

# Automatic detection of motion sequences for motion analysis

Bernhard Brüning  
CITEC Central Lab  
University of Bielefeld  
Germany 33615 Bielefeld  
+49 521 106-67258  
bbuening@uni-bielefeld.de

Christian Schnier  
Applied Informatics Group  
University of Bielefeld  
Germany 33615 Bielefeld  
+49 521 106-12229  
cschnier@uni-bielefeld.de

Karola Pitsch  
Applied Informatics Group  
University of Bielefeld  
Germany 33615 Bielefeld  
+49 521 106-12240  
karola.pitsch@uni-bielefeld.de

Sven Wasmuth  
CITEC Central Lab  
University of Bielefeld  
Germany 33615 Bielefeld  
+49 521 106-2937  
swachsmu@techfak.uni-bielefeld.de

## ABSTRACT

In order to understand and model the non-verbal communicative behavior of humans, qualitative techniques, such as Conversation Analysis, and quantitative techniques, such as 3D motion capturing, need to be combined. Although there has been some recent progress in annotation tools like ELAN or Anvil, there is still a lack of appropriate tool support that enables a concise simultaneous access to both types of data and that shows the relationship between them. Within this work, we present a pre-annotation tool that takes the results from off-the-shelf optical tracking systems, automatically fits an articulated skeleton model, and detects motion segments of individual joints. A sophisticated user interface easily allows the annotating person to find correlations between different joints, analyze the corresponding 3D pose in a reconstructed virtual environment, 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 or largely unconstrained engagement situations of humans and robots.

## Categories and Subject Descriptors

Multimodal Corpora Analysis Tool

## General Terms

Human Factors, Languages, Theory, Verification.

## Keywords

Motion capturing, Motion segmentation, annotation.

## 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 interactional procedures could be combined with quantifiable three-dimensional measures of body motions [1]. 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 [2] or Anvil [3], has recently started to integrate facilities for displaying time series data. ELAN and Anvil allow for linking text annotations with segments of digital media files. ELAN is specialized on Audio and Video media data and provides automatical 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 color highlighting trajectory visualization equals to the annotation color. However, in its current version the ability to handle data from multiple participants is missing and it offers only limited support for motion analysis. In this paper, we present our pre-annotation tool PAMOCAT that 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. To motivate our approach and to demonstrate how our tool could be integrated into a research cycle linking qualitative and quantitative methods, we will begin with a short example from the analysis of human-human interaction and the analytical issues that arise from it (section 2). Based on this, we will present our approach of robustly tracking multiple participants with motion capture technology (section 3), the basic ideas and user interface of our tool PAMOCAT (section 4) and explain some of its current analytical facilities (section 5). Specifically, we will introduce the notion of “key-intervals” as the basic concept of the tool. Finally, we will give some examples of how PAMOCAT could support data analysis (section 6) and will conclude with a short outlook regarding future work (section 7).

## 2. EXAMPLE: FROM VIDEO-BASED ANALYSIS TO MOTION CAPTURE DATA

In order to motivate our approach and the development of our tool, we begin with a short fragment from human-human interaction. We will reveal on the one hand analytical issues that arise when carrying out in-depth manual analysis of the participants’ interactional practices and show on the other hand the limitations of a video-based approach and how a corpus of combined video and motion capture data could help to overcome these limits. Let us consider the following short fragment, in which three participants in a semi-experimental setup are seated around a table and were asked to jointly plan a local recreation

area while manipulating a range of objects [1, 4]. Here, our analytical interest focuses on one particular aspect of the interactional organization: multimodal aspects of turn-taking and kinetic procedures of how to take the floor in multiparty conversations. Empirical investigation of similar situations has revealed that participants systematically use pointing gestures to objects in the local environment in order to announce and establish themselves as possible next speaker. Mondada [5] states that „[...] pointing gestures are precisely timed, being synchronized with the moment-by-moment organization of talk-in-interaction, with recipient oriented talk and bodily conducts, with appropriate arrangements of bodies and objects [...]“. Participants, who attempt to take the turn and position themselves as next speaker tend to bodily claim the floor before even starting to talk, and use as a systematic procedure „pre-initial turn pointing“[5]. Taking these findings further, we have been able to show that the precise localization of such pointing gestures in the interaction space matters [4]. The following fragment sheds some light on this phenomenon and reveals that a precise knowledge about hand positions, arm positions and body alignment of the participants is of particular importance:

```

01 A: |JA dann würden sie das ja noch verSCHLIMMERN |
    B-act: | ..... |pG-RH |
    A-gaz: | .....@C..... |
    C-gaz: | .....@A..... |
           *1

02 B-act: | .....pG (RH)..... |
    B: |dann bauen wa=n !HIER! hin..... |
    C-gaz: |...~B (RH).....@map..... |
    A-gaz: |...~C.....~B (RH)..... |
           *2a+b *3

03 C: |JA:- das wär auch ne option;|
    C-gaz: | .....@map..... |
  
