen durch die Zusammenarbeit mit Ihrem Partner noch aufgefallen?	Which additional ways of disturbance did you identify together with your partner?	(open question)
Wie natürlich schätzen Sie Ihre Kopfbewegungen in den Versuchsteilen ein? (im Eingewöhnungs-Gespräch am Anfang, in der Einzelaufgabe, in der Gemeinschaftsaufgabe)	How natural do you rate your head movements during the phases of the experiment (in the smalltalk phase, in the individual phase, during the group phase)?	very natural ... very unnatural
Was hat Ihnen besonders gut gefallen? Was hat Ihnen besonders schlecht gefallen? Was würden Sie ändern?	What did you like especially? What did you dislike especially? What would you change?	(open question)
Welche Anwendungen könnten Sie sich für ein solches Sensor-System (z.B. im Alltag) vorstellen	Which applications can you imagine for such a sensor system (e. g. during everyday life)?	(open question)

Thank you for your participation!

Table A.2.: Translations for the exhibition design study questionnaire – non-AR group.

Fragebogen zur Museumsplanung (Kopfbewegungssensor)

Geschlecht _____ Alter _____ Fachbereich/Beruf _____

Bitte beschreiben Sie kurz in eigenen Worten Ihren Eindruck beim Benutzen des Systems.

Fragen zu den Sensoren

Kannten Sie Ihren Interaktionspartner schon vor dem Experiment?	Ja <input type="radio"/>	Nein <input type="radio"/>			
Trugen Sie im Experiment eine Brille?	Ja <input type="radio"/>	Nein <input type="radio"/>			
Hatten Sie vor diesem Versuch schon einmal (oder mehrfach) an einer Studie teilgenommen, in der Ihre Kopfbewegungen gemessen wurden?	Ja <input type="radio"/>	Nein <input type="radio"/>			
Wie beurteilen Sie den Tragekomfort der Sensoren?	Sehr komfortabel <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr unkomfortabel <input type="radio"/>	<input type="radio"/>
Falls Sie das Tragen oder die Nutzung des Systems unkomfortabel fanden: Wie hat sich der fehlende Komfort geäußert?					
Wie stark hat Sie das Vorhandensein der Video-Kameras ihr Verhalten während des Experiments beeinflusst?	Sehr stark <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr wenig <input type="radio"/>	Gar nicht <input type="radio"/>
Wie stark hat die an Ihrem Kopf angebrachte Technik ihr Verhalten während des Experiments beeinflusst?	Sehr stark			Sehr wenig	Gar nicht
- der blaue Sensor am Kopf	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
- das Mikrophon	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fragen zur Aufgabe „Museumsplanung“

	Sehr leicht			Sehr schwer
Wie leicht fanden Sie die gestellte Aufgabe?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bei der Museumsplanung konnten sich die dargestellten Experimente gegenseitig auf verschiedene Arten stören. Welche „Störungs-Arten“ konnten Sie selbst identifizieren als Sie allein das Museum planten?				
Welche weiteren „Störungs-Arten“ sind Ihnen durch die Zusammenarbeit mit Ihrem Partner noch aufgefallen?				
Wie natürlich schätzen Sie Ihre Kopfbewegungen in den drei Versuchsteilen ein?	Sehr natürlich			Sehr unnatürlich
- im Eingewöhnungs-Gespräch am Anfang	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
- in der Einzelaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
- in der Gemeinschaftsaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Was hat Ihnen besonders gut gefallen (+)?
 Was hat Ihnen besonders schlecht gefallen (-)?
 Was würden Sie ändern (ä)?

Welche Anwendungen könnten Sie sich für ein solches Sensor-System (z.B. im Alltag) vorstellen?

Vielen Dank für Ihre Teilnahme!

A.1.3. Mapping of markers to interactive exhibits



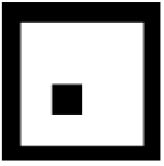

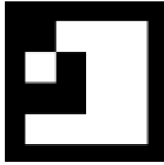

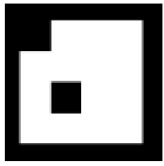

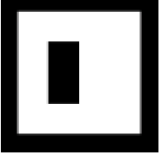
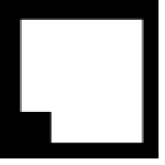

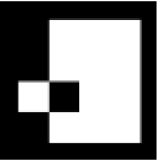




1	Extinguish candle by drumbeat (Löschen einer Kerze durch Paukenschlag)		9	Steadfast candle (Luftfluss um Hindernisse)	
2	Humming Stone (Summstein)		10	Color Mixtures (Farbmischungen)	
3	Optical Illusion: Swivel disk (Optische Scheiben)		11	Optical Illusion: Triangle in a House (Optische Täuschung:	
4	Water-sound dabbling bowl (Wasserklang-spritzschale)		12	Feel around in the dark (Lichtloses Tasten)	
5	Lasershow (Lasershow)		13	Huge soap bubbles (Riesige Seifenblasen)s	
6	Soundfigures of sand (Klangfiguren aus Sand)		14	Plasmadisk - electric discharges (Plasmascheibe - elektrische Entladungen)	
7	Listening (Lauschen)		15	Wind engine (Windmaschine)	
8	Optical Illusion: Arrows (Optische Täuschung: Pfeile)		16	Smelling Tree (Riechbaum)	

