In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (LREC), Malta, 2010, pp.63-69.

Linking Conversation Analysis and Motion Capturing: How to robustly track multiple participants?

Karola Pitsch[1,2], Bernhard Brüning[3], Christian Schnier[1], Holger Dierker[3], Sven Wachsmuth[1,3]

[1]Applied Informatics, [2]CoR-Lab & [3]CITEC
Bielefeld University, Faculty of Technology, P.O.Box 100 131, 33501 Bielefeld, Germany
E-mail: {kpitsch}{bbruenin}{cschnier}{hdierker}{swachsmu}@techfak.uni-bielefeld.de

Abstract

If we want to model the dynamic and contingent nature of human social interaction (e.g. for the design of human-robot-interaction), analysis and description of natural interaction is required that combines different methodologies and research tools (qualitative/quantitative; manual/automated). In this paper, we pinpoint the requirements and technical challenges for constituting and managing multimodal corpora that arise when linking Conversation Analysis with novel 3D motion capture technologies: i.e. to *robustly* track *multiple* participants over an *extended* period of time. We present and evaluate a solution to by-pass the limits of the current standard Vicon system (using rigid bodies) and ways of mapping the obtained coordinates to a human skeleton model (inverse kinematics) and to export the data into a format that is supported by standard annotation tools (such as ANVIL).

1. Introduction: Detecting interactional patterns across disciplines

In recent years, a range of initiatives has begun to enable robots and other technical systems to engage in more naturalistic forms of interaction with the human user. After important advances have been made both in detecting/sensing human conduct and creating human-like forms of system output, a central challenge today consists in enabling technical systems to participate in and deal with the dynamic nature of human social interaction: Systems need to observe – on a micro-level – human multimodal conduct, interpret it as meaningful in terms of the interactional organisation and react appropriately. While there is a longstanding tradition in the field of Ubiquitous Computing and Computer Supported Cooperative Work (CSCW) to include qualitative approaches, such as Ethnography and/or Ethnomethodological Conversation Analysis (EM/CA), into the development cycle of technical systems (e.g. Dourish, 2009; Luff et al., 2009), only recently researchers have begun to scoop from these same sources for the design of robot systems (Nishida et al., 2007; Kuzuoka et al., 2008; Pitsch et al., 2009). In particular, for the design of robot systems, EM/CA – with its fine-grained analysis of video data – is able to provide insights into the sequential organisation of interaction, reveal patterns of social conduct and investigate how one person's multimodal conduct both reacts to and shapes their co-participants' actions. On the one hand, this offers a rich basis for modelling the dynamic and contingent nature of social interaction; on the other hand, the ways in which a qualitative, video-based EM/CA is able to present its findings do not always match the sort of quantifiable information that is required for building computational algorithms. Against this background, we argue that interactional corpora – combining video recordings and new motion capture technologies – are required that allow researchers to use different methodologies and research tools (qualitative/quantitative; manual/automated) on the same data set (cf. Chen et al., 2006). However, with such an integrated methodological approach a range of new technical challenges arise regarding the constitution and management of multimodal corpora.

In this paper, we pinpoint the requirements and technical challenges that a combined approach brings to light with regard to establishing multimodal corpora (section 2), present our solution to solve these problems (section 3) and evaluate seemingly 'unnatural' aspects of our approach (section 4).

2. Corpus: Requirements and technical challenges

When planning and establishing a corpus that is designed to investigate multimodal turn-taking and other aspects of interactional organization in a group of two vs. three participants with a mixed approach of qualitative/ quantitative and manual/automated analysis, we have been largely informed by analytical experience from another ongoing interdisciplinary project (iTalk). We will use examples from this study to point out the requirements that the new corpus would need to fulfil.

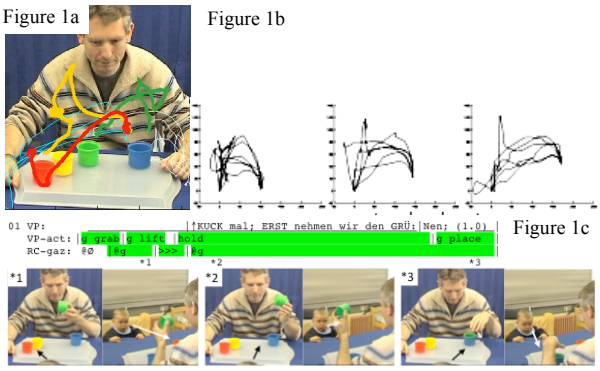


Figure 1: Parent demonstrating 'stacking cups' to his infant. (a) Video still with overlayed hand trajectories; (b) Normalized hand trajectories overlayed of several participants; (c) Transcript and stills from two cameras.

