PAMOCAT: Automatic retrieval of specified postures

Bernhard Brüning[1], Christian Schnier[2], Karola Pitsch[2], Sven Wachsmuth[1]
CITEC Central Lab[1], Applied Informatics Group[2], University of Bielefeld
Universitätsstr. 20, 33615 Bielefeld, Germany
E-mail: bbruening@uni-bielefeld.de, cschnier@uni-bielefeld.de, karola.pitsch@uni-bielefeld.de,
swachsmu@techfak.uni-bielefeld.de

Abstract

In order to understand and model the non-verbal communicative conduct of humans, it seems fruitful to combine qualitative methods (Conversation Analysis) and quantitative techniques (motion capturing). A tool for data visualization and annotation is important as it constitutes a central interface between different research approaches and methodologies. We have developed the pre-annotation tool "PAMOCAT" that detects motion segments of individual joints. A sophisticated user interface enables the annotating person to easily find correlations between different joints and to export combined qualitative and quantitative annotations to standard annotation tools. Using this technique we are able to examine complex setups with three persons in tight conversation. A functionality to search for special postures of interest and display the frames in an overview makes it easy to analyze different phenomena in Conversation Analysis.

Keywords: Motion capturing, Motion segmentation, annotation, posture retrieval, Motion decomposition.

1. Introduction

Despite important progress in the field of human-robot and human-agent interaction, robotic communication skills are still far from the smoothness of the social behavior of humans in natural conversation. In order to build more appropriate interaction models both – human-human and human-robot interaction scenarios – need to be analyzed and understood in detail, so that results can be fed back into the model. To do so, researchers currently begin to link qualitative sequential analysis of videotaped interaction data with quantitative approaches based on Motion Capture data, so that an in-depth understanding of interaction procedures can be combined with quantifiable three-dimensional measures of body motions (Pitsch et al. 2010). In order to carry out such combined analyses not only conceptual issues need to be discussed but also novel tools for supporting the visualization and analysis of the different types of data are required. Existing annotation software, such as ELAN or Anvil, has recently started to integrate facilities for displaying time series data. These tools allow for linking text annotations with segments of digital media files. ELAN is specialized on Audio and Video media data and provides forms of automatic annotation especially for audio signals. Anvil is additionally able to display the motion of a single person specialized on the plot from the axes of the position, velocity, acceleration, and a trajectory visualization. However, in its current version the ability to handle data from multiple participants is missing and it only offers limited support for motion analysis. Our pre-annotation tool PAMOCAT addresses these gaps: It is able to deal with data from multiple participants, to show their skeletons and corresponding motion, and to highlight motion activity for each Degree of Freedom (DOF) separately so that quick access to specific motion activities of a particular joint is possible. In particular, it allows to both visualize and analyze three-dimensional Motion Capture data and to export automatically generated annotations to existing annotation software such as ELAN. Basing upon research that has shown the importance of body movements for turn-taking (Mondada 2007) and in particular the precise localization of pre-turn-initial pointing gestures in the local environment (Schnier 2010), we investigate in this paper a first attempt at finding such interactionally relevant body postures on 3D-data (lean back, pointing, crossed arms, looking at co-participants).

2. PAMOCAT: Pre-Annotation tool for visualizing and analyzing motion capture data

We have developed a tool – "PAMOCAT – Pre Annotation Motion Capture Analyze Tool", to pre-annotate motion capture data. It gives an overview at which point in time the information recorded for the individual joints changes. The main window shows the 3D visualization from the recorded participants (cf. figure 3d) with additionally loaded 3D objects defining reference points for the recorded interactions, so that the motion can be analyzed in relation to it. Afterwards, the annotator is able to see the recorded data from any position. As the viewing direction can beeasy adjusted online, is not necessary to simultaneously inspect several videos from different angles. A window (figure 1 b) presents an overview of all DOF from all joints for a selected person which shows the motion sequences for each joint separately as key-intervals (chapter 3). One key-interval is represented by a horizontal line with a green point at the beginning and a red point at its end. Below is a plot of the corresponding angle, speed and acceleration (c). A permanently visible synchronized view of the recorded videos completes the screen. At the bottom is a slider that allows the user to move in time (figure 1 h). The annotation widget (figure 1 f) allows to manually add information or to edit the automatically

