Exploiting Motion Capture for Virtual Human Animation
Data Collection and Annotation Visualization

Alexis Heloir[1], **Michael Neff** [2], **Michael Kipp**[1]

[1]DFKI, Germany [2]UC Davis, USA
[1]firstname.surname@dfki.de [2]neff@cs.ucdavis.edu

Abstract

Motion capture (mocap) provides highly precise data of human movement which can be used for empirical analysis and virtual human animation. In this paper, we describe a corpus that has been collected for the purpose of modelling movement in a dyadic conversational context. We describe the technical setup, scenarios and challenges involved in capturing the corpus, and present ways of annotating and visualizing the data. For visualization we suggest the techniques of motion trails and animated re-creation. We have incorporated these motion capture visualization techniques as extensions to the ANVIL tool and into a procedural animation system, and show a first attempt at automated analysis of the data (handedness detection).

1. Motivation

Video has been the technology of choice for empirical movement analysts since it faithfully records the movements, facial expressions and spatial surroundings of the recorded subject. However, video has obvious limitations: the view angle cannot be changed after recording and any automatic analysis must use computer vision techniques to extract meaningful information like hand/face locations.

Motion capture is becoming an increasingly widely available resource for recording human movement. It allows researchers to supplement audio and video recordings with 3D reconstructions of a performer's movements. Normally, motion capture techniques reconstruct body movement as a stick figure skeleton, yielding angle data at each joint in the skeleton. The amount of data that can be captured is a function of the number and resolution of the cameras available, the number of subjects, the range of movement allowed and the amount of time available for cleaning and reconstructing the data. Nonetheless, motion capture provides a more precise 3D view of a subject's movement and also supports automated analysis of the data.

Although motion capture offers numerous advantages for motion analysis, new tools and visualization techniques are needed to fully exploit the potentials of this technology. In this paper, we present some of the trade-offs involved in building a corpus of conversational interactions that includes motion capture. Our intended application is virtual character models that can both talk and gesture. We discuss issues related to both capturing data and analysis. We also illustrate how an existing motion annotation tool ANVIL (Kipp, 2001; Kipp, 2010b; Kipp, 2010a) can be extended in order to take advantage of such data.

Our final corpus contains audio, video and motion capture data. Each modality provides different, important information for the analysis and synthesis process. Audio data provides the text that was spoken and the word timings. Motion capture data provides a 3D reconstruction of the motion, but in many standard applications such as ours, this reconstruction is at the fidelity of a stick figure. It does not capture the surface deformations of the performer, including facial expressions, muscle bulges and breathing. Video helps provide these missing pieces. Shooting from two angles, we can capture facial expressions of both interlocutors and also subtleties of body movement that may be missed in the motion capture.

2. Motion Capturing Dyadic Conversations

Building a corpus begins by determining the goals for its intended use, and from that, planning a set of scenarios to record, and choosing appropriate subjects. Our goals were to perform early, exploratory studies on gestures analysis and generation for two person (dyadic) conversations. This required obtaining a wide set of gesture variations. In this section, we describe one particular session in building our corpus.

2.1. Scenarios

We decided to use improvised scenarios as they placed less demand on our subjects by not requiring them to learn lines and also avoided introducing the bias of a pre-selected script. We chose subjects with extensive movement experience, both subjects had dance training and performance experience. Both were trained in Laban Movement Analysis. In general, we feel experience with verbal improvisation and physical acting is important for this kind of session, and offers the following benefits:

- subjects can better cope with the disturbing garment/setup required by motion capture,
- subjects can improvise coherent stories and interaction with minimal guidance,
- subjects can take directions well and adjust their performance to yield the desired data,
- the addtional training these subjects had in Laban Movement Analysis (LMA) (Laban, 1988) allowed them to be given directions in terms of LMA parameters, which allows precise changes in movement to be requested.

We recorded 23 separate sequences, each having a length of 1-2 minutes. In 19 of the 23 sequences, both actors were interacting. Performers were given minimal improvisation instructions, each focused on particular aspects of interaction:

- social status and levels of dominance, as suggested by Johnstonne (Johnstone, 1981),
- *valence* of the interaction,
- amount of *arousal* in the interaction,
- discussions where subjects *agree or disagree*.

