# Annotation of Human Gesture using 3D Skeleton Controls

**Quan Nguyen, Michael Kipp**

DFKI
Saarbrücken, Germany
{quan.nguyen, michael.kipp}@dfki.de

### Abstract

The manual transcription of human gesture behavior from video for linguistic analysis is a work-intensive process that results in a rather coarse description of the original motion. We present a novel approach for transcribing gestural movements: by overlaying an articulated 3D skeleton onto the video frame(s) the human coder can replicate original motions on a pose-by-pose basis by manipulating the skeleton. Our tool is integrated in the ANVIL tool so that both symbolic interval data and 3D pose data can be entered in a single tool. Our method allows a relatively quick annotation of human poses which has been validated in a user study. The resulting data are precise enough to create animations that match the original speaker's motion which can be validated with a realtime viewer. The tool can be applied for a variety of research topics in the areas of conversational analysis, gesture studies and intelligent virtual agents.

## 1. Introduction

Transcribing human gesture movement from video is a necessary procedure for a number of research fields, including gesture research, sign language studies, anthropology and believable virtual characters. While our own research is motivated by the creation of believable virtual characters based on the empirical study of human movements, the resulting tools are well transferrable to other fields. In our research area, we found virtual characters of particular interest for many application fields like computer games, movies or human-computer interaction. An essential research objective is to generate nonverbal behavior (gestures, body poses etc.) and a key prerequisite for this is to analyze real human behavior. The underlying motion data can be video recordings, manually animated characters or motion capture data. Motion capture data is very precise but requires the human subject to act in a highly controlled lab environment (special suit, markers, cameras), the equipment is very expensive and significant post-processing is necessary to clean the resulting data (Heloir et al., 2010). Traditional keyframe animation requires a high level of artistic expertise and is also very time-consuming. Motion data can also be acquired by manually annotating the video with symbolic labels on time intervals in a tool like ANVIL (Kipp, 2001; Kipp et al., 2007; Kipp, 2010b; Kipp, 2010a) as shown in Fig. 1. However, the encoded information is a rather coarse approximation of the original movement.

We present a novel technique for efficiently creating a gesture movement transcription using a 3D skeleton. By adjusting the skeleton to match single poses of the original speaker, the human coder can recreate the whole motion. Single pose matching is facilitated by overlaying the skeleton onto the respective video frame. Motion is created by interpolating between the annotated poses.

Our tool is realized as an extension to the ANVIL[1] software. ANVIL is a multi-layer video annotation tool where temporal events like words, gestures, and other actions can be transcribed on time-aligned tracks (Fig. 1). The encoded data can become quite complex which is why ANVIL offers typed attributes for the encoding. Examples of similar tools are ELAN[2] (Wittenburg and Sloetjes, 2006) and EXMARaLDA (Schmidt, 2004). Making our tool an extension of ANVIL fuses the advantages of traditional (symbolic) annotation tools and traditional 3D animation tools (like 3D Studio MAX, Maya or Blender): On the one hand, it allows to encode poses with the precision of 3D animation tools and, on the other hand, temporal information and semantic meaning can be added, all in a single tool (Figure 2). Note that ANVIL has recently been extended to also display motion capture data with a 3D skeleton for the case that such data is available (Kipp, 2010b).

In the area of virtual characters our pose-based data can immediately been used for extracting gesture lexicons that form the basis of procedural animation techniques (Neff et al., 2008) in conjunction with realtime character animation engines like EMBR (Heloir and Kipp, 2009). Moreover, empirical research disciplines that investigate human communication can use the resulting data and animations to validate their annotation and create material for communication experiments.

## 2. Human Gesture Annotation with a 3D Skeleton

Our novel method of gesture annotation is based on the idea that a human coder can easily "reconstruct" the pose of a speaker from a simple 2D image (e.g. a frame in a movie). For this purpose we provide a 3D stick figure and an intuitive user interface that allows efficient coding. The human coder visually matches single poses of the original speaker with the 3D skeleton (Figure 2). This is supported by overlaying the skeleton on the video frame and offering intuitive skeleton posing controls (Fig. 3). The system can then interpolate between the annotated poses to approximate the speaker's motion. Here we describe the relevant user interface controls, the overall workflow and how to export the annotated data.

