An Annotation Scheme for Conversational Gestures: How to economically capture timing and form

Michael Kipp
DFKI, Germany
kipp@dfki.de

Michael Neff
MPI Informatik, Germany
neff@mpi-inf.mpg.de

Irene Albrecht
MPI Informatik, Germany
albrecht@mpi-sb.mpg.de

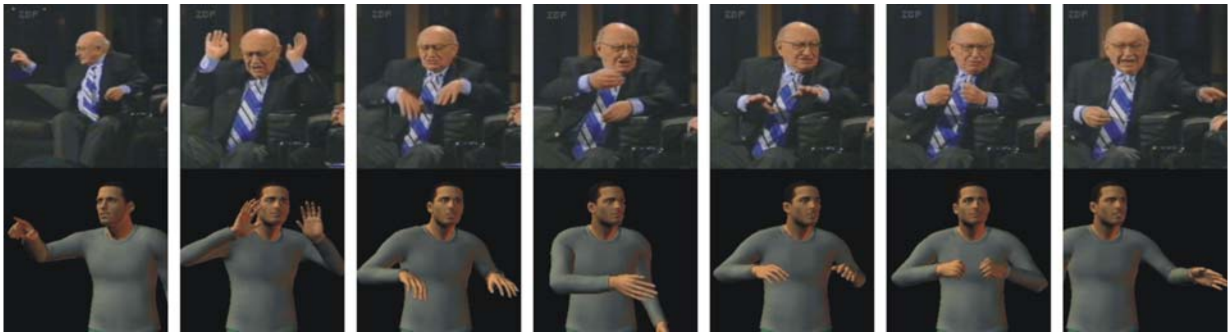


Figure 1. Selected frames of a source video (top) and a kinematic animation (bottom). The animation re-created the motion of the gesturing arm/hand of the original video from a manual annotation of the video which is based on our annotation scheme.

ABSTRACT

The empirical investigation of human gesture stands at the center of multiple research disciplines, and various gesture annotation schemes exist, with varying degrees of precision and annotation effort. We present a gesture annotation scheme for the specific purpose of automatically generating and animating character-specific hand/arm gestures, but with potential general value. We focus on how to capture temporal structure and locational information with relatively little annotation effort. The scheme is evaluated in terms of how accurately it captures the original gestures by re-creating those gestures on an animated character using the annotated data. This paper presents our scheme in detail and compares it to other approaches.

Author Keywords
Embodied Agents, Gesture Generation, Multimodal Interfaces

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Animated characters are useful in a wide range of applications like interfaces, games and movies. Generating nonverbal behavior for artificial bodies remains a challenging research task. One important technique for reproducing human-like gestures is to analyze original human behavior [7,9]. This can be done using motion capture or by manually annotating video data. While motion capture has unequalled precision, the video annotation approach has other advantages: it is an indirect observation method where people are less aware or unware of the observation, and arbitrary material (e.g. TV shows) can be analyzed, even of people otherwise unavailable. Moreover, the acquired data is usually encoded on an abstract level that can be understood and analyzed by conversational analysts, linguists, ethologists and computer animators alike, whereas motion captured data can only be interpreted with significant computational and human effort.

If the annotated data is to be used with an animation system that can create arbitrary motions for a humanoid character, the need for precise positional data becomes highly important, especially if you want to capture the specific style of a speaker. Speakers do not only differ in what and when they gesture, but also *where* they gesture. For instance, the “raised index finger” can be displayed quite shyly near the chest or dominantly above the head. We believe that such locational variation is integral to personal style. When encoding positional information, the question arises as to how faithfully that encoding reflects the original movement. Successfully re-creating the original motion from the encoded data would prove that something essential must have been captured by the annotation (see Figure 1).