Figure A.1.: Mapping of displayed experiments to a set of ARToolKit markers

A.2. Enhancing/Supporting: A multimodal display for the focus of attention

German	Translation	Answers
Bitte beschreiben Sie kurz in eigenen Worten Ihren Eindruck beim Benutzen des Systems.	Please describe shortly in your own words your impression of the use of the system	(open question)
Tragen Sie eine Brille?	Do you wear glasses?	Yes – No
Hatten Sie vor diesem Versuch schon einmal ein Augmented-Reality-System benutzt?	Have you ever used an Augmented Reality system previous to the experiment?	Yes – No
Wie beurteilen Sie den Tragekomfort des Systems?	How do you rate the wearing comfort of the system?	Very comfortable... very uncomfortable
Wie angenehm fanden Sie die Nutzung des Systems?	How convenient do you rate the usage of the system?	Very convenient ... very inconvenient
Falls Sie das Tragen oder die Nutzung des Systems unkomfortabel fanden: Wie hat sich der fehlende Komfort geäußert?	If you found the usage of the system uncomfortable: how did this lack of comfort show?	(open question)
Wie schnell haben Sie sich an das System gewöhnt?	How fast did you get used to the system?	very fast ... very slow not at all
Wie stark haben Sie die visuellen Hervorhebungen genutzt?	How much did you use the visual highlighting?	very much ... not at all
Wie stark haben Sie die auditiven Hervorhebungen genutzt?	How much did you use the auditory highlighting?	very much ... not at all
Wie hilfreich fanden Sie die visuellen Hervorhebungen?	How helpful did you find the visual highlighting?	very helpful ... not helpful distracting
Wie hilfreich fanden Sie die auditory Hervorhebungen?	How helpful did you find the auditory highlighting?	very helpful ... not helpful distracting
Was hat Ihnen besonders gut gefallen? Was hat Ihnen besonders schlecht gefallen? Was würden Sie ändern?	What did you like especially? What did you dislike especially? What would you change?	(open question)
Welche Anwendungen könnten Sie sich für ein solches System (z.B. im Alltag) vorstellen	Which applications can you imagine for such a system (e. g. during everyday life)?	(open question)

Thank you for your participation!

Table A.3.: Translations for the gaze game study questionnaire.

Fragebogen zur Anzeige des Aufmerksamkeitsfokus 2

Geschlecht _____

Alter _____

Bitte beschreiben Sie kurz in eigenen Worten Ihren Eindruck beim Benutzen des Systems.

	Ja	Nein		
Tragen Sie eine Brille?	<input type="radio"/>	<input type="radio"/>		
Hatten Sie vor diesem Versuch schon einmal ein Augmented-Reality-System benutzt?	Ja <input type="radio"/>	Nein <input type="radio"/>	Weiß nicht <input type="radio"/>	
Wie beurteilen Sie den Tragekomfort des Systems?	Sehr komfortabel <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr unkomfortabel <input type="radio"/>
Wie angenehm fanden Sie die Nutzung des Systems?	Sehr angenehm <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sehr unangenehm <input type="radio"/>
Falls Sie das Tragen oder die Nutzung des Systems unkomfortabel fanden: Wie hat sich der fehlende Komfort geäußert?				
Wie schnell haben Sie sich an das System gewöhnt?	Sehr schnell <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Gar nicht <input type="radio"/>
Wie stark haben Sie die visuellen Hervorhebungen genutzt?	Sehr stark <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Gar nicht <input type="radio"/>
Wie stark haben Sie die auditiven Hervorhebungen genutzt?	Sehr stark <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Gar nicht <input type="radio"/>
Wie hilfreich fanden Sie die visuellen Hervorhebungen?	Sehr hilfreich <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Gar nicht hilfreich Störend <input type="radio"/>
Wie hilfreich fanden Sie die auditiven Hervorhebungen?	Sehr hilfreich <input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Gar nicht hilfreich Störend <input type="radio"/>

Was hat Ihnen besonders gut, was besonders schlecht gefallen? Was würden Sie ändern?

Welche Anwendungen könnten Sie sich für ein solches System (z.B. im Alltag) vorstellen?

Vielen Dank für Ihre Teilnahme!