**Controlling the skeleton** 3D posing is difficult because it involves the manipulation of multiple joints with multiple

[1] http://www.anvil-software.de

[2] http://www.lat-mpi.eu/tools/elan/

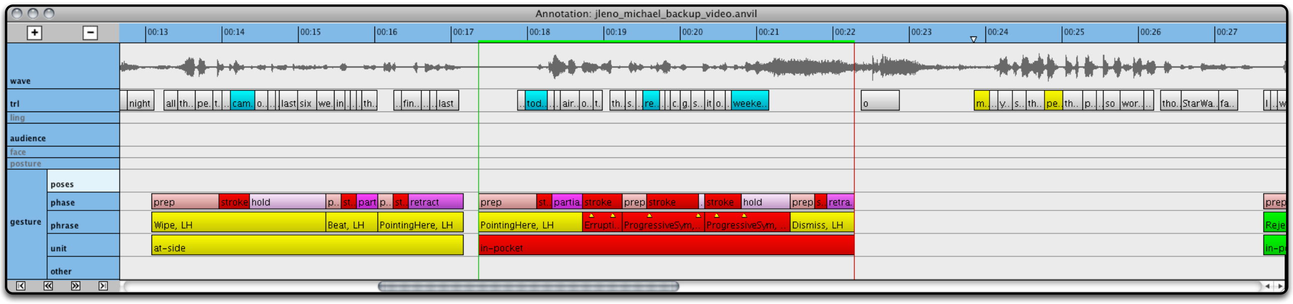
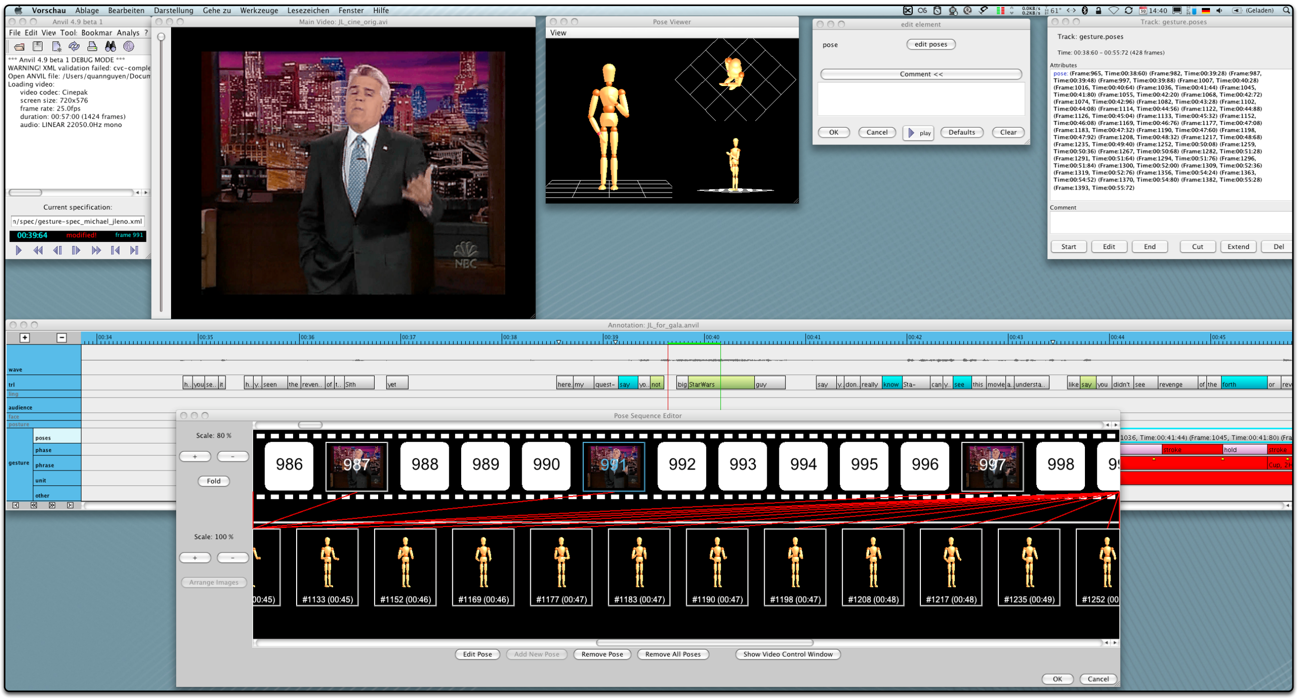


