Part of workshop "Multimodal Corpora: Beyond Audio and Video" at IVA 2013.

Augmented Reality as a Tool for Linguistic Research: Intercepting and Manipulating Multimodal Interaction

Karola Pitsch[1], Alexander Neumann[3], Christian Schnier[1,2], Thomas Hermann[3]

[1]Interactional Linguistics & HRI, [2]Applied Informatics, [3]Ambient Intelligence
Bielefeld University, Faculty of Technology, P.O.Box 100131, 33501 Bielefeld, Germany
{kpitsch}{alneuman}{cschnier}{thermann}@techfak.uni-bielefeld.de

Abstract. We suggest that an Augmented Reality (AR) system for coupled interaction partners provides a new tool for linguistic research that allows to manipulate the coparticipants' real-time perception and action. It encompasses novel facilities for recording heterogeneous sensor-rich data sets to be accessed in parallel with qualitative/manual and quantitative/computational methods.

Keyowrds. Augmented Reality, Multimodal Interaction, Semi-Experimental

1 Introduction

Linguistic research has increasingly become interested in the multimodal dimension of communication. It investigates how different modalities – talk, gaze, gesture, embodied actions – are intertwined. Their interplay has been conceived of as complex holistic *gestalts* and – integrating the material environment – as dynamic "contextual configurations". It is based on the participants' "mutual monitoring" and "online analysis (Goodwin 2000). Investigating authentic everyday communication, we can reconstruct the ways in which participants deploy these multimodal resources for solving specific interactional tasks. Analyses reveal how the interactional procedures are locally shaped and dynamically adjusted by the participants. While building collections of comparable cases allows us to study an interactional phenomenon in detail, it is difficult to tear apart the complex multimodal configurations. Which functions do different modalities assume for interactional coordination? To which extent could one modality be substituted by another one? How precise would timing need to be?

Authentic real-world interaction in which the participants interact under 'difficult' conditions has provided first insights into the dynamic adjustment of communicational resources once the "ecology of sign systems" [2] is challenged: When communicating in a foreign language, gestures receive a prominent role; in operation theatres, participants coordinate their actions although parts of their faces are hidden behind masks. Early developmental stages of a novel communication technology lead to perturbations, such as time lag or distorted views. This allows for investigating how participants adjust to these new constraints, but the interactional conditions cannot be manipulated reliably by the researcher. Psycholinguistics has a longstanding tradition

in experimenting with communicational parameters: limiting visual access to the co-participants' head, hands or material resources; eliminating prosodic contours or syntactical information; etc. However, these procedures are limited in that they use static physical manipulations or pre-recorded stimuli excluding real-time interaction.

We suggest that an Augmented Reality (AR) system for coupled interaction partners provides a new tool for linguistic research that allows us to manipulate the coparticipants' real-time perception and action. Also, it encompasses novel facilities for recording sensor-rich data sets which can be accessed in parallel with qualitative/manual and quantitative/computational methods. In this paper, we present our new conceptual approach, the resulting multimodal corpus, and point to challenges and solutions in developing a consistent corpus across various sensor-modalities. We provide a first example of how such a corpus can be analytically deployed.

2 Dissecting Multimodal Communication

2.1 Augmented Reality: The ARbInI system

We developed an AR-system ('ArbInI') to intercept and manipulate in real-time the audiovisual perception of interacting coparticipants [1]. The participants wear head-mounted displays (HMDs), which block the user's direct view, but include an integrated camera which projects its recording on the user's screen and thus constitutes their visual perception. Users are given a set of wooden blocks with machine-readable patterns serving as material handles for virtually displayed objects. Sharing an interaction space, the participants' perception of the world is a virtual one. Their talk is recorded via headset microphones. This way, the users' audio-visual data streams can be captured and manipulated in real-time. As both participants are connected to the same computer, information from one user can be displayed to the other.

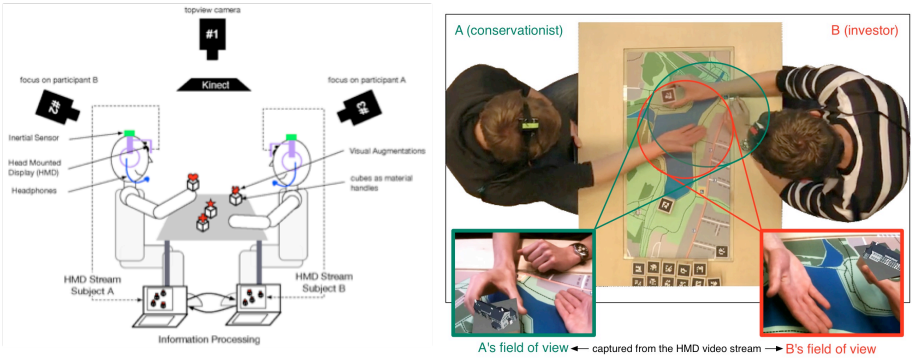

Figure 1a. AR System (dyadic). **1b.** AR System in use with Obersee scenario.