Bibliography

- Allwood, J. and Cerrato, L. (2003). A study of gestural feedback expressions. In *Proc. of the First Nordic Symposium on Multi-modal Communication. Copenhagen*, pages 7–20.
- Alves Fernandes, B. and Fernández Sánchez, J. (2008). Acceptance of an augmented reality system as a visualization tool for computer-aided design classes. *Digital Education Review*, 16:66–86.
- Argyle, M. (1975). *Bodily communication*. Methuen, London.
- Argyle, M. (1988). *Bodily communication*. Methuen.
- Argyle, M., Lefebvre, L., and Cook, M. (1974). The meaning of five patterns of gaze. *European journal of social psychology*, 4(2):125–136.
- Arthur, K. (2000). *Effects of field of view on performance with head-mounted displays*. Dissertation, University of North Carolina.
- Azuma, R., Baillet, Y., Behringer, R., Feiner, S., Julier, S., and MacIntyre, B. (2001). Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, pages 34–47.
- Azuma, R. et al. (1997). A survey of augmented reality. *Presence-Teleoperators and Virtual Environments*, 6(4):355–385.
- Baars, B. J. H. (2007). *Cognition, brain, and consciousness: introduction to cognitive neuroscience*. Elsevier Acad. Press, Amsterdam [u.a.].
- Baier, G. (2001). *Rhythmus – Tanz in Körper und Gehirn*. Rowohlt Taschenbuch Verlag, Reinbeck.
- Baird, K. and Barfield, W. (1999). Evaluating the effectiveness of augmented reality displays for a manual assembly task. *Virtual Reality*, 4(4):250–259.
- Barkhuysen, P. (2008). *Audiovisual prosody in interaction*. PhD thesis, Universiteit van Tilburg.
- Barrett, G. and Thornton, C. (1968). Relationship between perceptual style and simulator sickness. *Journal of Applied Psychology*, 52(4):304–308.
- Basapur, S., Xu, S., Ahlenius, M., and Lee, Y. (2007). User expectations from dictation on mobile devices. *Human-Computer Interaction. Interaction Platforms and Techniques*, pages 217–225.
- Bavelas, J., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

- Bavelas, J., Coates, L., and Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580.
- Bazzanella, C. and Damiano, R. (1999). The interactional handling of misunderstanding in everyday conversations. *Journal of Pragmatics*, 31(6):817–836.
- Bear, J., Dowding, J., and Shriberg, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics.
- Biber, D. (1991). *Variation across speech and writing*. Cambridge Univ Pr.
- Billinghamurst, M. and Kato, H. (2002). Collaborative augmented reality. *Communications of the ACM*, 45(7):64–70.
- Birdwhistell, R. (1970). *Kinesics and context: Essays on body motion communication*. Univ of Pennsylvania Pr.
- Bonaiuto, J. and Thórisson, K. (2008). Towards a neurocognitive model of turn taking in multimodal dialog. *Embodied communication in humans and machines*, page 451.
- Bovermann, T., Hermann, T., and Ritter, H. (2006). Tangible data scanning sonification model. In Stockman, T., editor, *Proceedings of the International Conference on Auditory Display (ICAD 2006)*, pages 77–82, London, UK. International Community for Auditory Display (ICAD), Department of Computer Science, Queen Mary, University of London.
- Bovermann, T., Koiva, R., Hermann, T., and Ritter, H. (2008). TUImod: Modular objects for tangible user interfaces. In *Proceedings of the 2008 Conference on Pervasive Computing*.
- Brooks, R. and Meltzoff, A. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543.
- Bui, T., Heylen, D., and Nijholt, A. (2004). Combination of facial movements on a 3d talking head. *Computer Graphics International, 2004. Proceedings*, pages 284–290.
- Bull, P. and Connelly, G. (1985). Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3):169–187.
- Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM New York, NY, USA.
- Caudell, T. and Mizell, D. (1992). Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, volume 2, pages 659–669. IEEE.
- Cerrato, L. and Skhiri, M. (2003). Analysis and measurement of communicative gestures in human dialogues. *Proc. of AVSP 2003*, pages 251–256.