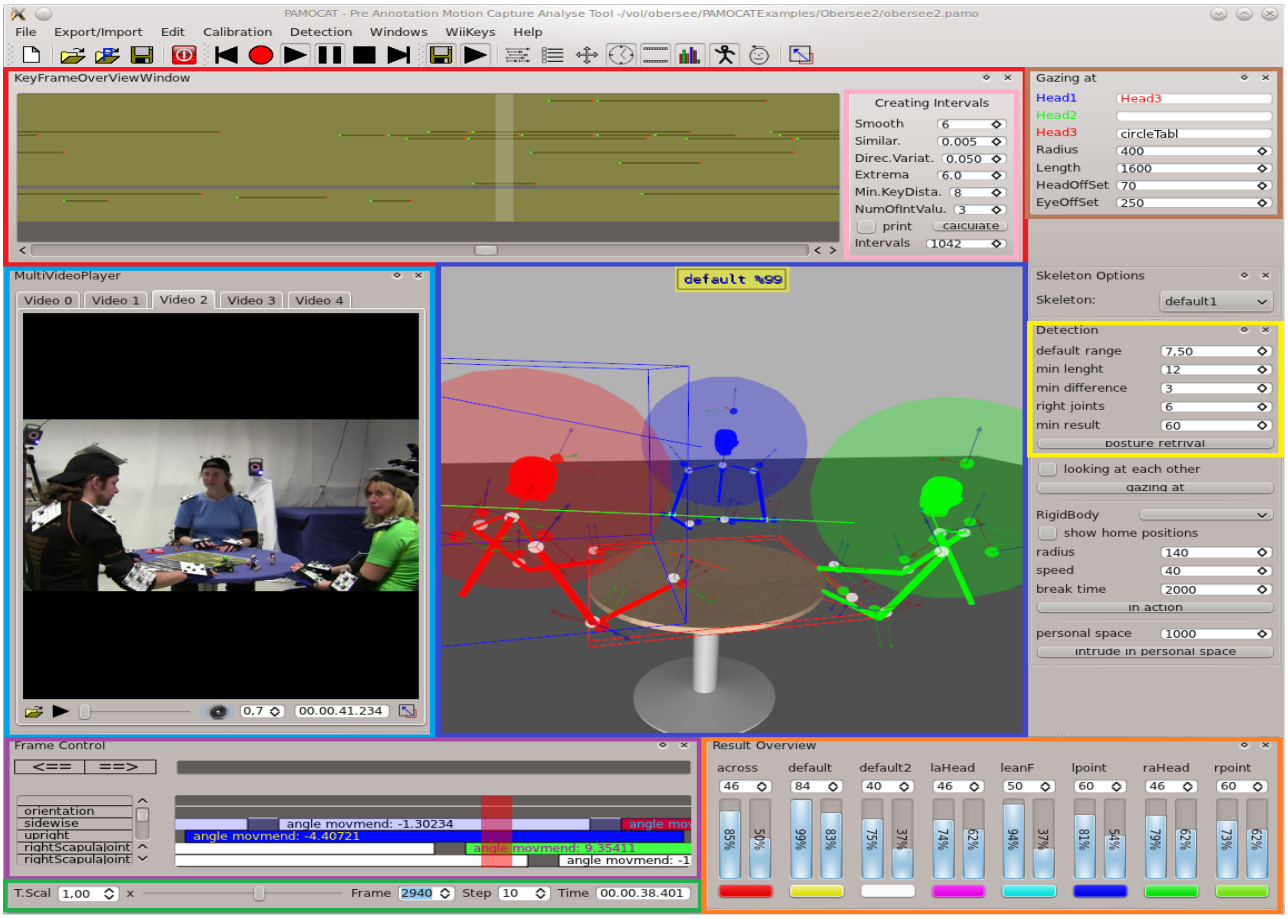

Figure 1 GUI with VideoPlayerFunctions (a), KeyIntervalOverViewWidget (b), DetectionEdit (c), MoCapView (d), Annotation Widget (e), KeyIntervalCalculationParameters (f),TimeSlider (h) GazingAtViewWidget (i) and ResultOverViewWidget (j)

generated or previously added information. In the following there will be a more detailed description on these features.