These audiovisual streams, the position of the virtual objects and the users' head movements are recorded synchronously. External cameras (HDV, Kinect) capture the scene from a top-view and from different side angles (Fig.1a, Table 1).

2.2 Manipulating Interactional Conditions

We created a corpus of spontaneous co-present interactions of two (tentatively three) participants. In several studies using the same scenario, we have begun to systematically manipulate specific features of the interactional conditions. This introduces a disturbance for the participants ranging from limiting their possibilities for 'mutual monitoring' (ii) and manipulating their visual perception (iii, iv) to altering the timing of different resources (v). Currently, the corpus contains the following conditions:

(i) **Face-to-Face** (12 dyads): As a baseline, the participants interact in a conventional face-to-face setting, only equipped with the inertial head sensors.

(ii) **AR-Baseline** (10 dyads, 1 triad): The HMDs provoke a limited field of view of 42° instead of the 180° of human perception. Thus, the ArbInI system manipulates the participants' 'mutual monitoring' as a basis on which further mechanisms operate.

(iii) **AR-PartiallyVisibleObjects** (2 dyads): While having access to all material handles of the objects, for each participant only a limited number of objects is shown which are invisible to the respective coparticipant.

(iv) **AR-SwappedObjects** (2 dyads): For two or three material handles, the virtual objects are swapped (maximal vs. minimal differences) for the participants.

(v) **AR-Desync** (ongoing): Auditory and visual information is desynchronized.

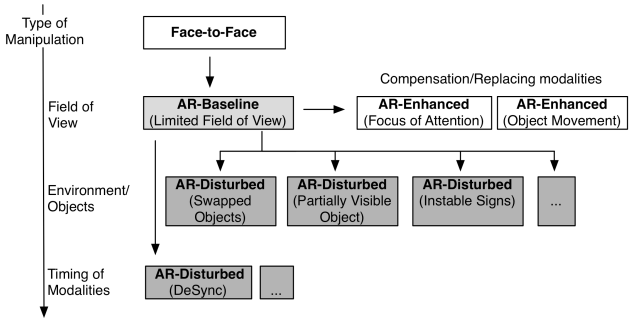

Figure 2. Manipulating Interactional Conditions

Additonally, we explore to which extent communicational resources relevant in face-to-face interaction could be shifted to and displayed in other modalities [4]. Therefore, we (vi) **visually augment the coparticipant's visual focus of attention** (8 dyads) and (vii) introduce **sonification of object movement** [3] (7 dyads).

3 Scenario & Study design

Participants were asked to jointly plan the redesign of a local recreation area. To stimulate negotiation, they assumed opposed roles of 'investor', 'mayor'/'conservationist'. They were seated vis-à-vis across a table with equal access to a map of the area (Fig.1b) and given 18 objects, which could be used as planning concepts and placed on the map to mark out the spaces for specific investments. The interaction took about 20 minutes followed by 10 minutes of free conversation.

4 Corpus

For a combined qualitative/manual and quantitative/automated analysis, a multimodal corpus of parallel videobased and sensor-rich data is needed.

4.1 Primary Data: Synchronizing heterogeneous data sources

A particularly rich set of sensor data is recorded: the users' visual and auditory perception (Firewire cameras on HMD, Microphones), head movements and position (BRIX gyroscope and accelerometer), and the objects and their positions as they occur in both participants' fields of view (ARToolkit). Additionally, the activities on the table are recorded from a top view using a 640x480 depth image (Kinect), and three (resp. four) external HDV cameras are focused on each participant and the scene.