- Chen, L., Harper, M., and Quek, F. (2002). Gesture patterns during speech repairs. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 155–160. IEEE.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*, volume 119. The MIT press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.
- Cook, M. (1977). Gaze and mutual gaze in social encounters: How long—and when—we look others" in the eye" is one of the main signals in nonverbal communication. *American Scientist*, 65(3):328–333.
- Crystal, D. (2008). *A dictionary of linguistics and phonetics*. Blackwell, Malden, Mass.
- Darwin, C., Ekman, P., and Prodger, P. (2002). *The expression of the emotions in man and animals*. Oxford University Press, USA.
- de Kok, I. and Heylen, D. (2009). Multimodal end-of-turn prediction in multi-party meetings. In *ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces*, pages 91–97, New York, NY, USA. ACM.
- de Vries, S. and Padmos, P. (1997). Steering a simulated unmanned aerial vehicle using a head-slaved camera and HMD. In *Proceedings of SPIE*, volume 3058, page 24.
- DeCarlo, D., Stone, M., Revilla, C., and Venditti, J. (2004). Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38.
- DePaulo, B., Lanier, K., and Davis, T. (1983). Detecting the deceit of the motivated liar. *Journal of Personality and Social Psychology*, 45(5):1096.
- Dierker, A., Bovermann, T., Hanheide, M., Hermann, T., and Sagerer, G. (2009a). A multimodal augmented reality system for alignment research. In *International Conference on Human-Computer Interaction*, pages 422–426, San Diego, USA.
- Dierker, A., Mertes, C., Hermann, T., Hanheide, M., and Sagerer, G. (2009b). Mediated attention with multimodal augmented reality. In *ICMI-MLMI '09: Proceedings of the 2009 international conference on multimodal interfaces*, pages 245–252, New York, NY, USA. ACM.
- Dierker, A., Pitsch, K., and Hermann, T. (2011). An augmented-reality-based scenario for the collaborative construction of an interactive museum. Technical report, Bielefeld University.
- Dittmann, A. and Llewellyn, L. (1968). Relationship between vocalizations and head nods as listener responses. *Journal of personality and social psychology*, 9(1):79.
- Doherty-Sneddon, G., Anderson, A., O222Malley, C., Langton, S. ., Garrod, S., and Bruce, V. (1997). Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, 3(2):105–125.

- Dolezal, H. (1982). *Living in a world transformed: Perceptual and performatory adaptation to visual distortion*. Academic Press.
- Drascic, D. and Milgram, P. (1996). Perceptual issues in augmented reality. In *Proceedings of SPIE - the international society for optical engineering*, pages 123–134. Citeseer.
- Drew, P. (1997). [] open-class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, 28(1):69–101.
- DuFon, M. (2002). Video recording in ethnographic sla research: Some issues of validity in data collection. *Language Learning & Technology*, 6(1):40–59.
- Duncan Jr., S. (1969). Nonverbal communication. *Psychological Bulletin*, 72(2):118.
- Duncan Jr., S. and Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10(3):234–247.
- Egeth, H. and Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual review of psychology*, 48(1):269–297.
- Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.
- Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W., Pitcairn, T., Ricci-Bitti, P., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712.
- Ekman, P., O'Sullivan, M., Friesen, W., and Scherer, K. (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, 15(2):125–135.
- Ellis, S., Breant, F., Manges, B., Jacoby, R., and Adelstein, B. (1997). Factors influencing operator interaction with virtual objects viewed via head-mounted see-through displays: viewing conditions and rendering latency. In *Proc., Virtual Reality Ann. Internat. Symp. (VRAIS 97)*, pages 138 – 145. Published by the IEEE Computer Society.
- Evans, D. (2003). *Emotion – a very short introduction*. Very short introductions ; 81. Oxford Univ. Press.
- Feiner, S., Macintyre, B., and Seligmann, D. (1993). Knowledge-based augmented reality. *Commun. ACM*, 36(7):53–62.
- Finch, G. (2002). *Linguistic terms and concepts*. Palgrave, Basingstoke.
- Fink, G. A. (1999). Developing HMM-Based Recognizers with ESMERALDA. In *TSD '99: Proceedings of the Second International Workshop on Text, Speech and Dialogue*, pages 229–234, London, UK. Springer-Verlag. Available from: <http://portal.acm.org/citation.cfm?id=647237.720414>.
- Frischen, A., Bayliss, A., and Tipper, S. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694.

- Frohlich, D., Drew, P., and Monk, A. (1994). Management of repair in human-computer interaction. *Human-Computer Interaction*, 9(3-4):385–425.
- Furchner, I. (2009). Gespräche im alltag - alltag im gespräch: Die konversationsanalyse. In Müller, H. M., editor, *Arbeitsbuch Linguistik – eine Einführung in die Sprachwissenschaft*, UTB ; 2169 : Sprachwissenschaft, pages 532 S. : Ill., graph. Darst. Schöningh.
- Goffman, E. (1959). The presentation of self in everyday life. *Anchor*.
- Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological Inquiry*, 50(3-4):272–302.
- Goodwin, C. (1987). Forgetfulness as an interactive resource. *Social Psychology Quarterly*, pages 115–130.
- Goto, M., Itou, K., and Hayamizu, S. (1999). A real-time filled pause detection system for spontaneous speech recognition. In *Sixth European Conference on Speech Communication and Technology*.
- Graf, H., Cosatto, E., Strom, V., and Huang, F. (2002). Visual prosody: Facial movements accompanying speech. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 396–401. IEEE.
- Großekathöfer, U. and Lingner, T. (2005). Neue Ansätze zum maschinellen Lernen von Alignments. Master's thesis, Bielefeld University.
- Hadar, U., Steiner, T., and Clifford Rose, F. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228.
- Hadar, U., Steiner, T., Grant, E., and Rose, F. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2):35–46.
- Hanheide, M. (2006). *A Cognitive Ego-Vision System for Interactive Assistance*. phdthesis, Technische Fakultät – Universität Bielefeld. Available from: <http://bieson.uni-bielefeld.de/volltexte/2007/1032/>.
- Hanheide, M., Lohse, M., and Dierker, A. (2010). SALEM – Statistical Analysis of Elan files in Matlab. In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pages 121–123, Malta.
- Healey, P. and Thirlwell, M. (2002). Analysing multi-modal communication: Repair-based measures of communicative co-ordination. *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, page 83.
- Heeman, P. and Allen, J. (1999). Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.
- Heidemann, G., Bekel, H., Bax, I., and Saalbach, A. (2004). Hand gesture recognition: Self-organising maps as a graphical user interface for the partitioning of large training data sets. In J. Kittler, M. P. and Nixon, M., editors, *Proc. ICPR 2004*, volume 4, pages 487–490, Cambridge, UK. IEEE CS-Press.