2.1 Free choice of view position on the motions

The posture assumed by a participant can sometimes be difficult to inspect from a particular perspective. To be able to perceive every detail the annotator can use normal 3D viewer navigation (like in other 3D software normally used) or a walking through mode with a Nintendo Wiimote controller. Additionally, we can use a stereo 3D visualization for a good immersion in the virtual world on the recorded motions. The annotator is better able to estimate the distance for each recorded participant and each joint in relation to the rest of the body.

2.2 Kinematic of skeleton with rigid bodies

We implemented a skeleton representation with inverse kinematics to calculate the joint angles and the kinematics to represent the skeleton (Brüning 2008). The rigid bodies are visualized with a coordinate system by arrows pointing in all three dimensions; the skeleton can be visualized with a kinematic skeleton or by links between the rigid bodies. It is possible to show to the rigid bodies the label names and geometry like a rigid body (cf. figure 2).

2.3 Trajectories of the body parts

As gestures and body motions are ephemeral phenomena, it is helpful for the analyst to visualize specific motion trajectories (cf. Pitsch et al. 2009). Our software is able to create such motion trajectories (see figure 3), either for all or for selected rigid bodies, in selected time intervals or during a specific time span. Using this feature we can easily see the interaction area of each joint. The density of the created trajectories (created to analyze this over the whole time span) shows the areas where the home positions of each joint are. The whole trajectory represents the area of the interaction space.

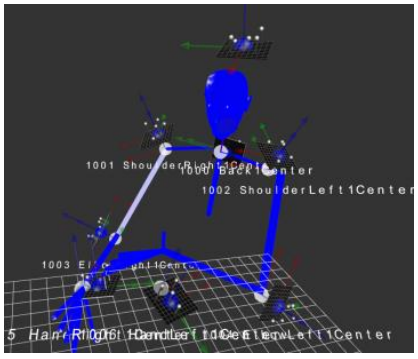

Figure 2 Skeleton with the related rigid bodies and labels

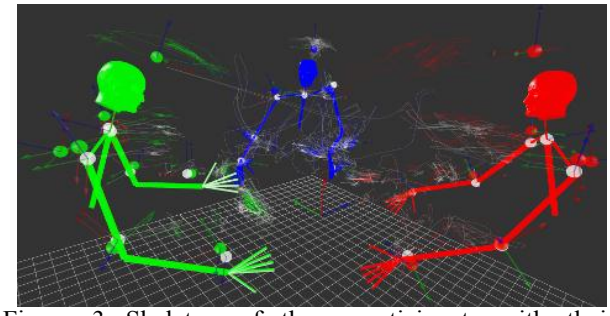

Figure 3 Skeleton of three participants with their Trajectories representing the interaction area of each

2.4 Parallel inspection of video recordings and motion capture data

The motion capture data and the video data are synchronized. The number of videos is not limited by the software, other videos can be switched on by a mouse click. With a GUI element called ”TimeShiftSlider” there is a free control over the available time interval (see figure 1). To find specific constellations there is the option to change the playback speed of the motion capture data, so a record of 20 minutes could be inspected for example in only 2 minutes (“TimeScalFactor” x10). This factor can be adjusted in 0.1-intervals, depending on the situation in the records. If the wanted phenomena are found they can be analyzed in detail in slow motion up to a TimeScaleFactor of 0.1.

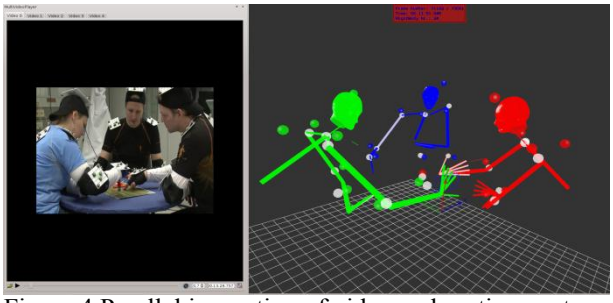

Figure 4 Parallel inspection of video and motion capture

2.5 Ability of analyzing the recorded motion in relation to a virtual environment of the real scenario

The annotator is able to retrieve the information where the recorded person is orienting to on the table or to another co-participant. Not only the motion itself is of interest in some cases, it could also be that the motion in relation to an object is of interest. For another example we conducted a study in a local arts museum, where the motion was related to more than one artwork that were placed in the recording area (Pitsch 2011). In this study there was an interacting robot that reacted depending on how close the participants came to the robot that gave explanations related to the art. The head of the participants and of the acting robot were tracked. To be able to analyze the motion in relation to the environment, we modeled the recorded area (one room with the artwork) and loaded it into the 3D virtual visualization together with the motion of the participants and the robot. Thus, the annotator is able to see the motion of the recorded participant in the virtual environment from any view point (with real depth information through 3d stereo). The information when the recorded participants are oriented towards the art or towards the robot is now automatically available for further analysis.