Table 1. Primary sensor data and marker-based data

Sensor	Data	#	Hz	Comment
Firewire Camera	640x480 Screencapture	2	25	Participants' field of view
Microphone	WAV audio	2	44100	Participants' speech
BRIX - Gyroscope	16 bit time series	2	30	Head movement velocity
BRIX - Accelerometer	16 bit time series	2	30	Head position
ARToolkitPlus	XML	2	25	Position/visibility of augmented objects
Microsoft Kinect	640x480 depth image	1	30	Top view of scene
HDV Camera	1080p HD Video	3	50	External view

With these heterogeneous data, different sample frequencies, unrelated time stamps and system-internal/-external streams, particular challenges lie in their syncronization:

(1) In principle, all computer-based information could be linked via a global system time stamp. However, even when some sensors record with an identical sample frequency (e.g. Kinect, BRIX), they do not necessarily produce a similar recording time. At moments, their time stamps vary for up to a half of the sample rate. To deal with this problem, we (i) preserve the original recording time of every sensor reading and (ii) extend it by a time delta which indicates the 'age' of the readings. This also prevents incomplete data sets and the need for a 'dummy' token in the model.

(2) As the system-external audiovisual recordings do not contain a global time stamp, a specific synchronization event is required to link them to the system-internal data. Therefore, we developed a 'digital clapper board' (Fig.3a): A BRIX sensor toolkit [5] is turned into a hardware device to emit a specific light/sound pattern which is triggered by the system and visible/audible for the external HDV cameras. It is used at the start and end of a trial to also cope with framerate inconsistencies. The Unix timestamp is logged and used to define the global time and framerate of each stream.

4.2 Post-processing

As raw sensor data contain ‘noise’, they need to be (1) validated and (2) unified before being syncronized. These steps are conducted consecutively, their methods and results are documented and stored separately.

(1) **Data Validation**: Real-time vision-based tracking accuracy depends on the video image quality. Motion blur caused by moving HMDs and markers lead to 'tracking gaps' and missing/false results. Similar effects occur when inertial sensors experience physical impact. We apply a median filter to smooth oscillating sensor readings and to detect outliers based on velocity or orientation (verified manually).

(2) **Data Unification**: The recordings are converted to a unified discrete data representation with a resolution of 25Hz and a Unix timestamp. Numerical readings and annotations are stored as CSV data tables, video files as H264-encoded mp4 and audio recordings as uncompressed WAV files. Depth images are stored in a binary file where every frame contains the Unix timestamp of its creation time.

4.3 Secondary Data: Automated features and manual annotations

Using the sensor-based primary data, we derived a set of features representing more complex interactional phenomena, such as the movement and visibility of objects or speech activity. Further features can be added.

Table 2: Features derived from primary data

Input Data	Methods	Output Data
Object world coordinates	Erosion, basic arithmetics, derivation	Object movement
Object Screen coordinates; HMD captures	Erosion, basic arithmetics	Object visibility
Audio Recordings	Amplitude threshold, activity detection	Speech activity

Manual annotations are created using the annotation tool ELAN (http://tla.mpi.nl/tools/ tla-tools/elan/), including e.g. the action structure, a verbal transcript, annotation for deixis, gaze direction or the participants’ manipulation of objects.

4.4 Accessing and representing sensor data

While audiovisual data and manual annotations can be used with ELAN, machine-readable sensor data require specialized methods. We developed scripts to convert our time series data to an ELAN-compatible format to be represented with their existing tools. The primary and secondary data are explored with the quantitative data analysis facilities provided by scientific computing tools (e.g. NumPy for Python, www.numpy.org/) to filter data (sensors, time periods or data quality) and to carry out quantitative analysis. As a more intuitive access to the data, we created a data viewer based on matplotlib, a python plotting library and compatible with the data representation in NumPy (Fig. 3b). It fuses the marker visibility with a setup scheme and allows to inspect which participant has visual access to which objects at a given moment in time.

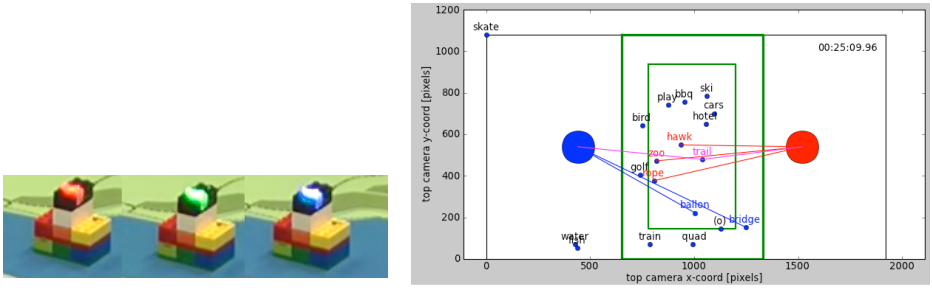

Figure 3a. 'Digital clapper board' emitting three light beams and 2800Hz sound for 0.2 sec. with a 1 sec. pause. **3b.** Data viewer

5 Linking qualitative and quantitative analysis: The case of 'co-orientation'

As in the AR-Baseline condition (ii), the participants' abilities of "mutual monitoring" are restricted, we are interested in understanding how they can coordinate their talk and embodied actions with each other. We focus on one basic interactional task, i.e to establish co-orientation towards an object. While in the literature, the importance of gaze following to achieve "joint attention" has been highlighted, this resource is limited here. How does this circumstance effect the communicational procedures?