The iTalk project (www.italkproject.org) aims at

enabling robots to learn within and from the interaction with a human partner. Given the analogy of limited cognitive capabilities both in robots and young infants, our starting point consists in understanding the ways in which parents demonstrate actions to their young infants as a model for the design of the robot system (Rohlfing et al., 2006). In order to build the robot system, we need to know e.g. how participants structure their actions, which features are constitutive for tutoring, how the recipients react to the demonstration and how this, in turn, influences on the presenter's demonstration (Pitsch et al., 2009; Vollmer et al. 2009). Therefore, we have built and are analysing – with different research methodologies – a video corpus, in which 128 parents are demonstrating a set of actions to their infants aged 8 to 36 months. In this line of research, Ethnomethodological Conversation Analysis offers an interactional perspective on the task and is able to reveal with its qualitative-manual analysis the strategies and methods used by the participants, to uncover relevant multimodal features/cues and to find interactional patterns and systematic relationships between the co-participants' actions. At the same time, this approach is limited e.g. in describing the presenter's manual actions in terms of the concrete shape of the hand trajectory performed in a given demonstration. Interestingly, these shapes differ considerably in the corpus, which becomes visible once a semi-automatic computational 2D hand tracking is applied to the video data delivering time-stamped x,y-coordinates of the parent's hand motions (Fig. 1a, 1b). While EM/CA is able to reveal the interactional causes and effects of the variability in the hand trajectories (linked to the child's focus of attention, Fig. 1c), mathematical and statistical methods can describe these trajectories in a way that they become suitable for building computational algorithms that allow a robot to distinguish certain types of actions. At the same time, relevant interactional categories evolve from CA-analysis, which, then, can be systematically transcribed/annotated with corpus tools (such as ELAN, ANVIL) and be subject to a computational investigation of correlations between the different interactional variables on the entire corpus (Pitsch et al., 2009).

Not only does this example give a case for closely interrelated qualitative-quantitative analysis, but it also provides us with central *requirements* when establishing a new interactional corpus that is designed for the same area of research: We need to be able to capture (i) the timely interplay of *several* (two or more) participants, (ii) their talk, gaze, body posture, gestures, head, arm and body motions, and (iii) interactional episodes that take about 30 minutes of time. As – for the parent-child-corpus – we only dispose of video recordings of the interaction, we had to develop a motion tracker in order to be able to precisely describe and analyze the hand trajectories. While this has proven extremely useful for our case (and might be oriented towards the sensors that current robot systems are equipped with), analysis is limited with regard to the features that can be tracked robustly and by the fact that it can only deliver 2D information (information about depths is missing). Thus, (iv) for the new corpus both video and 3D motion capture data are required.

However, if we attempt to use current state of the art 3D motion capture technologies for recording data with the requirements presented above, we are facing a crucial *technical challenge*: How can we *robustly* track *multiple* participants over an *extended* period of time?

Existing optical motion capture technologies, such as the Vicon system, have been originally developed for capturing human motions in the fields of sports and health sciences or for animating virtual characters in movies and computer games. Small reflective markers (spheres) are attached to particular places of the human body, tracked simultaneously by a set of (at least 10) infrared cameras and mapped to a generic model skeleton. In these cases, generally *one* single participant is recorded for a *short* period of time. In recent times, researchers have begun to use such systems also for recording multi-party interaction (Chen et al. 2006; Battersby et al. 2008). However, once we attempt to use the system to track two or three participants during an interaction period of e.g. 30 minutes, we encounter a range of problems: (i) Due to visual obstruction, the system easily loses the individual markers during the recording process. (ii) This leads either to incomplete and thus problematic data or an extensive post-processing phase is required, in which markers need to be re-assigned and labeled. We have conducted a set of internal trials, which revealed that 1 minute of recording time requires about 60 minutes of post-processing for one participant – impossible to handle for large corpora.