- Heinrich, M., Fleer, D., Liguda, C., Mödecker, L., and Mrugalla, A. (2010). Intelligent systems project: AR-based applications. Technical report, Faculty of Technology, Bielefeld University.
- Hellwig, B. and Uytvanck, D. (2004). EUDICO Linguistic Annotator (ELAN) Version 2.0. 2 manual. *Nijmegen-NL, Max Planck Institute for Psycholinguistics*.
- Hermann, T. (2002). *Sonification for Exploratory Data Analysis*. PhD thesis, Bielefeld University, Bielefeld, Germany.
- Hermann, T. (2008). Taxonomy and definitions for sonification and auditory display. In Katz, B., editor, *Proc. 14th Int. Conf. Auditory Display (ICAD 2008)*, pages 1–8, Paris, France.
- Hermann, T., Neuhoff, J., and Hunt, A., editors (2011). *The Sonification Handbook*. Logos Verlag, Berlin, Germany.
- Hermann, T. and Pitsch, K. (2009). C5: Alignment in AR-based cooperation. In *Funding Proposal for the 2nd period of the Collaborative Research Centre SFB 673 "Alignment in Communication"*, pages 357–378. Bielefeld University.
- Hermann, T. and Sagerer, G. (2005). C5: Alignment in AR-based cooperation. In *Funding Proposal for the Collaborative Research Centre SFB 673 "Alignment in Communication"*, pages 361–376. Bielefeld University.
- Heylen, D. (2005). Challenges ahead. head movements and other social acts in conversation. *Virtual Social Agents*.
- Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(3):1–27.
- Heylen, D., Bevacqua, E., Tellier, M., and Pelachaud, C. (2007). Searching for prototypical facial feedback signals. In *Intelligent Virtual Agents*, pages 147–153. Springer.
- Holloway, R. (1997). Registration error analysis for augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):413–432.
- Hood, B., Willen, J., and Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, 9(2):131.
- Jaimes, A. and Sebe, N. (2007). Multimodal human computer interaction: A survey. *Computer vision in human-computer interaction*, pages 1–15.
- Jefferson, G. (1974). Error correction as an interactional resource. *Language in society*, 2(1974):181–199.
- Kalkofen, D., Mendez, E., and Schmalstieg, D. (2007). Interactive focus and context visualization for augmented reality. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–10. IEEE Computer Society.
- Kandel, E. R. H. (1996). *Neurowissenschaften – eine Einführung*. Spektrum, Akad. Verl., Heidelberg [u.a.].

- Kaplan, F. and Hafner, V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2):135–169.
- Kato, H. and Billinghurst, M. (1999). Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*, volume 99, pages 85–94. San Francisco, CA.
- Kendon, A. (2002). Some uses of the head shake. *Gesture*, 2(2):147–182.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge Univ Pr.
- Kendon, A. and Cook, M. (1969). The consistency of gaze patterns in social interaction. *British Journal of Psychology*, 60(4):481–494.
- Kim, K., Lee, S., and Kim, H. (2004). A wireless measurement system for three-dimensional ocular movement using the magnetic contact lens sensing technique. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 1, pages 2287–2289. IEEE.
- Kiyokawa, K., Billinghurst, M., Hayes, S., Gupta, A., Sannohe, Y., and Kato, H. (2002). Communication behaviors of co-located users in collaborative ar interfaces. In *Proceedings International Symposium on Mixed and Augmented Reality*, pages 139–148. IEEE Computer Society.
- Kleinke, C. (1986). Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78.
- Knapp, M. and Hall, J. (2009). *Nonverbal communication in human interaction*. Wadsworth Pub Co.
- Knapp, M. L. (1978). *Nonverbal communication in human interaction*. Holt, Rinehart and Winston, New York [u.a.].
- Knight, J. and Baber, C. (2005). A tool to assess the comfort of wearable computers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 47(1):77.
- Koesling, H. (2003). *Visual Perception of Location, Orientation and Length: An Eye-Movement Approach*. PhD thesis, Bielefeld University.
- Kollenberg, T., Neumann, A., Schneider, D., Tews, T.-K., Dierker, A., and Koesling, H. (2009). The influence of head-mounted displays on head and eye movements during visual search. Technical report, Faculty of Technology, Bielefeld University, Bielefeld, Germany.
- Kollenberg, T., Neumann, A., Schneider, D., Tews, T.-K., Hermann, T., Ritter, H., Dierker, A., and Koesling, H. (2010). Visual search in the (un)real world: How head-mounted displays affect eye movements, head movements and target detection. In C.H. Morimoto, I. H. E., editor, *Symposium on Eye Tracking Research & Applications*, pages 121–124, New York, NY, USA. ACM.