Annotation schemes for human movement can be classified according to the amount of detail they capture, where high detail seems to be proportional to high annotation cost and a low level of abstraction. On one side of the spectrum lies the Bern system [2,3], where a large number of degrees of freedom are manually annotated, thus resembling modern motion capture techniques. While it results in fine grained, purely descriptive and reliably coded data which can be reproduced easily with a synthetic character, annotation effort is immense. In addition, the resulting data is hard to interpret. It does not abstract away from even minor variations and the amount of data is so massive that it is hard to put it in relation to the accumulated knowledge about gesture structure and form found in the literature. On the other end of the spectrum, lies Conversational Analysis, where the written speech transcription is used as a basis and gestures are annotated by inserting brackets in the text for beginning and end of the gesture [4]. Gesture form is captured by either a free-form written account or by gestural categories which describe one prototypical form of the gesture. Such information would be too informal or too imprecise for automatic character animation. Thus, a key decision in annotation is: how much do you abstract? Or, how large are your equivalence classes?

We propose a scheme that makes a conscious compromise between purely descriptive, high-resolution approaches and abstract interpretative approaches. We restrict ourselves to hand/arm movement to identify the most essential features of a gesture before moving to other body parts. Our scheme encodes positional data but relies on an intelligent "time slicing", based on the concept of movement phases, to determine the most relevant time points for position encoding. It is based on the observation that transition points between phases correspond to key frames in traditional animation. Moreover, we use the concept of a gesture lexicon, well known in Conversational Analysis, where each lexeme contains some generalized information about form. Lexemes can be taken as prototypes of recurring gesture patterns. When encoding lexeme type for an annotated gesture in the video material all this general data is implicitly encoded as well.

TARGET SCENARIO

Our annotation scheme aims at the specific application of gesture generation for an animated character. However, we think that the annotation scheme will be of general interest in the interdisciplinary fields of multimodal and gesture research. The needs that arise from *animating* gestures on the basis of manual annotation provide good guidance on the essential descriptive parameters of human gestures.

The generation approach we aim at "imitates" a human speaker's gesture behavior using statistical models and a database of sample gestures, both extracted from video annotations [7]. For this application, the annotation scheme must capture the temporal and spatial structure of a gesture, and its relation to speech. Since original gesture samples are re-used in generation, the annotation should make it possible to re-create original gestures in synthetic animation. On the other hand, the annotation should be as economical as possible in terms of annotation effort.

Our video corpus consists of selected video clips from two TV talk shows, featuring two different speakers.

ANNOTATION SCHEME

While gestures appear to be quite arbitrary in form at first glance various researchers found them to have fairly stable form, even if they are not clear emblems [5]. Conversational gestures have no clear meaning and may even be a byproduct of speech. However, there seem to be shared lexica or inventories of conversational gestures [12]. For instance, the metaphoric gesture "progressive" [11], where a speaker makes a circular movement with the hands, seems to occur when talking about progress, movement or the future [1]. Another universal gesture is the "open hand" where the speaker holds the open hand in front of the body, showing the palm [4,11]. While such forms appear to be universal, there is still much inter-speaker and intra-speaker variation in terms of the *exact* position of the hands and their ensuing trajectory. To investigate and capture these variations was one driving force of our work.

We use the Anvil video annotation tool [6] for our purposes, which allows annotation on multiple tracks. Coding consists of adding annotation elements which can be complex attribute-value objects.

Capturing Temporal Structure

We capture the temporal structure of a gesture by first identifying the basic movement phases [4,8,11]:

preparation > hold > stroke > hold > retraction

where the stroke is the most energetic part of the gesture while the preparation moves to the stroke's starting position. Holds are optional still phases which can occur before and/or after the stroke. Kita et al. [8] identified *independent holds* which can occur instead of a stroke. The retraction returns to a rest pose (e.g. arms hanging down, resting in lap, or arms folded). Kita et al. refined the notion of stroke by defining a multiple stroke that includes small beat-like movements that follow the first stroke, but seem to belong to the same gesture. In our scheme, a stroke contains a "number" attribute to capture the number of within-stroke movements.