```

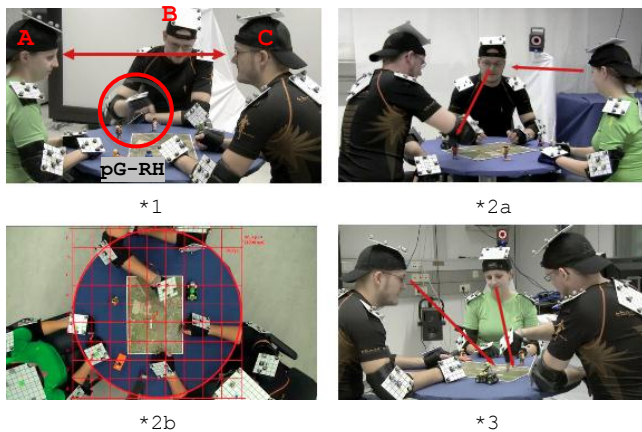


Figure 1 Localization of pointing gestures

Prior to this fragment, participants A and C enter into a stable two-party conversation for a longer stretch of time. Their body posture is aligned and they are mutually oriented to each other (cf. \*1). At this point in time, participant B assumes a status as observer. Then, in line 01 participant B projects to again enter the discussion and to take the floor by a pre-initial turn pointing (shaded gray). As pointing practices and turn-taking practices are deeply embedded, it can be shown that his pointing orients to a transition relevance place (TRPs mark places in the current turn where turn exchanges occur because of a completeness of multimodal cues (cf. Sacks et al. 1974)) [8] at the end of A’s turn.

What are the interactional consequences of B’s pointing action? Our analysis reveals that the structural order of the ongoing course of interaction is affected by the localization of the pointing gesture. In this case B’s pointing takes place in a shared space of action between him and his co-participant C. C reacts to B’s action by interrupting his dialogue with A and orients to the new attention focus represented by B’s right-hand pointing gesture (B(RH))(cf. 02 shaded green, \*2a+b). Afterwards participant A treats C’s digressing view as a relevant orienting device to change her own focus of attention and also follows B’s pointing gesture (cf. \*3). We observe that the precise timing and spatial placement of the pointing gesture seems to be consequential for the question which participant will comment on the pointer’s action. If the pointer does not simultaneously address its turn to a specific co-participant, that participant tends to react who firstly re-orientates to the new attention focus (cf. 03). In a first analytical step, using only the video data, we overlayed a position mask over the video data (cf. 2b) and manually annotated the hand positions of each participant at any point in time. This allowed us, for single cases, to specify pointing gestures with regard to the local occurrence. However, if we want to use such findings for interactional modeling in e.g. human-robot-interaction, we need to describe the different aspects of this interactional practice in greater detail and examine them over a large corpus basis:

- What are typical interaction spaces for joint action, in which participants e.g. collaboratively manipulate objects?
- To which extent does the speed and precise trajectory of the hand movements matter in the described set of practices?
- What are the participants’ global home positions? When do they leave it and return to it?

To answer these questions, motion capture data describe the kinetic aspects with greater precision and are able to provide measures for their speed, acceleration and posture. An automated motion analysis would allow to detect certain types of gestures or activities over a larger corpus basis. Also, to view and inspect the recorded data from any position and to display the precise trajectories of the participants are a promising advantage for the analytical investigation of this phenomenon.