3. The "Obersee Corpus": Suggestions for robustly tracking multiple participants

When establishing our corpus designed to investigate multimodal interactional organisation with a mixed methodological approach, we needed to find ways to by-pass the limits imposed by the current Vicon system, i.e. to *robustly* track *multiple* participants over an *extended* period of time. In what follows, we present our solution which involves both changes in hardware and new algorithms for transforming the raw motion capture data.

3.1 Study Design

As the corpus should allow for investigating a range of different aspects of multimodal interactional organization, we choose a semi-experimental set-up that would engage groups of participants in (a) a free conversation and (b) a task-related interaction which requires the use of material objects and gesticulation. At the same time, we needed to both control the situation in a way to allow for comparison between different groups of participants and to be open enough to allow for spontaneous social interaction. Therefore, we invited in total 15 groups of participants (6 dyades, 9 triades) to

engage in a 20 minutes conversation, in which they were supposed to discuss and come up with a solution for a redesign of the local lake (the "Obersee") into a new recreation area. We asked them to each assume a certain role (financial investor, local mayor, Greenpeace activist) and provided them with a map of the area as well as a set of toy objects (such as inline skater, diver, quad, barbeque) that they could use for inspiration and (re-)position on the map. Afterwards, they were asked to remain seated while the experimenter had to check the recording, get the questionnaires to be filled out, which provided us with further 10 minutes of free conversation.

3.2 Technical Setup

We recorded these interactions with four HD video cameras, ten Vicon T20 cameras and an additional microphone hanging from the ceiling (Fig.2). While the video footage was stored individually, the Vicon data was (i) firstly gathered and processed by a Vicon MX Giganet server, (ii) then sent to a PC using the Vicon Nexus software V1.4.112 to detect (patterns of) Vicon markers and to calculate their position and orientation and (iii) finally sent to another PC for saving the data. This setup allowed us to by-pass the limits of recording time and amount of data imposed by the Vicon Nexus software while using its pattern recognition facilities.

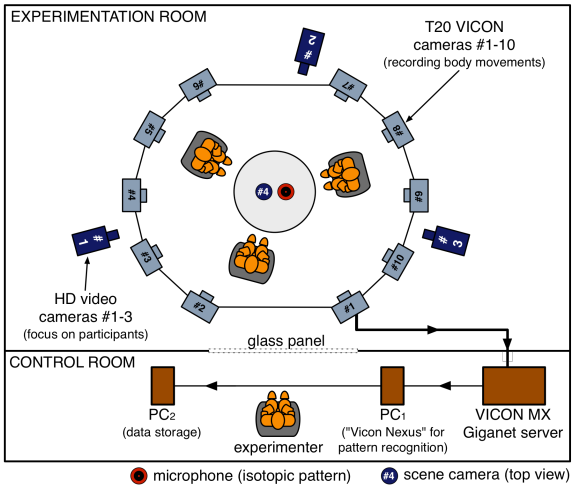


Figure 2: Technical setup

With this approach, we lose the function provided by the Vicon system of producing synchronized video and 3D motion capture data. We compensate for it by making one participant clap a slate at the beginning of the session, which creates a distinctive signal that can – afterwards – be automatically detected in the different media sources. In addition, a visual calibration pattern was positioned in the middle of the scene, so that we are able to calculate 3D information from the video footage.

3.3 Rigid bodies for robustly tracking three participants

In order to deal with the problem of losing markers and a resulting extensive post-processing, we decided to use – instead of individual markers – so-called "rigid bodies". A rigid body consists of a pattern of several markers that are spatially arranged in a particular way and can be distinguished from other rigid bodies (Fig. 3 and 4). It has a unique ID assigned by the marker, which, in turn, denotes the corresponding body part, so that it can be assigned to a position and an orientation in 3D space. The main advantage resides in the fact that – in case markers get lost – they can be automatically reassigned to each body limb by the system. Also, in the case of marker loss, chances are high that at least one or two markers (out of the set of five) are continuously tracked, so that limited information about the whereabouts of that particular body part will still be available. Consequently, no extensive manual post-processing is required.[1]