- Kollin, J. (1993). A retinal display for virtual-environment applications. In *SID International Symposium Digest of Technical Papers*, volume 24, pages 827–827. Society for information display.
- Kramer, G. H. (1994). *Auditory display – sonification, audification, and auditory interfaces – [International Conference on Auditory Display, Santa Fe Institute, October 28 - 30, 1992]*, volume 18 of *Santa Fe Institute studies in the sciences of complexity*. Addison-Wesley.
- Krause, D. (2005). *Luhmann-Lexikon – eine Einführung in das Gesamtwerk von Niklas Luhmann mit über 600 Lexikoneinträgen einschließlich detaillierter Quellenangaben*. Soziologie fachübergreifend. Lucius & Lucius.
- Krauss, R., Garlock, C., Bricker, P., and McMahon, L. (1977). The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, 35(7):523.
- Kurzweil, R. (1990). *The age of intelligent machines*. MIT Pr., Cambridge, Mass. [u.a.].
- Lange, E. (2005). Issues and Questions for Research in Communicating with the Public through Visualizations. *Trends in Real-Time Landscape Visualization and Participation: Proceedings at Anhalt University of Applied Sciences, Dessau, Germany, May*, pages 26–28.
- Langton, S., Watt, R., and Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59.
- Lee, J. (2008). Hacking the nintendo wii remote. *Pervasive Computing, IEEE*, 7(3):39–45.
- Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *Intelligent Virtual Agents*, pages 243–255. Springer.
- Leech, G. (1993). Corpus annotation schemes. *Literary and linguistic computing*, 8(4):275–281.
- Levelt, W. and Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics*, 2(2):205–218.
- Lingley, A. R., Ali, M., Liao, Y., Mirjalili, R., Klonner, M., Sapanen, M., Suihkonen, S., Shen, T., Otis, B. P., Lipsanen, H., and Parviz, B. A. (2011). A single-pixel wireless contact lens display. *Journal of Micromechanics and Microengineering*, 21(12):125014. Available from: <http://stacks.iop.org/0960-1317/21/i=12/a=125014>.
- Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1526–1540.
- Lohse, M. (2010). *Investigating the influence of situations and expectations on user behavior: empirical analyses in human-robot interaction*. PhD thesis, Doctoral Thesis. Bielefeld University, Technical Faculty, Germany.
- Luhmann, N. (1984). *Soziale Systeme – Grundriß einer allgemeinen Theorie*. Suhrkamp, Frankfurt am Main.

- Martin, J. (1970). Suspicion and the experimental confederate: A study of role and credibility. *Sociometry*, pages 178–192.
- McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7):855–878.
- Mendez, E., Kalkofen, D., and Schmalstieg, D. (2006). Interactive context-driven visualization tools for augmented reality. In *Proc. IEEE/ACM International Symposium on Mixed and Augmented Reality ISMAR 2006*, pages 209–218.
- Mertes, C. (2008). Multimodal augmented reality to enhance human communication. Master's thesis, Bielefeld University. Available from: <http://bieson.ub.uni-bielefeld.de/volltexte/2009/1414/>.
- Mertes, C., Dierker, A., Hermann, T., Hanheide, M., and Sagerer, G. (2009). Enhancing human cooperation with multimodal augmented reality. In *International Conference on Human-Computer Interaction*, pages 447–451, San Diego, USA. Springer.
- Milgram, P. and Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, E77-D(12).
- Milgram, P., Yin, S., and Grodski, J. (1997). An augmented reality based teleoperation interface for unstructured environments. In *Proc. American Nuclear Society 7th Topical Meeting on Robotics and Remote Systems*. Citeseer.
- Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371.
- Miram, W. and Krumwiede, D. (1985). *Informationsverarbeitung*. Materialien fuer den Sekundarbereich II : Biologie. Schroedel, Hannover, neubearb edition.
- Morency, L. (2006). Watson head tracker: Real-time head pose estimation and tracking, eye gaze estimation and gesture recognition from usb or stereo camera. Available from: <http://sourceforge.net/projects/watson/> [cited 2011-06-21].
- Morency, L. (2009). Co-occurrence graphs: contextual representation for head gesture recognition during multi-party interactions. In *Proceedings of the Workshop on Use of Context in Vision Processing*, pages 1–6. ACM.
- Morency, L., Rahimi, A., Checka, N., and Darrell, T. (2002). Fast stereo-based head tracking for interactive environments. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002. Proceedings*, pages 390–395.
- Morency, L., Sidner, C., Lee, C., and Darrell, T. (2005). Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24. ACM.
- Mortensen, C. D. (1972). *Communication: the study of human interaction*. McGraw-Hill, New York.