2.6 At which point in time does motion activity occur?

As shown in figure 5a (the key interval overview widget) the tool gives an overview at which points in time motion activity occur. With this GUI element it is easy to see which participant is mostly active at which time. With a plot and a decimal display of the angle, speed and acceleration, detailed information of the selected joints is available. The automatic identification of motion activity allows to detect relevant segments on a larger corpus without the need of identifying them manually. In the case that the annotator is searching for activity at some particular DOF, for example head orientation, he is easily able to select the joint. The selected joint is highlighted by a blue transparent line (cf. figure 5a) and it is possible to scroll with the time shift slider swiftly to all frames with activity. The annotator can now swiftly find the key interval of interest containing the relevant information of activity for each joint. The key intervals represent joint activity to an extreme in speed (cf. figure 6b), and from an extreme to no joint activity (more details section 3). When the head orientation changes from the home position to move to the right side, the software will create two key intervals.

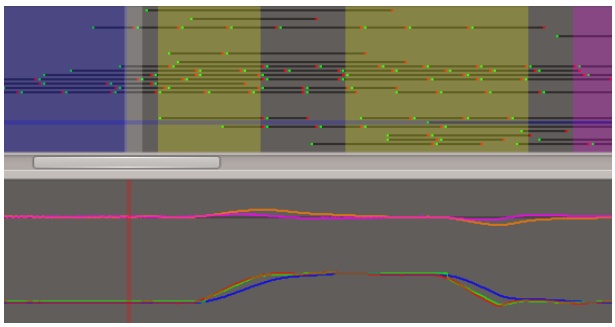

Figure 5 (a) Key-interval over view, the blue transparent line is a single selected DOF and detected posture highlights in colors (b) below there is the additionally plot of the angle, speed and acceleration for a selected single DOF

2.7 Support and exchange of data

To be useful for as many researchers as possible, we directed the development of the tool to be wide spread for exchange of data. The tool was development on two difference optical infrared tracking-systems, it started on the ART Tracking-system and is now also usable with the Vicon Nexus Tracking-system. Depending on the situation and environment a different tracking-system may be chosen for better results. PAMOCAT – as mentioned before – is not the only annotation tool, and is seen as a pre annotation tool with many abilities in the field of motion capturing, leaving video and audio signal

analysis to tools like ANVIL and ELAN. Therfore we have an ELAN and an ANVIL exporter with import and export functionality.

3. Joint motion decomposition using / defining key-intervals

In order to make the motion easy to annotate and to detect labeled motion sequences automatically, we need to decompose natural motion. To do so, we decompose the human motion into key-intervals. A key interval belongs to one DOF. It consists of a starting time, a length, a starting angle, and an ending angle. To decompose the motion from the entire skeleton, the concept of key interval is used. Each single DOF is individually analyzed with regard to speed and acceleration to reduce the values that have to be compared by the analysis during labeling (for example not all values of a shoulder joint with 3 DOF have to be compared in the case that one DOF contains an active key-interval). Let's assume a use case, where the annotator's interest is, for example, only focused on the participant's head orientation. He can now easily find a time frame where this DOF is active. In case of a similar speed over a number of frames, the angle information is stored in a key interval (cf. Figure 6).

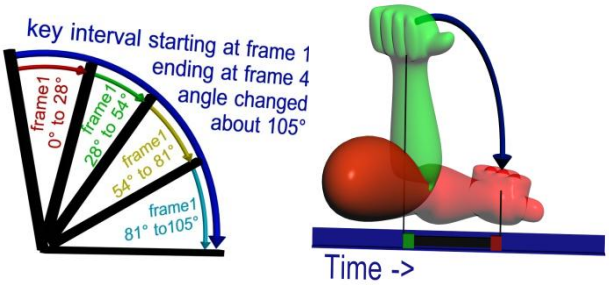

Figure 6 (a) A key-interval saves the same information of four frames in the case the speed is similar (b) key intervals activity in the elbow joint starting at the green position and ends in the red position after some time.