### 3. HOW TO ROBUSTLY TRACK MULTIPLE PARTICIPANTS

In a first step we needed to find a way to record multiple participants over an extended period of time in a way that allows for time-efficient post-processing. Normally, motion capture data has to be revised, which is a very time intensive work. It is only practicable for short motion sequences. Thus, we had to initially focus on robustly tracking multiple participants over an extended period of time [1]. In this case we have to deal with motion in more than one direction. Some experiments show that depending on the camera position the motion labeling task is significantly easier or harder [6]. Furthermore, in video the real joint positions are difficult to define and the annotating person or automatic algorithms needs to deal with video problems like noise, or finding and labeling the limbs. Since motion capture systems are getting more and more common, many of these problems can be bypassed so that the main issues are on the automatic pre-labeling which typically costs a significant post processing effort. We use a

commercial optical tracking-system (Vicon MX) which is based on infrared cameras. Instead of the usual individual markers, which are attached to the participants' body, we use the system with so-called rigid bodies (see figure 2). A rigid body is a pattern of – in our case – a set of 5 single markers in a unique pattern mounted on a base plane. This configuration differs for each rigid body. These rigid bodies allow to individually track the limbs of multiple participants, so that – even if a marker cannot be detected at some moment – it can be identified as belonging to a certain joint once it reoccurs, so that no post labeling is required. However, in the usual case of a single marker tracking system, the markers would need to be manually assigned to the limbs of each participant at the beginning of the recording and/or once the system has lost the marker during the recording. Under such conditions, the typical post-processing time is nearly a factor of 10 times longer as the recorded time for each recorded participant.

For our rigid bodies, the size of the plane depends on the camera distance, the size of the recorded interaction area and the number of cameras. We significantly improved a previous planar design of the rigid bodies [1]. To make the size of their base plane smaller, we built a 3D pattern instead of a 2D pattern shown in figure 2.

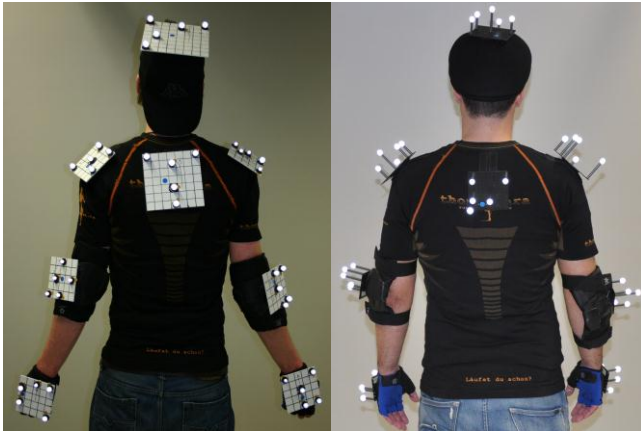


Figure 2 (a) 2D rigidbody placed at the participant

(b) 3D Version

In summary, the rigid body tracking method allows – in direct comparison to the common single-marker based tracking – to track multiple participants over a longer period of time without the need of pre- and post-labeling. This builds an essential prerequisite for conversation-analytical research with the following presented annotation tool “PAMOCAT”.

#### 4. PAMOCAT: PRE-ANNOTATION TOOL FOR VISUALIZING AND ANALYZING MOTIONCAPTURE DATA

We have developed a tool – “PAMOCAT – Pre Annotation Motion Capture Analyze Tool”, to pre-annotate the motion capture data. It gives an overview at which point in time the information recorded for the individual joints changes, it is able to add information about the joint angle difference, speed, acceleration and movement in relation to the world and it gives a plot of joint angles combined to joints from other participants. In our tool we calculate the orientation of the skeleton joints from all participants in real time. Afterwards, the annotator is able to see