- Munhall, K., Jones, J., Callan, D., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility – head movement improves auditory speech perception. *Psychological Science*, 15(2):133–137.
- Munteanu, C., Baecker, R., Penn, G., Toms, E., and James, D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 493–502. ACM.
- Nakatani, C. and Hirschberg, J. (1993). A speech-first model for repair detection and correction. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 46–53. Association for Computational Linguistics.
- Narchet, F., Meissner, C., and Russano, M. (2011). Modeling the influence of investigator bias on the elicitation of true and false confessions. *Law & Human Behavior*.
- Neumann, A. (2011). Design and Implementation of Multi-modal AR-based Interaction for Cooperative Planning Tasks. Master's thesis, Bielefeld University.
- Noller, P. and Callan, V. (1989). Nonverbal behavior in families with adolescents. *Journal of nonverbal behavior*, 13(1):47–64.
- O'Conaill, B., Whittaker, S., and Wilbur, S. (1993). Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-computer interaction*, 8(4):389–428.
- O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., and Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, 8(2):177–192.
- Ong, S., Yuan, M., and Nee, A. (2008). Augmented reality applications in manufacturing: a survey. *International journal of production research*, 46(10):2707–2742.
- O'Sullivan, T., Hartley, J., Saunders, D., and Fiske, J. (1983). *Key concepts in communication*. Methuen, London.
- Pallant, J. (2005). *SPSS survival manual*. Open Univ. Press.
- Papagiannakis, G., Singh, G., and Magnenat-Thalmann, N. (2008). A survey of mobile and wireless technologies for augmented reality systems. *Computer Animation and Virtual Worlds*, 19(1):3–22.
- Patterson, R., Winterbottom, M., and Pierce, B. (2006). Perceptual issues in the use of head-mounted visual displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(3):555.
- Pickering, M. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.
- Pitsch, K. (2010). CA and the design of multimodal human-robot-interaction: On 'pause & restart' and nodding as communicational resources. Paper presented at the International Conference on Conversation Analysis 2010 (ICCA).

- Poggi, I., D'Errico, F., and Vincze, L. (2010). Types of Nods. The polysemy of a social signal. In *Proceedings of the 7th international conference on language resources and evaluation*, pages 17–23.
- Pölönen, M., Järvenpää, T., and Häkkinen, J. (2010). Comparison of near-to-eye displays: subjective experience and comfort. *Journal of Display Technology*, 6(1):27–35.
- Posner, M. and Cohen, Y. (1984). Components of visual orienting. *Attention and performance X*, pages 531–556.
- Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286.
- Reitmayr, G. and Schmalstieg, D. (2004). Collaborative augmented reality for outdoor navigation and information browsing. In *Proc. Symposium Location Based Services and Tele-Cartography*. Available from: http://www.ims.tuwien.ac.at/publication_detail.php?ims_id=124.
- Rolland, J., Biocca, F., Barlow, T., and Kancherla, A. (1995). Quantification of adaptation to virtual-eye location in see-thru head-mounted displays. In *Proc., Virtual Reality Ann. Internat. Symp.*, page 56. Published by the IEEE Computer Society.
- Rosenthal, M., State, A., Lee, J., Hirota, G., Ackerman, J., Keller, K., Pisano, E., Jiroutek, M., Muller, K., and Fuchs, H. (2001). Augmented reality guidance for needle biopsies: A randomized, controlled trial in phantoms. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001*, pages 240–248. Springer.
- Rudnicky, A., Hauptmann, A., and Lee, K. (1994). Survey of current speech technology. *Communications of the ACM*, 37(3):52–57.
- Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.
- Scaletti, C. (1994). Sound synthesis algorithms for auditory data representations. In *Auditory display*. Gregory Kramer.
- Schachter, S. (1951). Deviation, rejection, and communication. *The Journal of Abnormal and Social Psychology*, 46(2):190.
- Schegloff, E. (1987). Some sources of misunderstanding in talk-in-interaction. *Linguistics*, 25(1):201–218.
- Schegloff, E., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, pages 361–382.
- Schillingmann, L., Wrede, B., and Rohlfing, K. J. (2009). A Computational Model of Acoustic Packaging. *IEEE Transactions on Autonomous Mental Development*, 1(4):226–237. Available from: <http://dx.doi.org/10.1109/TAMD.2009.2039135>.
- Schnier, C., Pitsch, K., Dierker, A., and Hermann, T. (2011a). Adaptability of communicative resources in ar-based cooperation. In *Gesture and Speech in Interaction (GESPIN)*, Bielefeld. GESPIN 2011. Available from: <http://gespin.uni-bielefeld.de/?q=node/66>.