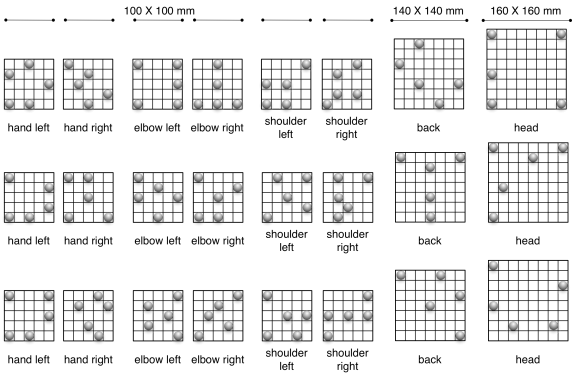


Figure 3: Three Sets of 8 rigid bodies worn by the participants

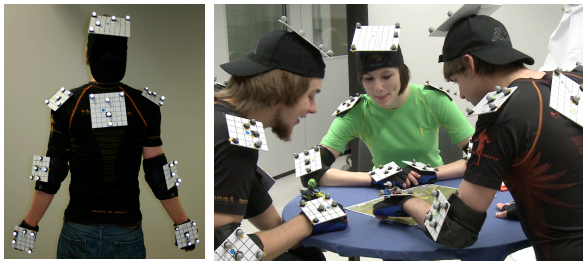


Figure 4: Participants wearing rigid bodies

In order to capture the most central movements of the human body, which are supposed to be interactionally relevant in a seated face-to-face setting, we used eight rigid bodies per person. These were attached to the head, back, left/right shoulders, left/right elbows and left/right hands (Figure 3). As we wanted to robustly track three participants simultaneously, we had to provide a set of 24 rigid bodies that were clearly distinguishable from each other. While we started with a systematic arrangement of markers on a 10 cm x 10 cm grid (allowing for a 5 by 5 grid), we soon had to increase the grid size to 16 cm x 16 cm (8 by 8 grid) to be able to create enough patterns that the Vicon Nexus software

[1]Systematic evaluation of this approach will be undertaken.

would robustly recognize as distinct. [2]
While the size of the rigid bodies was determined by the technical feasibility and robustness, we were concerned to keep their size as small and unobtrusive for the participants as possible. Initial pre-trials suggested that participants would rather tolerate the larger rigid bodies attached to their back and the top of their head, and could cope with the 10cm x 10cm grids at the other positions if they were fixed appropriately (e.g. by using thin fingerless biker gloves). Being aware that rigid bodies could potentially influence the participant's "natural" conduct in the experiment, we used a questionnaire to evaluate their experience of our setup (Section 4).

3.4 Skeleton representation and inverse kinematics

While our approach to by-pass the limits of the standard Vicon system (rigid bodies, "Vicon Nexus" software for detecting patterns of markers and giving their location and orientation, external data storage) allows us to robustly capture three participants over a long period of time, we have to find ways to map the rigid body's coordinates to a human skeleton model to calculate the joint angles.

To calculate the joint angles of the tracked person, we use a mathematical representation of the human skeleton based on the Denavit-Hartenberg-Convention developed and used in the field of robotics. It describes the transformation of a single joint with one degree of freedom to the next adjacent joint. For this, it uses four elementary transformations: $A_i = R_{z\,i-1}, * T_{zi-1} * T_{xi} * R_{xi}$

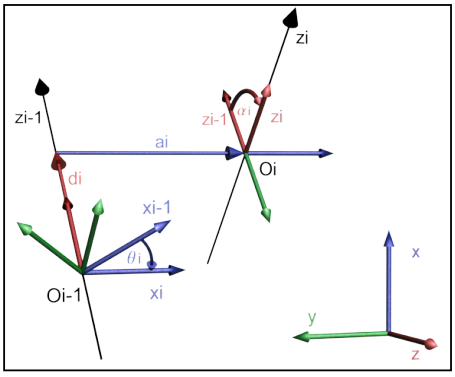


Figure 5: The four elementary transformation from one joint axis to the next adjacent joint

The transformation A_i contains a rotation around the previous z-axis, a translation along the previous z-axis, another translation along the current x-axis and a rotation around the current x-axis. Such a transformation can be used to model either a complete human skeleton or a single arm etc. (Fig.5). A single rotation joint is represented as a cylinder and has the ability to rotate around the z-axis which is parallel to the height of the cylinder (see Fig.6 where e.g. the shoulder has got three joints represented as cylinders).