Figure 1: Regular annotations in ANVIL are displayed as color-coded boxes on the annotation board. Every annotation can contain multiple pieces of symbolic information on the corresponding event.



```
*** Anvil 4.9 beta 1 DEBUG MODE ***
WARNING! XML validation failed: cvc-comple
Open ANVIL file: /Users/quannguyen/Docum
Loading video:
     video codec: Cinepak
     screen size: 720x576
     frame rate: 25.0fps
     duration: 00:57:00 (1424 frames)
     audio: LINEAR 22050.0Hz mono
```

Current specification: n/spec/gesture-spec_michael_jleno.xml

00:39:64 modified! frame 991

Track: gesture.poses

Time: 00:38:60 – 00:55:72 (428 frames)

Attributes

pose: (Frame:965, Time:00:38:60) (Frame:982, Time:00:39:28) (Frame:987, Time:00:39:48) (Frame:997, Time:00:39:88) (Frame:1007, Time:00:40:28) (Frame:1016, Time:00:40:64) (Frame:1036, Time:00:41:44) (Frame:1045, Time:00:41:80) (Frame:1055, Time:00:42:20) (Frame:1068, Time:00:42:72) (Frame:1074, Time:00:42:96) (Frame:1082, Time:00:43:28) (Frame:1102, Time:00:44:08) (Frame:1114, Time:00:44:56) (Frame:1122, Time:00:44:88) (Frame:1126, Time:00:45:04) (Frame:1133, Time:00:45:32) (Frame:1152, Time:00:46:08) (Frame:1169, Time:00:46:76) (Frame:1177, Time:00:47:08) (Frame:1183, Time:00:47:32) (Frame:1190, Time:00:47:60) (Frame:1198, Time:00:47:92) (Frame:1208, Time:00:48:32) (Frame:1217, Time:00:48:68) (Frame:1235, Time:00:49:40) (Frame:1252, Time:00:50:08) (Frame:1259, Time:00:50:36) (Frame:1267, Time:00:50:68) (Frame:1282, Time:00:51:28) (Frame:1291, Time:00:51:64) (Frame:1294, Time:00:51:76) (Frame:1296, Time:00:51:84) (Frame:1300, Time:00:52:00) (Frame:1309, Time:00:52:36) (Frame:1319, Time:00:52:76) (Frame:1356, Time:00:54:24) (Frame:1363, Time:00:54:52) (Frame:1370, Time:00:54:80) (Frame:1382, Time:00:55:28) (Frame:1393, Time:00:55:72)

Comment

Figure 2: Our ANVIL extension allows to encode human poses with the precision of 3D animation tools. This information complements ANVIL's conventional coding which stores temporal and symbolic information.