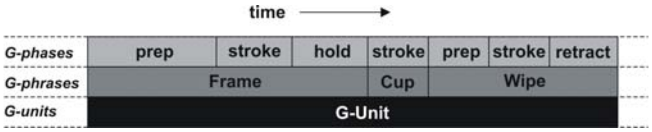


Figure 2. Gesture Annotation on Three Anvil Tracks.

To annotate phases in Anvil, the coder specifies beginning and end times of a phase as well as phase type (prep, stroke, etc.) and stroke number. On a second track, the coder combines phases into gestures, also called gesture phrases (Figure 2). In this way, we store the gesture's internal temporal structure, most importantly begin/end times of the stroke or independent hold. On a third track, we combine gestures into gesture units. A gesture unit is a sequence of contiguous gestures in which the hands do not return to a rest pose until the end of the last gesture [4,11]. This allows us to examine a speaker's g-unit structure. For instance, the average number of gestures, patterns of recurring lexeme sequences etc.

Capturing Spatial Form

In order to capture the spatial form we aimed at the best compromise between exactness and economy. For the sake of economy we make two important assumptions: (1) the most "interesting" configurations occur exactly at the beginning and at the end of a stroke, and (2) bihanded gestures are symmetrical. Although many gestures are actually asymetrical, most of them can be approximated quite well with symmetrical versions.

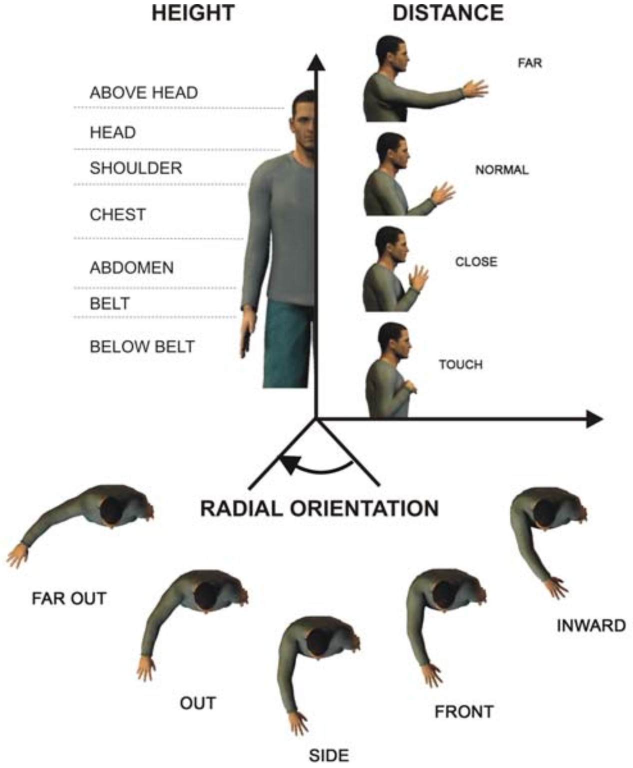


Figure 3. Our three dimensions for hand position.

The first two parameters encoded are handedness and whether the trajectory of the hand(s) in the stroke phase is straight or curved. Next, we have to capture the start and end positions of the hands/arms for the stroke. For a single position we encode three dimensions for hand location (Figure 3) and encode elbow inclination in a fourth dimension (Figure 4). The dimensions were chosen such that (1) we have sufficient granularity for later animation and (2) it is quick and reliable to annotate video, which explains the selection of landmarks like "shoulder", "belt" and intuitive terms like "normal". For bihanded gestures, we also encode hand-to-hand distance for added precision by marking the hands on the video screen; we extended Anvil to handle this new kind of "spatial annotation" (Figure 5). The hand-to-hand distance is normalized by dividing it by the shoulder width which must be encoded each time the size of the displayed speaker changes due to camera movement.