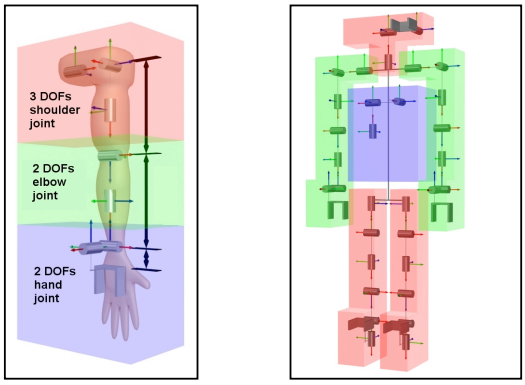


Figure 6: (a) Arm and (b) complete body representation in the Denavit Hartenberg Convention

Based on the mathematical description of the skeleton, we have developed algorithms that firstly calculate the positions of the joints out of the rigid body coordinates. Secondly, we proceed with inverse kinematics, in which the angles of each joint are calculated using the tangent = sinus/cosine = adjacent/opposite = y0/x0 (Spong et al., 2006; Brüning et al., 2008).

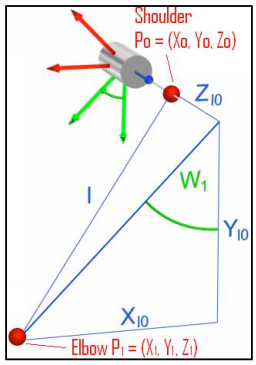


Figure 7: Inverse kinematics – Calculation of a single joint angle from a local joint coordinate system

From these calculations, we obtain the angles for one joint, which we then have to apply for all joints for each individual participant. When applying this procedure, we start by localizing the human body in space (i.e. the marker attached to the participant's back) and from there proceed by calculating step by step each further joint.

[2] While the rigid bodies and their locations were robustly tracked, initial investigation of the motion capture data showed slight problems for four markers, where – at moments – the orientation of the marker could not be precisely tracked. This can be caused by different factors (positioning of cameras, obstruction, the marker itself) and more detailed analysis of the causes will be required. At the current state, we used entire plastic plates as the basis for the rigid bodies. In a next iteration, we might consider cutting out the 'unused' space to reduce their obtrusiveness for the user. This, however, will need further consideration regarding mirror-invariance in the patterns.

3.5 Displaying data and integration into existing annotation tools

Once we have obtained the angles for all joints, we can display a skeleton of the human participant showing its posture at a given moment in time during the interaction (Fig.8a). The motion capture data also allows us to display and analyze in 3D the motion trajectories that the participants perform (Fig.8b).

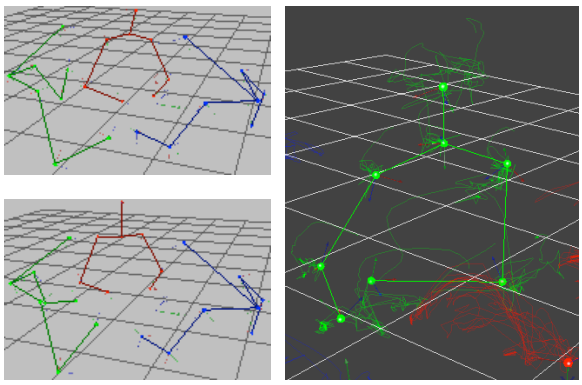


Figure 8: (a) Representations from the current scene and (b) including motion trajectories of one participant

In order to link the motion capture data with the video and sound files, we export the data obtained into a format that is supported by standard annotation tools (such as ANVIL) which are used by Conversation Analysts for transcription and annotation. To do so, we have developed a program that exports the motion capture data to the BVH (biovision hierarchy) format which is supported in the current version of the ANVIL annotation tool (Kipp et al., 2010). This file format consists of two main parts: one containing a description of the hierarchical order of the joints that describe the skeleton with the offsets from one joint to another; the other one comprising the angles of all joints written in the order of their hierarchical arrangement.

However, at the current state, ANVIL only supports motion capture data displaying *one* human; extensions will be required to also include the appropriate display of the interactional organization between *multiple* participants.

4. User Experience: How obtrusive are rigid bodies for the participants?

When developing our approach of using rigid bodies we were concerned with the question to which extent these objects might be – when being attached to the human participants – uncomfortable to wear and obtrusive for interacting or grabbing objects. While initial pre-trials suggested this approach to be acceptable, we wanted to evaluate the participants' experience more systematically. Therefore, after the experiment, we asked all participants to fill out a short questionnaire collecting information about their experience with regard to participating in (semi-)experimental studies, being videotaped and having used motion capture systems before.