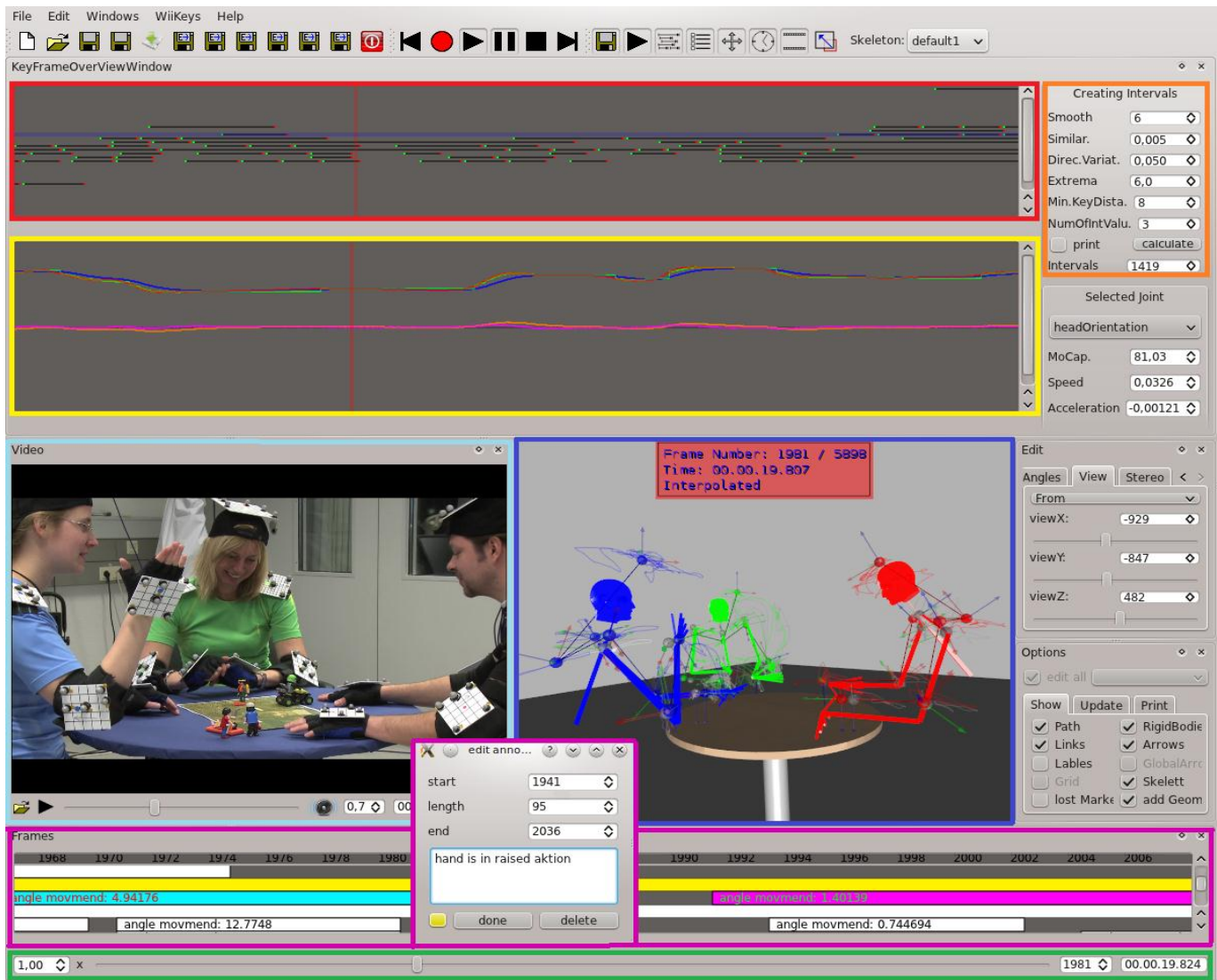
the recorded data from any position. It is not necessary to simultaneously watch many videos from different directions. Instead, the viewing direction can easily be adjusted. Additionally, a window presents an overview of all DOF (Degree of Freedom) from all joints for a selected person which shows the motion sequences for each joint separately, and a GUI element that renders the angle, speed and acceleration for one or more selected joints. An always visible synchronized view of the recorded videos completes the screen. The technical basis of the application is OpenSG2 and QT4. OpenSG2 is a library developed for clustered rendering as typically used in big VR installations. QT4 is a library to create GUI (Graphical User Interface) elements like buttons, combo box or load dialogs. The GUI consists of a main window and additional docking windows. The main window shows the 3D visualization from the recorded participants (see figure 3d) with additionally loaded 3D objects defining reference points for the recorded interactions. Below another docking window defines a slider that allows the user to move in time (h), so that the user is able to scroll very fast through the frames of interest. The special regions of interest can be represented in a separate docking window, a so-termed key-interval-overview window (b). It displays all key-intervals found (chapter 5). One key-interval is represented by a horizontal line with a green point at the beginning and a red point at the end. Below is a plot of the corresponding angle, speed, acceleration and the reconstructed angle (c). In this plot only the values for one joint are shown; with the choice box each joint can be selected (see Figure 3 GUI). The annotation widget (f) allows to manually add information or to edit the automatically generated or previously added information. In the motion capture view the recorded motion can be put in relation to a 3D model of the environment, so that the motion can be analyzed in relation to it.

### 5. JOINT MOTION DECOMPOSITION USING / DEFINING KEY-INTERVALS

In order to make the motion easy to annotate and, in the future, to detect labeled motion sequences automatically, we need to decompose natural motion. To do so, we decompose the human motion into key-intervals.