4. Posture learning and retrieval

Posture retrieval is a frequent operation in data annotation and analysis. By looking at points in time where a posture reappears, it can be examined in which situations a gesture or posture is used, how many probands reacted with the same gesture, or what different meanings special gestures could have. This procedure constitues an enormous advantage for the analytical work of Conversation Analysts: (i) The operation saves a lot of time so that it is more easy to verify findings on a broader corpus. (ii) The detection of one particular posture is more precise because the integrated search engine of PAMOCAT allows to fix the key characteristics of one particular posture. Therefore, the user marks a specific posture in the corpus data and automatically retrieves other points in time with a similar posture that need to be checked.

4.1 Technical basic background

4.1.1 Learning

To learn the postures we need some examples from differently sized human probands that are manually annotated. The joint limits from these postures need to be trained, and the relevant joints need to be discovered. If some person is pointing with one arm, the posture of the other arm is not so important. Which joints are relevant depends on the situation. If the probands are standing the arm would typically hang down, but if they are gathered around a table talking, one of the probands might support his upper body or maybe his head with his/her arms. As a result there is a lot of variation which is not relevant for the pointing gesture. Here the learning of a posture is an iterative process over all annotated postures and all joints of the skeleton. In the case that the joint value is outside of the limits of the joint, this joint will update the size (make the range of the joint bigger) of the range depending on the deviation from the limits. In case the joint value is inside the valid range the limits will be updated depending on its deviation from the center (make the range of the joint smaller).

4.1.2 Retrieval

The searching for the postures comes down to checking if all joint values are in the valid ranges learned before. A similarity score is calculated by weighting each joint by the importance coefficient. To decide which posture of the previously annotated ones is most similar to the current posture, the maximum similarity score of all previously learned postures is calculated. In case that a posture is not already annotated the value will be much smaller, because joints would be out of the valid ranges from the learned postures.

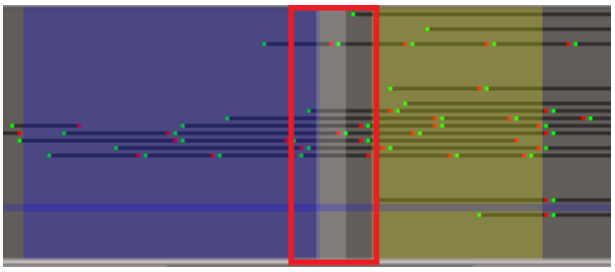

Figure 7 key-interval overview widget and the posture highlights (the colors are set default and can be later chosen for each posture)

4.2 GUI embedded functionality

In order to be useful our self-implemented posture retrieval is integrated in the pre-annotation tool PAMOCAT, to have a userfriendly control over this functionality in an advanced motion capture analysis tool. The annotator can manually select some training examples for a posture (see the dialog in Figure 8), and is presented with an overview of frames with similar postures. The annotator can change the starting range of

Figure 8 posture add dialog to train postures and to affect the importance (angle or range) of the joints depending on the posture

all joints and add later posture for training. The control which joint has which importance factor for a posture is optionally configurable. The visualization of the retrieval is a real time calculated similarity score that is displayed in an overlay at the top on the motion capture view (cf. figure 1d). For a general overview which posture was classified over the hole records each detected posture is colored in the frame with a color (cf. figure 7). The color can be be chosen manually in an result overview area (cf. figure 9). This contains all saved postures to detect, with the name and the classification result for all 3 different features (correct joints, distance in angles to the basic posture, are the important joints for the posture correct). Results are shown in figure 9 for 8 different postures with all results for the current frame (highlighted in white in figure 7). To optimize the results there are three parameters for variation, that influence the results of interest. The first is to filter only the postures that are recognized for a certain number of frames. The second parameter states that the classified result must be better than the others and as a defined value. The third parameter is that the correct joint value must be above as a defined limit.