In particular, two aspects are of interest here. We asked whether the participants felt disturbed during their interaction (i) by being videotaped and (ii) by the rigid bodies attached to their different body parts. Analysis reveals that in general, participants feel only 'slightly' disturbed by the recording equipment with a *similar* distribution between (i) being video-taped and (ii) having rigid bodies attached to their bodies (2x s-field- X^2-Pearson´s chi square test, with α=0,05). This result confirms our initial observations from the pre-trials and suggests that the rigid bodies do not seem to create more a unauthentic situation than video recordings – with the latter being recognized as a standard method of data acquisition in research.

Considering the answers for the motion capture in detail, we find that the participants' disturbance with regard to hand and elbow markers shows a tendency for slight disturbance while they feel hardly, i.e. 'less than slightly', disturbed with head, shoulder and back markers (Fig. 9).

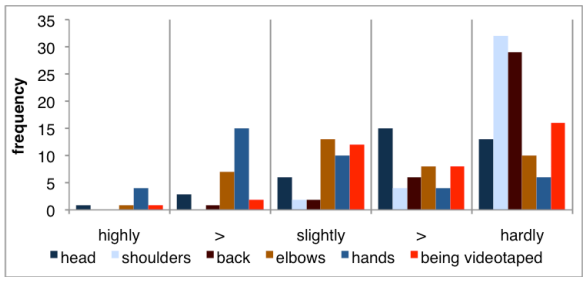
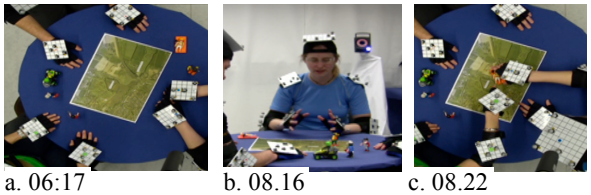


Figure 9: How disturbed do participants feel when (i) being videotaped and (ii) using rigid body markers?

These results and their analogy with the video recording situation suggest that our approach of using rigid bodies for overcoming the problem of robustly tracking multiple participants might able to generate – both technically and socially – valid interactional data.

In addition to asking users about their experience, close examination of the video recordings should allow to explore in more detail the rigid bodies' impact on the users' conduct in situ and the potential form of disturbance they might cause. Initial analysis of one group reveals that participants, at the beginning of the experiment, appear to position their hands rather flat on the table and without much manual actions or motions (Fig. 10a). This, however, changes step by step as the interaction unfolds. Around 8 minutes in the recording – when the participants are immersed in their roles and tasks – the first instance of gesticulation can be observed (Fig. 10b), and participants begin to bring their hand (and markers) close to some body else's hand (and markers) while manipulating objects on the plan (Fig. 10c). At this time, also vertical hand positions begin to occur, which suggests that participants are not particularly concerned (any more) with the question of the rigid bodies' adherence or trackability (Fig. 10d). After 17 minutes, participants can be seen to approach their hands even closer to the co-participant's hands (Fig. 10e) and to also reach to the other side of the table while

crossing their co-participants' arms and markers (Fig. 10f).



a. 06:17 b. 08.16 c. 08.22

d. 08.25 e. 17.52 f. 17.54

Figure 10: Participants' gestures and manipulation of objects changing over time

These observations suggest that during the first 5 to 8 minutes of an experiment participants seem to use more controlled hand motions and gestures, while after this initial period both their individual motions and their collaboration with others become more vivid. However, the question whether this effect is linked to the general situation of being observed or whether it might be caused specifically by the rigid bodies needs further consideration. Further analysis will also need to include other groups of participants, to investigate the motion of different body parts and begin to link the participants' motions to the concrete interactional tasks being carried out.

5. Conclusion and Future Work

In this paper, we have presented a system that is able to *robustly* track and record *multiple* participants over an *extended* period of time (30 minutes) with a 3D motion capture system. Linking this data to four HD video recordings, we are able to establish a multimodal corpus that is suitable for a combined qualitative/quantitative corpus analysis. The recorded data from the different sources can be analyzed using both Conversation Analysis and mathematical/statistical methods.