Consider an example, in which the participants are at the transition from one activity – suggesting an object to be built which is then negotiated – to the next one. In our example (Fig.4), both participants A and B are oriented to the same (only) object in the middle of the map when discussing about the object 'playground' (#1). As B places it, A affirms this choice and scrutizines the map, while B proceeds to orient to the stack of objects and thereby projects a new activity (#2). Thus, at this moment, the participants are oriented to two different activities. This duality is resolved towards establishing a joint focus of attention, when B utters the discourse marker "so" (projecting a new activity) and directs his hand to an object in the stack. A's gaze follows to the stack (#3). Thus, a combined procedure employing a verbal marker and a hand movement appear to help the coparticipant orient to B's action and to establish a joint focus of attention. As B suggests to build a hotel, they both orient to it (#4).

To detect such moments of interactional sequential organization on the basis of machine-readable data, we need to include the participants' perception, talk and manual actions. Investigating the secondary data on 'object visibility' (Table 2), we find:

(#1) A and B are each oriented to one object, and their intersecting set is 'one' too.

(#2) B looks at 12-15 objects, while A looks at 0 objects.

(#3) For both A and B about 10-15 objects are in the field of view, and in subsequent moments about 8 objects appear in their intersecting set.

(#4) A and B are each oriented to one object, and their intersecting set is 'one' too.

Using such data, we are able to assess the structure of the participants' actions, which we can use for pre-structuring the corpus or, in combination with other modalities, to identify sequential organization structures.

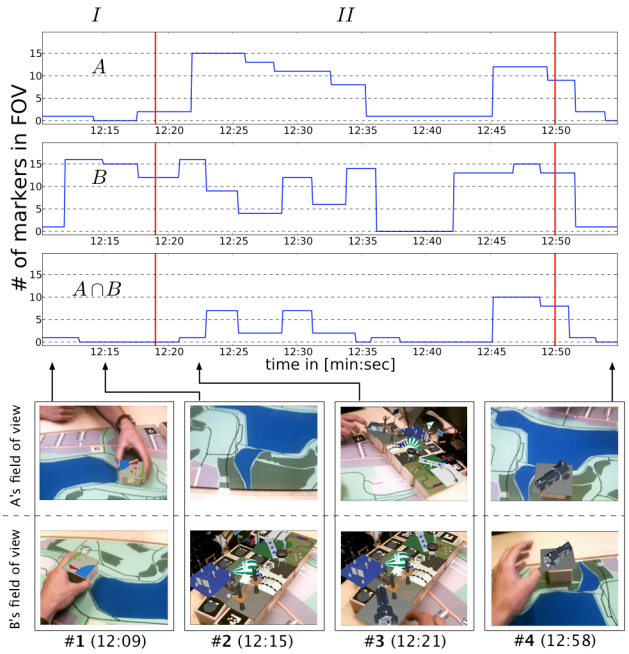

Figure 4. Example of objects in participants' view

6 Future Work

Future work consists in deriving further secondary features, continuing qualitative/ quantitative analysis, and exploring further manipulation and compensation methods.

Acknowledgments. The authors gratefully acknowledge the financial support from the CRC 673 'Aligment in Communication' funded by the German Research Foundation (DFG). Karola Pitsch also acknowledges the financial support from the Volkswagen Foundation. We thank Katharina Geretzky for her support during the studies.

7 References

1. Dierker, A., Mertes, C., Hermann, T., Hanheide, M., Sagerer, G. (2009): Mediated attention with multimodal augmented reality. In: ICMI-MLMI 2009, 245-252.
2. Goodwin, C. (2000): Action and embodiment within situated human interaction. In: JP 32, 1489-1522.
3. A. Neumann, T. Hermann (2013). Interactive Sonification for Collaborative AR-based Planning Tasks for Enhancing Joint Attention. In: ICAD.
4. Schnier, C., Pitsch, K., Dierker, A., Hermann, T. (2011). Collaboration in Augmented Reality: How to establish coordination and joint attention? ECSCW, 405-416.
5. Zehe, S. (2012): BRIX. An Easy-to-Use Modular Sensor and Actuator Prototyping Toolkit. In: SeNAmi, 823-829.