5. Verification

The functionality of the posture search was verified by a manual annotation of the automatically detected postures. It was a verification of one person from three participants in a conversation. The record from our database (15 records) from the "Obersee" scenario (Pitch 2010) we identified different postures that are relevant to the

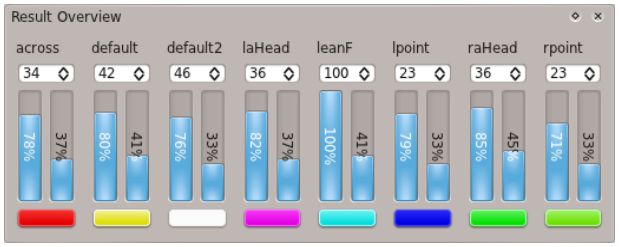

Figure 9 detection result live view is showing all results for the chosen 8 postures (across= arms crossed, default 1 and 2 = two differente resting posture, laHead = left arm to head, raHead = right arm to head, leanF = lean front, lpoint = left arm pointing gesture, rpoint = right arm pointing gesture).

conversation analysis scenario. These are by default (doing nothing), "lean back", "pointing" and "crossed arms" gesture. Some interesting lower level postures are "looking at subject B or C" and "looking at the table". As seen in the table 1 the detection and classification works very well (means her orientation was to other subject or to the table). It displays the information at which point in time there could be relevant information without losing closely related postures.

Posture	**Total**	**Classification**
Looking right	20	20
Looking left	25	25
Looking table	24	24
Default 1	19	16 + 3 default 2
Default 2	25	3 + 22 default 1
Pointing right arm	18 + 3 found additionaly	21
Pointing left arm	1	
Lean front	5 + 2 found additionaly	7
Right arm to head	4 + 4 found additionaly	8
left arm to head	1	1

Table 1: First verification

For the verification we additionally tested two very similar postures (default 1 and default 2) if they could be distinguishable from each other. The result is that the pre verification person (person who searched for postures to test manually) could not annotate the postures exactly (very similar postures), and the classification of the software was correct. Interesting was that the posture retrieval found additional frames in that a small similar postures was.

6. Conclusion

First results of the posture retrieval indicate that the added functionality provides a very useful input for annotation and analysis of motion capture data. As a next step, we aim at the search for more complex posture patterns that define gestures in different contexts. A comprehensive

search engine for familiar postures in everyday face-to-face communication builds a basal prerequisite for the verification of qualitative results of analyses and links both approaches – qualitative and quantitative – in a useful way.

7. Acknowledgements

This research is supported by the Center of Excellence 'Cognitive Interaction Technology' (CITEC, EC277), the project C5 'Alignment in AR-based cooperation' in the SFB/CRC 673 and the Volkswagenstiftung/Dilthey Fellowship 'Interaction & Space'.

8. References

Auer, E., Russel, A. Sloetjes, H., Witternurg, P., Schreer, O., Masneri, S., Schneider, D. & Toepel, S., 2010. *ELAN as flexible annotation framework for sound and image processing detectors*. In N. Calzolari, B. Maegaard, J. Mariani, J. Odjik, K. Choukri, S. Piperidis, M. Rosner, & D. Tapias (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation (pp. 890-893). LREC.

Brüning, B., Latoschik, M. E., Wachsmuth, I. 2008, *Interaktives MotionCapturing zur Echzeitanimation virtualler Agenten* Proceedings VRAR2008.

Brüning, B., Schnier, C., Pitsch, K., Wachsmuth, S., 2011. *Automatic detection of motion sequences for motion analysis*, ICMI Workshop multimodal corpora.

Heloir, A., Neff, M., Kipp, M., 2010, *Exploiting Motion Capture for Virtual Human Animation: Data Collection and Annotation Visualization.* In Proceedings of LREC Workshop on "Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality", ELDA.

Kendom, A., 2004. Gesture: *Visble Actions as Utterance*, Cambridge University, p.158.

Mondada, L., 2007. *Multimodal resources for turn-taking: pointing and the emergence of possible next speakers*. In Discourse Studies, 9, 2, pp. 194-225.

Pitsch, K., Brüning, B., Schnier, C. and Wachmuth, S., 2010. *Linking Conversation Analysis and Motion* Capturing: "How to robustly track multiple participants?". In Proceedings Workshop on Multimodal Corpora, LREC.

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 (Gesture and Speech in Interaction) 2009, Poznan, Poland, 7 pages.

Pitsch, K., Wrede, S., Seele, J.Ch., Süssenbach, L. (2011): *Attitude of German Visitors towards an Interactive Art Guide Robot*. HRI 2011.

Schnier, C., 2010. Turn-Taking: *Interaktive Projektionsleistungen über Kinesische Displays*. Masterarbeit Universität Bielefeld, p. 93