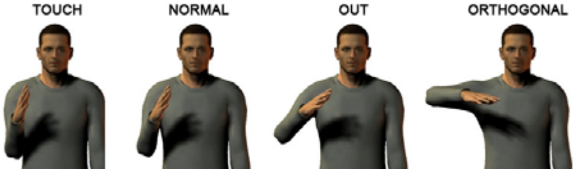
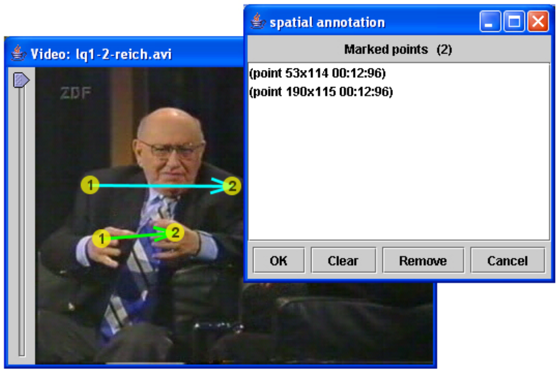


Figure 4. A fourth dimension encodes elbow inclination.

In summary, for each stroke based gesture we encode 2 positions where each position is expressed by 5 attributes. Adding handedness and trajectory gives us 12 attributes to code for the spatial form of a gesture. Independent holds only require 1 position, for the beginning of the hold.



(point 53x114 00:12:96)
(point 190x115 00:12:96)

Figure 5. Annotating 2D points in Anvil: Shoulder width (top arrow) and hand-to-hand distance (bottom arrow).

Capturing Membership to Lexical Category

A number of parameters are determined by the gesture's lexeme, including: handshape, palm orientation and exact trajectory. For each lexeme, these parameters can be either fixed (definitional parameter), restricted to a range of values, or arbitrary. To annotate lexemes on the phrase track, we rely on a simplified version of the gesture lexicon collected in [7] where 79% agreement in lexeme coding experiments is reported. Typical lexemes include: RaisedIndexfinger, Cup (open hand), FingerRing (thumb touches index finger) and Progressive (circular movement). We found 31 and 35 different lexemes for our two speakers with an overlap of 27 lexemes between the two.

Capturing the Relationship To Speech

Once shape and lexeme are determined, the gesture must be connected to speech. When annotating real data, we found that the claim that gesture stroke and lexical affiliate always co-occur [11] is often wrong. Therefore, we encode co-occurrence and lexical affiliate in different attributes. Co-occurrence is not trivial. The gesture stroke has a temporal extension and may overlap with many co-occurring words. Choosing every overlapping word does not reflect our intuition of gesture-word co-occurrence. We use the following heuristics to automatically annotate co-occurrence: From the words overlapping with the stroke, choose (1) the word carrying the emphasis, if present, or else (2) the last word. Lexical affiliation is a more difficult task. We rely on the gesture literature and sometimes intuition when it comes to connecting gestures to the speech's semantics (cf. [7]). The lexeme usually gives some direction: for pointing gestures look for personal pronouns like "you", "his" etc., for the metaphoric "Cup" gesture look for the closest noun, for the metaphoric "Progressive" gesture look for the closest verb or noun that expresses movement or temporal relation.

EVALUATION BY RE-CREATION

Any transcription scheme must be measured by two factors. First, how well the annotation reflects the original motion (usually dependent on application or experiment). Second, how reliably the annotation can be performed by human coders. While we have not yet tested reliability, we propose a method for the first criterion: re-creating the gestures with an animated agent [2]. Using only the pure annotation information already produced satisfying results. Adding information that had been manually collected for specific gesture lexemes (hand shape/orientation, trajectory) produced animations that very precisely matched the original motions. See Figure 1 for an impression of our re-creation experiments.