#### 5.1 Key-intervals

A key interval belongs to one joint. 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. In the case of the upper body, the skeleton has 24 DOF, while the full body has 41 DOF. 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, in which 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 motion or DOF is active. There are 5 variables (see table 1) that can be adjusted for an additional detailed analysis from the resulting key intervals calculated with the default values. The algorithm for key intervals is separately applied on each DOF (or elementary joint). The result is a compressed motion.



In case of a similar speed over a number of frames, the angle information is stored in a key interval only once for each joint in the skeleton with these DOFs (see Figure 4). The figure shows the arm movement from a starting time to an ending time during an angle change. The movement of the arm in a closed-loop feedback could also be decomposed to one single key-interval. Normally the speed at the beginning rises slowly, reaches a maximum and then decreases again. This maximum is an interesting information for the annotator, so that the human motion gets decomposed into

**Figure 3** GUI with **VideoPlayerFunctions** (a), **KeyIntervalOverViewWidget** (b), **AngleView** (c), **MoCapView** (d), **Annotation Widget** and **Annotation Dialog** (e), **KeyIntervalCalculationParameters** (f) and **TimeSlider** (h)

**Table 1. Adjustable variables**

adjustable variables	description
Smoothing	smoothing factor to reduce noise
Speed threshold	defines the interesting speed
Acceleration threshold	that specifies the minimal acceleration distance to zero
Sliding window size	to detect if the signal is increasing or decreasing, or has a minimum or maximum
Minimal interval length	Define the minimal length of the intervals

more than only one key interval. These motions are now saved as a key animation. Typically, a key animation is used to define naturally appearing motions for virtual characters. The quality depends on the number of key frames. Here we are going the opposite way from natural motion to discrete events that can be further analyzed. Depending on the adjustable variables the motion becomes more or less compressed while losing more or less information. The default values are improved and recalculated to produce the best results for the recorded scenarios. It is a costly process in which different constellations from the adjustable variables are computed (one calibration for one recording scenario should be enough). The parameters can be adjusted through a direct visualization loop of the reconstructed key intervals. It is possible to visually compare a skeleton animated with the reconstructed motion against the original motion. Differences are further shown in an angle plot, the reconstructed motion in a green plot and the original motion in a red plot.

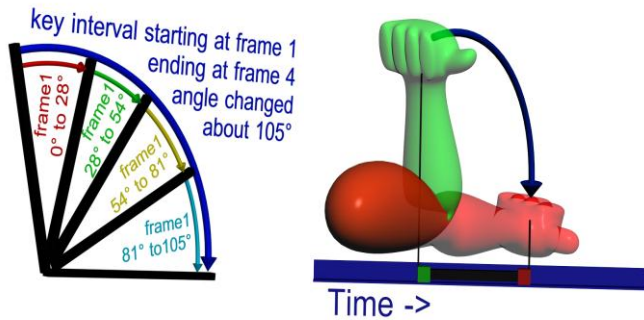


Figure 4 (a) A key-interval saves the same information of four frames in the case the speed is similar (b) key intervals activity in the elbow joint starting at the green position and ends in the red position after some time.

## 5.2 What are the advantages from key frame compared to trajectory analysis?

Instead of taking a look at the trajectories of single body parts and afterwards transforming or decomposing them into splines, we are analyzing the local activities of each joint transformed into key frame activities. The idea is that the local information is more important for the detection of a movement than the global trajectories. Additionally, less data needs to be handled in the local data representation. Each trajectory has 3 DOF for the position and 3 DOF for the orientation. The local skeleton representation has less than 6 DOF, typically, depending on the viewed skeleton parts (elbow to Hand only 2 DOFs, shoulder to Elbow 3 DOF, excluding the Hand itself has 6 DOFs), an important information reduction for a real time interaction system.

## 6. Using PAMOCAT in the research cycle

How can PAMOCAT be integrated into the research cycle? How can it support annotation and empirical analysis considering both qualitative and quantitative research? – In what follows we will give a short overview of the ways in which PAMOCAT is useful both for (i) visualizing data and (ii) analyzing data.

### 6.1 Visualization of data

#### 6.1.1 Trajectories of the body parts

As gestures and body motions are ephemeral phenomena, it is helpful for the analyst to visualize and materialize specific motion trajectories. Our software is able to create such motion trajectories, 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 (see section 2).