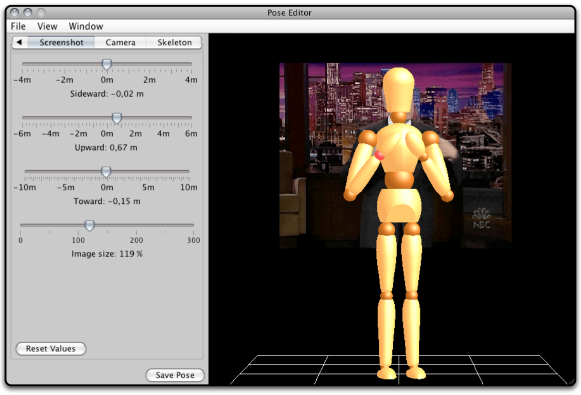
degrees of freedom. The two methods of skeleton manipulation are forward kinematics (FK) and inverse kinematics (IK). Pose creation with FK, i.e. rotating single joints, is slow. In contrast, IK allows the positioning of the end effector (usually the hand or wrist) while all other joint angles of the arm are resolved automatically. In our tool, the coder can pose the skeleton by moving the hands to the desired position (IK). The coder can the fine-tune single joints using FK, changing arm swivel, elbow bent and hand orientation (Figure 4). For IK it is necessary to define kinematic chains which can be done in the running system. By default, both arms are defined as kinematic chains, thus both arms can be manipultated by the user. The underlying skeleton can be freely defined using the standard *Collada*[3] format.

**Limitations** Our controls are limited to posing arms. The coder cannot change the head pose or specify facial expressions. Also, there are no controls for the upper body to adjust e.g. the shoulders for shrugging or produce hunched over or upright postures. Also, no locomotion or leg positioning are possible.

**Pose matching** For every new pose, our tool puts the current frame of the video in the background behind the skeleton (Figure 3). This screenshot serves as reference for the to be annotated pose. Automatic alignment of skeleton and screenshot is performed by marking the shoulders: the tool then puts the screenshot in a correct position to match the skeleton size. To check the current pose in 3D space, the *pose viewer window* (Figure 5) offers three adjustable views (different camera position + angle). Additionally, the user can adjust the camera in the main editor window.

**From poses to motion** The skeleton is animated by interpolating between poses in realtime. Using this animation the

[3] `https://collada.org`



Sideward: –0,02 m

Upward: 0,67 m

Toward: –0,15 m

Image size: 119 %

Figure 3: The *pose editor* window takes the current frame from the video and places it in the background for direct reference.

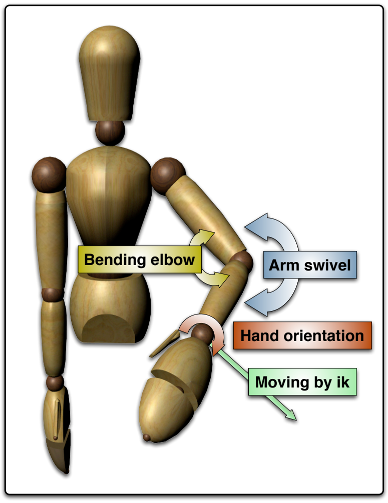


Figure 4: The coder can pose the skeleton by moving the end effector (green). To adjust the final pose the coder can correct the arm swivel (blue), bend the elbow (yellow) or change the hand orientation (red).

coder can validate whether the specified movement matches the original motion. The coder can always improve the animation by adding new key poses. In the *sequence view window* thumbnails of the poses are shown to allow intuitive viewing and editing (Figure 6(a) and Figure 6(b)).

**Data export** Apart from the regular ANVIL file, the poses are stored in a format that allows easy reuse in animation systems but can also be analyzed by behavior analysts. We support the two standard formats: *Collada* and *BML* (Behavior Markup Language)[4]. BML is a description language for human nonverbal and verbal behavior. *Collada* is a standard to exchange data between 3D applications, supported by many 3D modeling tools like Maya, 3D Studio MAX or Blender. The animation data can be used to animate own skeletons in these tools or for realtime animation with animation engines like *EMBR*.

[4] `http://wiki.mindmakers.org/projects:bml:main`

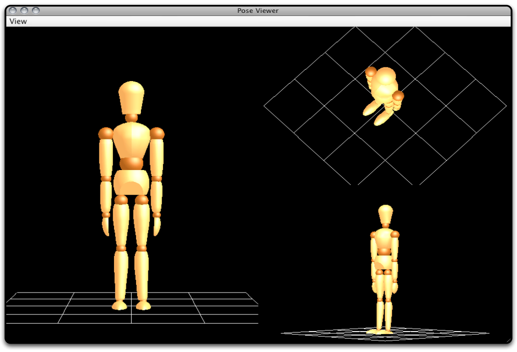


Figure 5: The *pose viewer* provides multiple views on the skeleton. Additionally the user can zoom and rotate the camera in each view seperatly.