- Schnier, C., Pitsch, K., Dierker, A., and Hermann, T. (2011b). Collaboration in augmented reality: How to establish coordination and joint attention? In *European Conference on Computer-Supported Cooperative Work*, Aarhus, Denmark.
- Shannon, C. and Weaver, W. (1962). *The mathematical theory of communication*, volume 19(9). University of Illinois Press Urbana.
- Shriberg, E. and Stolcke, A. (1996). Word predictability after hesitations: A corpus-based study. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1868–1871. IEEE.
- Sidner, C., Lee, C., Kidd, C., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.
- Spexard, T. P., Siepman, F. H. K., and Sagerer, G. (2008). A memory-based software integration for development in autonomous robotics. In *International Conference on Intelligent Autonomous Systems*, pages 49–53, Baden-Baden, Germany.
- Stiefelhagen, R. (2002). Tracking focus of attention in meetings. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces-Volume 00*, page 273. IEEE Computer Society.
- Sutherland, I. (1968). A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pages 757–764. ACM.
- Swann, J. (2004). *A dictionary of sociolinguistics*. Edinburgh Univ. Press, Edinburgh.
- Tang, A., Owen, C., Biocca, F., and Mou, W. (2003). Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 73–80. ACM.
- Taylor, T. J. (1992). *Mutual Misunderstanding: Scepticism and the theorizing of language and interpretation*. Duke University Press Books.
- Thomas, B., Krul, N., Close, B., and Piekarski, W. (2003). Usability and Playability Issues for ARQuake. In *Entertainment computing: technologies and applications: IFIP First International Workshop on Entertainment Computing (IWEC 2002), May 14-17, 2002, Makuhari, Japan*, page 455. Springer Netherlands.
- Ulbricht, C. and Schmalstieg, D. (2003). Tangible augmented reality for computer games. In *Proceedings of the Third IASTED International Conference on Visualization, Imaging and Image Processing*, pages 950–954.
- Vatavu, R., Pentiu, Ș., and Chaillou, C. (2005). On natural gestures for interacting in virtual environments. *Advances in Electrical and Computer Engineering*, 24(5).
- Vecera, S. and Johnson, M. (1995). Gaze detection and the cortical processing of faces: Evidence from infants and adults. *Visual Cognition*, 2(1):59–87.
- Verdonik, D. (2010). Between understanding and misunderstanding. *Journal of Pragmatics*, 42(5):1364–1379.

- Vettin, J. and Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2):93–115.
- Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese* 1. *Journal of Pragmatics*, 32(8):1177–1207.
- Watson, J. and Hill, A. (1984). *A dictionary of communication and media studies*. Edward Arnold, London.
- Watzlawick, P., Bavelas, J. B., and Jackson, D. D. (1971). *Menschliche Kommunikation – Formen, Störungen, Paradoxien*. Huber, Bern.
- Wikipedia (2012). Attention — wikipedia, the free encyclopedia. [Online; accessed 31-January-2012]. Available from: <http://en.wikipedia.org/w/index.php?title=Attention&oldid=474115292>.
- Wöhler, N.-C. (2009). Maschinelles Lernen von Kopfgesten mit Ordered Means Models.
- Wöhler, N.-C., Großekathöfer, U., Dierker, A., Hanheide, M., Kopp, S., and Hermann, T. (2010). A calibration-free head gesture recognition system with online capability. In *International Conference on Pattern Recognition*, Istanbul, Turkey.
- Wrede, S., Fritsch, J., Bauckhage, C., and Sagerer, G. (2004a). An XML based framework for cognitive vision architectures. In *Proc. Int. Conf. on Pattern Recognition*, volume 1, pages 757–760.
- Wrede, S., Hanheide, M., Bauckhage, C., and Sagerer, G. (2004b). An active memory as a model for information fusion. In *Proc. 7th Int. Conf. on Information Fusion*, volume 1, pages 198–205. Available from: <http://citeseer.ist.psu.edu/wrede04active.html>.
- Wrede, S., Hanheide, M., Wachsmuth, S., and Sagerer, G. (2006). Integration and coordination in a cognitive vision system. In *Int. Conf. on Computer Vision Systems*.
- Yngve, V. (1970). On getting a word in edgewise. In *sixth regional meeting Chicago linguistic society*, pages 567–578.
- Yoshida, H. and Smith, L. (2008). What's in View for Toddlers? Using a Head Camera to Study Visual Experience. *Infancy*, 13(3):229–248.
- Young, S. (1996). A review of large-vocabulary continuous-speech. *Signal Processing Magazine, IEEE*, 13(5):45.
- Zaefferer, D. (1977). Understanding misunderstanding: a proposal for an explanation of reading choices. *Journal of Pragmatics*, 1(4):329–346.
- Zhou, F., Duh, H., and Billinghurst, M. (2008). Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In *7th IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR 2008)*, pages 193–202. IEEE.