#### 6.1.2 Rigidbodies in relation to the real skeleton

The rigid bodies are visualized with a coordinate system through arrows; the skeleton can be visualized with a kinematic skeleton

or through links between the rigid bodies. An important feature in the visualization is, to know at which position the real skeleton joints are in relation to the rigid bodies (naturally the joint and rigid body positions are not equals). Additionally it is important to have the distance and related position of the hands in relation to the body.

#### 6.1.3 Free choice of the view position on the motion

The posture assumed by a participant is generally not well visible from every perspective. To be able to get 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 are using 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 single joint in relation to the rest of the body.

#### 6.1.4 Parallel inspection of video recordings and motion capture data

The motion capture data and the video data are synchronized. It is possible to change the play time speed (for a slow detailed analysis or a fast overview). 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.

## 6.2 Analysis of data

#### 6.2.1 At which time does some motion activity occur?

As shown in figure 3 (b) “the key interval overview widget” the tool gives an overview at which points in time there is motion activity. With this GUI element it is easy to see which participant is mostly active at which time, and might be the current speaker. With a plot and a decimal display of the angle, speed and acceleration there are detailed information of the selected joints available. With regard to the analytical example in chapter 2 the advantage becomes directly evident. The automatic identification of motion activity allows easily to detect relevant segments on a larger corpus without the need of identifying them manually.

#### 6.2.2 Where is activity at a particular joint?

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 gets highlighted through a blue transparent line 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 extremum in the speed, and from an extremum to no joint activity (this is of particular interest with regard to our second question in the introductory example (cf. chapter 2)). When the head orientation changes from the home position to move to the right side, the software will create two key intervals.

### 6.2.3 Detailed analysis which joints and/ or DOF are used in specific gestures

For a detailed analysis of performed gestures (e.g. a pointing gesture), the tool shows the sequences of each DOF to be able to understand the activity of every joint of the entire skeleton. There is a detailed view at which time which joint is active maybe in combination with other joint. It could be seen as a hierarchical (skeleton) description of the whole movement decomposed into sub movements down to the level of single DOFs. Especially the timing of the starting activity of a joint is highly relevant with regard to our introductory example. For example, in the case of pre-initial turn pointing (section 2) the timespan between the onset of the pointing gesture and the first verbal expression is very short and hard to observe in video data.

### 6.2.4 Add or edit automatically generated annotations manual

In the annotation area it is possible to add, delete or edit annotations (in normal case the automatically generated annotations), and to change the color to highlight special elements. The software is seen as a pre-annotation tool, because it provides useful structural hints for a semantically motivated final annotation. Other tools like Anvil and ELAN have integrated an advanced annotation area (allowing for zooming or scaling), but it seems to be important to be able to annotate directly.

### 6.2.5 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 looking at the table or is looking at another interacting partner. 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 was placed in the recording area [7]. 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 was 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 and the virtual environment from any view point (with real depth information through 3d stereo). The information when the recorded participants are looking at the art or at the robot is now automatically available for further analysis.

## 7. FUTURE WORK

The direction from the tool will go more into automatic annotation or pre-annotation. A major goal is that manual annotated motions should become automatically labeled, afterwards, or that you can create a schema of correlating joints.

For example you could tell the system to label a part of the motion as “pointing” if the head and hand are oriented nearly in the same direction. Or if there is a sinusoidal signal on one DOF like it would be during head shaking or head nodding. Developing directions additionally go into the area of real time detection of motions learned before. A system like a virtual agent or a robot then could be able to react on these motions from the participant in a human computer interaction.

## 8. ACKNOWLEDGMENT

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’.

## 9. REFERENCES

- [1] 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.
- [2] 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
- [3] 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.
- [4] Schnier, C., 2010. Turn-Taking: *Interaktive Projektionsleistungen über Kinesische Displays*. Masterarbeit Universität Bielefeld, p. 93.
- [5] Mondada, L., 2007. *Multimodal resources for turn-taking: pointing and the emergence of possible next speakers*. In Discourse Studies, 9, 2, pp. 194-225.
- [6] Parameswaran, V., Chellappa, R., 2003. *View invariants in Human Action Recognition*, In Proceeding Computer Vision and Pattern Recognition, IEEE.
- [7] Pitsch, K., Wrede, S., Seele, J.Ch., Süßenbach, L. (2011): Attitude of German Visitors towards an Interactive Art Guide Robot. HRI 2011.
- [8] Sacks, H., Schegloff, E. A. and Jefferson G., 1974. *A Simplest Systematics for the Organization of Turn Taking for Conversation*. In: *Language*, 50, pp. 696–735.