## 3. Evaluation

In an evaluation study we examined the intuitiveness and efficiency of our new annotation method. We recruited eight subjects (21-30 years) without prior annotation or animation experience. The task was to annotate a given gesture sequence (123 frames). Subjects were instructed with a written manual and filled in a post-session questionnaire. Subject took 19 mins on average for the gesture sequence (123 frames = approx. 5 sec). At least 13 poses were annotated. We compared annotation times with the performance of an expert (one of the authors) which we took as the optimal performance. In addition, the expert performed conventional symbolic annotation (Fig. 7(a)). What is clear is that symbolic and skeleton annotation are similar in terms of time, even though the symbolic annotation is much coarser in resulting data. The learning curve of the non-expert subjects needed a ”normalization” because the different poses were of differing complexity.

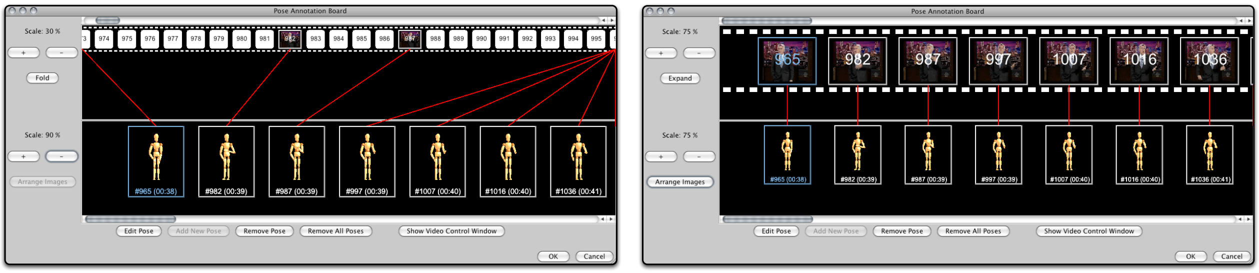
The complexity of a pose is measured by looking at the difference between two neighboring poses. The following formula defines complexity $C$:

$$C = T + R$$

where $T$ is the covered distance of both end-effectors between the given pose and a constant base pose, and $R$ is the sum of all joint modifications. A joint modification is the angle difference between two joint orientations (Nguyen, 2009). This means, that a pose has the highest complexity if both arms are moved in the widest range and all joints are rotated by the maximal degree.

The normalized curve (Fig. 7(b)) nicely shows that even within a single pose, a significant learning effect is observable which indicates that the interface is intuitive.

The subjective assessment of our application (by questionnaire) was very positive. It showed that the application was accepted and easy to use. Subjects often described the application as ”plausible and intuitive”. Our application seems to be, at least in regard to subjective opinions, an intuitive interface. For instance the subjects appreciated the film stripe looks of the *sequence view window* in the sense that the functionality was directly clear. Additionally the annotation with this 3D extension was rated as ”easily



(a) Expanded and zoomed view of the pose annotation board

(b) Folded view of the pose annotation board

Figure 6: Expanded view (left) and folded view (right) of the *pose annotation board*. The white areas symbolize unannotated frames. The coder can add new key poses to improve the animation by clicking on these areas. Areas with images represent annotated key frames. To keep an overview of all annotated key frames the coder can fold this stripe to only see key frames. Additionally, the view can be zoomed in or out to have an overview about all frames at a glance.

accessible". One reason was the result can bee seen directly. The possibility to see the animation of the annotated gesture immediately was "highly motivating."

We conclude that our method is regarded as intuitive in subjective ratings and appears to be highly learnable and efficient in coding. Note the impressive performance of the expert coder who was able to code a whole gesture in approx. 1 minute.