RELATED WORK

In this section we focus on two highly related schemes (for a general overview see [13]). The Bern scheme [2,3] 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. In comparison, our scheme needs a maximum of 12 attributes for a gesture's positional information. FORM is a more recent descriptive gesture annotation scheme [10]. 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. By contrast, we measured an average effort of only 1 hour per minute of video for our scheme. We explain this stark difference by our very focused approach to gesture annotation. While FORM encodes every movement of independent body parts, we hypothesize that the stroke (or independent hold) alone carries the definitional part of the gesture. 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.

CONCLUSION

We presented an effective gesture annotation scheme for gesture generation that appears to be a good compromise between detail and economy. Re-creating animations showed that the scheme captures the original motions quite well. We consciously restricted the project to arm/hand movement, ignoring the rest of the body for the sake of simplicity. However, other body parts should be included in the future. Another future issue is to test coding reliability.

We think that the main reason why our annotation so successfully captures gestures in an economic way is that it consciously focuses the annotation effort by exploiting the concept of gesture phases. The coder first identifies those time points most worth investing annotation work in and only then encodes the time-consuming positional data. Another "trick" is to move recurring patterns to a lexicon of gestures. By identifying the lexeme of a gesture, the coder specifies a number of features that need not be transcribed. While our annotation scheme has obvious drawbacks in what it does not capture (handshape, asymmetry, etc.) it is straightforward to extend if necessary. However, part of our intent in creating this scheme was to find the most economical solution for descriptive gesture annotation.

ACKNOWLEDGEMENTS

The research presented here was partially funded by the German Ministry of Research and Technology (BMBF) under grants 01 IMB 01A (VirtualHuman). The responsibility for the contents of this paper lies with the authors.

REFERENCES

1. Calbris, G. *Semiotics of French Gesture*. Indiana University Press, Bloomington, Indiana, 1990.
2. Frey, S. *Die Macht des Bildes*. Verlag Hans Huber, Bern, 1999.
3. Frey, S., Hirsbrunner, H. P., Florin, A., Daw, W. and Crawford, R. A Unified Approach to the investigation of Nonverbal and Verbal Behavior in Communication Research. In: *Current Issues in European Social Psychology*, Doise, W. and Moscovici, S. (eds.), Cambridge University Press (1983), 143-199.
4. Kendon, A. *Gesture: Visible Action as Utterance.* Cambridge University Press, 2004.
5. Kendon, A. An Agenda for Gesture Studies. In: *The Semiotic Review of Books* **7** (3), 1996, 8-12.
6. Kipp, M. Anvil - A Generic Annotation Tool for Multi-modal Dialogue. In *Proc. Eurospeech 2001*, 1367-1370.
7. Kipp, M. *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation.* Dissertation.com, Boca Raton, FL, USA, 2004.
8. Kita, S., van Gijn, I. and van der Hulst, H. Movement Phases in Signs and Co-speech Gestures, and Their Transcription by Human Coders. In *Gesture and Sign Language in Human-Computer Interaction*, Wachs-muth, I. and Fröhlich, M. (eds.). Springer (1998), 23-35.
9. Kopp, S., Tepper, P. and Cassell, J. Towards integrated microplanning of language and iconic gesture for multimodal output. In: *Proc. Intl. Conf. Multimodal Interfaces 2004*, 97-104.
10. Martell, C. FORM: An Extensible, Kinematically-Based Gesture Annotation Scheme. In *Proc. ICSLP 2002*, 353-356.
11. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, 1992.
12. Webb, R. *Linguistic Properties of Metaphoric Gestures*. PhD thesis, University of Rochester, New York, 1997.
13. Wegener Knudsen, M., Martin, J.-C., Dybkjær, L., Machuca Ayuso, M., Bernsen, N.O., Carletta, J., Heid, U., Kita, S., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van Elswijk, G., Wittenburg, P. *Survey of Multimodal Annotation Schemes and Best Practice*. ISLE Deliverable D9.1, 2002.