Our approach consists of by-passing the limits of the current standard Vicon system (using rigid bodies) and ways of mapping the obtained coordinates to a human skeleton model (inverse kinematics) and to export the data into a format that is supported by standard annotation tools (such as ANVIL). With regard to traditional motion capturing the following main differences can be summarized as follows:

Aspects	**Motion Capturing**	
	Traditional	**Rigid bodies**
Preparation (w/o subject)		- Build rigid bodies
Preparation (with subject)	- Attach 18 markers per user - Map markers to the body parts	+ Attach 8 rigid bodies per user
Comfort for subjects	+ More comfortable	- Less comfortable
Stability of tracking	- Markers lost easily - Once marker is lost, the system doesn't know the position of that body part until the post processing	+ Set of 5 markers more stable to track + Once rigid body is lost, it can be automatically reassigned to each body limb by the system
Data saving	- After recording. Time consuming	+ Real-time
Post processing	- Map marker (that got lost) to the corresponding body part	+ None. Rigid body is always attached to a specific body part

A first evaluation of the setup suggests that the use of rigid bodies does not create more an unauthentic situation than do video recordings.

Next steps consist in further evaluating the impact of the rigid bodies on the user's conduct, and we aim to establish automated ways of detecting typical motions to allow for more automated ways of corpus annotation.

6. Acknowledgements

The authors gratefully acknowledge the financial support from the Cluster of Excellence "Cognitive Interaction Technology" and the EU-funded project "iTalk".

7. References

Battersby, S.A., Lavelle, M., Healey, Patrick G.T. & McCabe, R. (2008): *Analysing Interaction: A comparison of 2D and 3D techniques.* In: Proceedings Workshop on Multimodal Corpora (LREC 2008), 73-76.

Brüning, B. (2008): *Entwicklung eines Motion Capture Recorders für einen virtuellen Agenten auf der Basis eines optischen Trackingsystems*, Bielefeld University: Diploma thesis.

Brüning, B., Latoschik, M. E., Wachsmuth, I. (2008), *Interaktives Motion Capturing zur Echtzeitanimation virtueller Agenten*, In VRAR.

Chen, L., Rose, R. T., Qiao, Y., McNeill, D., & Harper, M. (2006). *VACE multimodal meeting corpus*. In S. Renals & S. Bengio (Eds.), Machine learning for multimodal interaction. (pp. 40-51). Heidelberg: Springer.

Dourish, P. (2001): *Where the Action Is*. Foundations of Embodied Interaction. MIT Press.

Kipp, M. (2010) *Multimedia Annotation, Querying and Analysis in ANVIL*. In: M. Maybury (ed.) Multimedia Information Extraction, Chapter 19, MIT Press.

Kuzuoka, H., Pitsch, K., Suzuki, Y., Kawaguchi, I., Yamazaki, K., Kuno, Y., et al. (2008). *Effects of restarts and pauses on achieving a state of mutual gaze between a human and a robot.* In CSCW 2008, 201-204.

Luff, P., Pitsch, K., Heath, C., Herdman, P., & Wood, J. (2009). *Swiping paper and the second hand: Mundane artefacts, gesture and collaboration.* Journal of

Personal and Ubiquitous Computing, 213-224.
Nishida, T. (2007). *Conversational informatics: An engineering approach*. Wiley.
Pitsch, K., Kuzuoka, H., Suzuki, Y., Süssenbach, L., Luff, P., & Heath, C. (2009). *"The first five seconds". Contigent stepwise entry into an interaction as a means to secure sustained engagement.* In RO-MAN 2009, 985-991.
Pitsch, K., Vollmer, A.-L., Fritsch, J., Wrede, B., Rohlfing, K., & Sagerer, G. (2009). *On the loop of action modification and the recipient's gaze in adult-child-interaction.* In GESPIN 2009. Poznan, Poland.
Rohlfing, K., Fritsch, J., Wrede, B., & Jungmann, T. (2006). *How can multimodal cues from child-directed interaction reduce learning complexity in robots?* Advanced Robotics, 20(10), 1183-199.
Spong, M. W., Hutchinson, S., Vidyasagar, M. (2006). Robotik Modeling And Control, Wiley.
Vollmer, A.-L., Lohan, K., Fischer, K., Nagai, Y., Pitsch, K., Fritsch, J., Rohlfing, K.J. & Wrede, B. (2009): *People Modify Their Tutoring Behavior in Robot-Directed Interaction for Action Learning*. In. IDCL 2009.