## 4. Related Work

To the best of our knowledge, this is the first tool using 3D skeletons to transcribe human movement. Previous approaches for transcribing gesture rely on symbolic labels to describe joint angles or hand positions. In own previous work (Kipp et al., 2007), we relied on the transcription of hand position (3 coordinates) and arm swivel to completely specify the arm configuration (without hand shape). We could show that our scheme was more efficient than the related Bern and FORM schemes, although it must be noted that those schemes offer a more complete annotation of the full-body configuration.

The Bern scheme (Frey et al., 1983) is an early, purely descriptive scheme which is reliable to code (90-95% agreement) but has high annotation costs. For a gesture of, say, 3 seconds duration, the Bern system encodes 7 time points with 9 dimensions each (counting only the gesture relevant ones), resulting in 63 attributes to code. FORM is a more recent descriptive gesture annotation scheme(Martell, 2002). It encodes positions by body part (left/right upper/lower arm, left/right hand) and has two tracks for each part, one for static locations and one for motions. For each position change of each body part the start/end configurations are annotated. Coding reliability appears to be satisfactory but, like with the Bern system, coding effort is very high: 20 hours coding per minute of video. Of course, both FORM and the Bern System also encode other body data (head, torso, legs, shoulders etc.) that we do not consider. However, since annotation effort for descriptive schemes is generally very high, we argue that annotation schemes must be targeted at this point to be manageable and have research impact in the desired area.

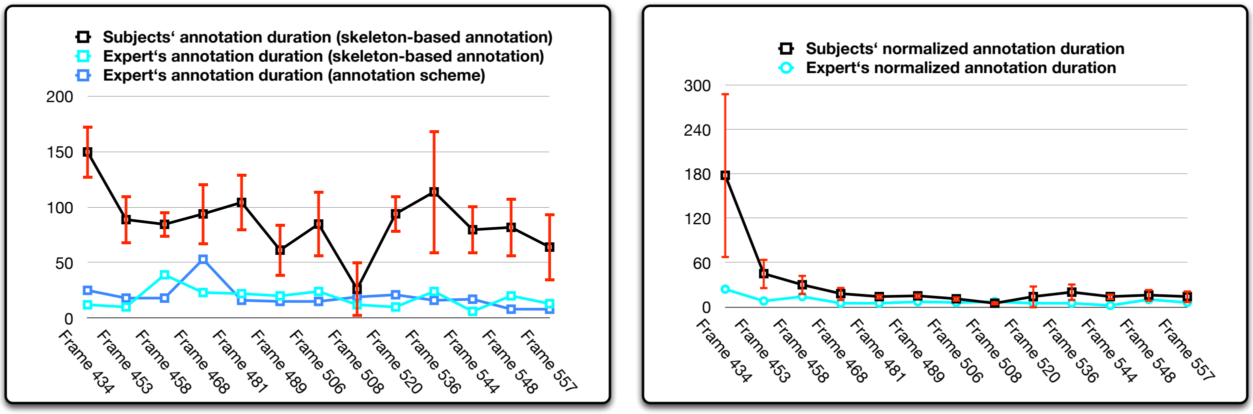
Other approaches import numerical data for statistical analysis or quantitative research. For instance, Segouat et al. import numerical data from video, which are generated by image processing, in ANVIL to analyze the possible correlation between linguistic phenomena and numerical data (Segouat et al., 2006). Crasborn et al. import data glove signals into ELAN to analyze sign languages gestures (Crasborn et al., 2006). ANVIL offers the possibillity to import and visualize motion capture data for analysis (Kipp, 2010b). It shows motion curves of e.g. the wrist joint's absolute position in space, their velocity and acceleration. These numerical data are useful for statistical analysis and quantitative research (Heloir et al., 2010). However, our extension supports the annotation of poses and gestures, so that the annotated gesture or pose can be reproduced and reused to build a repertoire of gesture from a specific speaker.

## 5. Conclusions

We presented an extension to the ANVIL annotation tool for transcribing human gestures using 3D skeleton controls. We showed how our intuitive 3D controls allow the quick creation, editing and realtime viewing of poses and animations. The latter are automatically created using interpolation. Apart from being useful in a computer animation context, the tool can be used for quantitative research on human gesture in fields like conversation analysis, gesture studies and anthropology. We also argued that the tool can be used in the field of intelligent virtual agents to build a repertoire of gesture templates from video recordings. Future work will investigate the use of more intuitive controls for posing the skeleton (e.g. using multitouch or other advanced input devices) and automating part of the posing using computer vision algorithms for detecting hands and shoulders. Additionally, we plan to provide more controls for the manipulation of shoulders (shrugging), leg poses or body postures.

## Acknowledgements

Part of this research has been carried out within the framework of the Cluster of Excellence *Multimodal Computing and Interaction* (MMCI), sponsored by the German Research Foundation (DFG).



(a) Average annotation time

(b) Average improvement in the course of annotation

Figure 7: The left diagram shows the annotation duration per frame (successive frames of a single gesture). This was measured for all subjects (black line shows the means, red bars indicate standard deviation) and for one expert where we compared our skeleton-based annotation with the standard ”annotation scheme” method. In the right diagram all durations are normalized against the *complexity C* of a pose. Only then do we see a clear learning effect after a few poses.

## 6. References

Onno Crasborn, Hans Sloetjes, Eric Auer, and Peter Wittenburg. 2006. Combining video and numeric data in the analysis of sign languages with the elan annotation software. In *Proceedings of the 2nd Workshop on the Representation and Processing of Sign languages: Lexicographic matters and didactic scenarios*.

S. Frey, H. P. Hirsbrunner, A. Florin, W. Daw, and R. Crawford. 1983. A unified approach to the investigation of nonverbal and verbal behavior in communication research. In W. Doise and S. Moscovici, editors, *Current Issues in European Social Psychology*, pages 143–199. Cambridge University Press, Cambridge.

Alexis Heloir and Michael Kipp. 2009. Embr — a realtime animation engine for interactive embodied agents. In *IVA '09: Proceedings of the 9th International Conference on Intelligent Virtual Agents*, pages 393–404, Berlin, Heidelberg. Springer-Verlag.

Alexis Heloir, Michael Neff, and Michael Kipp. 2010. Exploiting motion capture for virtual human animation: Data collection and annotation visualization. In *Proc. of the Workshop on ”Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality”*.

Michael Kipp, Michael Neff, and Irene Albrecht. 2007. An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. *Journal on Language Resources and Evaluation - Special Issue on Multimodal Corpora*, 41(3-4):325–339, December.

Michael Kipp. 2001. Anvil – a Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of Eurospeech*, pages 1367–1370.

Michael Kipp. 2010a. Anvil: The video annotation research tool. In Jacques Durand, Ulrike Gut, and Gjert Kristofferson, editors, *Handbook of Corpus Phonology*. Oxford University Press.

Michael Kipp. 2010b. Multimedia annotation, querying and analysis in anvil. In Mark Maybury, editor, *Multimedia Information Extraction*, chapter 21. MIT Press.

Craig Martell. 2002. FORM: An extensible, kinematically-based gesture annotation scheme. In *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP 02)*, pages 353–356, Denver.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style. *ACM Transactions on Graphics*, 27(1):1–24, March.

Quan Nguyen. 2009. Werkzeuge zur IK-basierten Gestenannotation mit Hilfe eines 3D-Skeletts. Master's thesis, University of Saarland.

T Schmidt. 2004. Transcribing and annotating spoken language with exmaralda. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora*.

Jérémie Segouat, Annelies Braffort, and Emilie Martin. 2006. Sign language corpus analysis: Synchronisation of linguistic annotation and numerical data.

Brugman H. Russel A. Klassmann A. Wittenburg, P. and H. Sloetjes. 2006. Elan: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*.