The recording started with a warm-up sequence where subjects were told to talk about what they did the day before. Subsequent sequences are summarized and briefly described in the following table:

Dominance	
Corrupt judge and briber	judge is low status
Corrupt judge and briber	judge is high status
New neighbors meeting	both high status
New neighbors meeting	both low status
Boss fires employee	boss is high status
Boss fires employee	boss is low status
Valence	
Old friends meet	they are happy to meet
Uncomfortable meeting	they dislike each other
Arousal	
Sketch of the dead parrot	high arousal
Sketch of the dead parrot	low arousal
Agree/disagree	
Coffee is better with a cigarette	disagree
Brad Pitt should be president	agree
Mac computers are better	disagree
Jay Leno is an alien	agree
The best way to eat an egg (small end/big end first)	disagree

2.2. Technical Setup

We were interested in capturing the following modalities for two subjects simultaneously: speech, posture and gesture, including hand shape. Our corpus was recorded in a motion capture lab equipped with 12 optical Vicon MX 40+ cameras and two digital HD video cameras. This system tracks the 3D locations of refelctive markers attached to the subjects. Speech and facial expressions were captured using digital video cameras aimed at each subject. Motion capture was used to record both body motion and hand shape, the latter being particularly challenging. The difficulty of recording finger motion using optical motion capture, especially with a limited number of cameras, comes from the high probability of visual occlusions between crossing and/or overlapping fingers. Full body and hand capture can be attempted using three different strategies:

One strategy consists of recording the hand motion and the body motion separately. The performer must wear different markers for each capture session and the two sets of data must be spliced together afterwards using temporal warping algorithms. This method has been successfully demonstrated by Majkowska et al. for choreographed Mudras dance (Majkowska et al., 2006). Unfortunately our scenarios heavily rely on improvised performances in which the hand poses are unknown. Indeed, one of our goals is to study changes in hand shape. Therefore, this technique didn't fit our requirements.

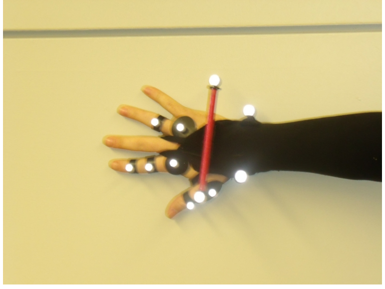

Figure 1: Marker constellation used for hand shape and finger orientation.

A second strategy uses a combination of an optical motion capture system for the body motion and a glove equipped with bend sensors for the hands. This technique has the advantage of being robust to finger occlusions and has been successfully employed for recording Sign Language sequences (Heloir et al., 2005). However, data gloves have several drawbacks: they record motion at lower frequency than optical system (approx. 60Hz vs 120Hz), the sensors have non-linear behavior when approaching flexion limits, the gloves need to be recalibrated at regular intervals, they are expensive and many systems require wires.

A third strategy consists of using a limited set of optical markers on the hand to capture a portion of its movement, and then inferring the remainder of the hand shape. This technique has been used extensively in the motion picture industry and has proved to give acceptable, although not optimal, results. Recent research work took advantage of the joint inter-dependencies of the human hand to perform hand motion capture with a limited set of markers for grasping tasks (Chang et al., 2007). The third method was chosen because it made use of existing equipment and allowed for the simultaneous capture of hand and body movement.

After some experiments, we found that seven markers on the hand were enough to provide a faithful reconstruction of the hand's overall shape in most instances. The marker constellation for one hand is depicted in Fig. 1. We used two markers for the thumb, two markers for the ring finger and three markers for the index.

2.3. Lessons Learned

The recording of the 23 sequences took six to seven hours. Two hours were necessary to brief and prepare the two subjects. Once recorded, postprocessing of the motion capture data took one week for a single person working full time. Not surprisingly, the reconstruction of the hand motion required the most manual correction. Only for some sparsely occurring intervals (approx. 3% of the time), hand motion reconstruction could not be achieved due to occlusion.

3. Annotation, Analysis and Visualization

In our work, we are concerned with the phase structure of gesture (Kita et al., 1998). A given gesture can be broken down into a set of phases: preparation, hold, stroke and retraction. The whole gesture is considered the next level of

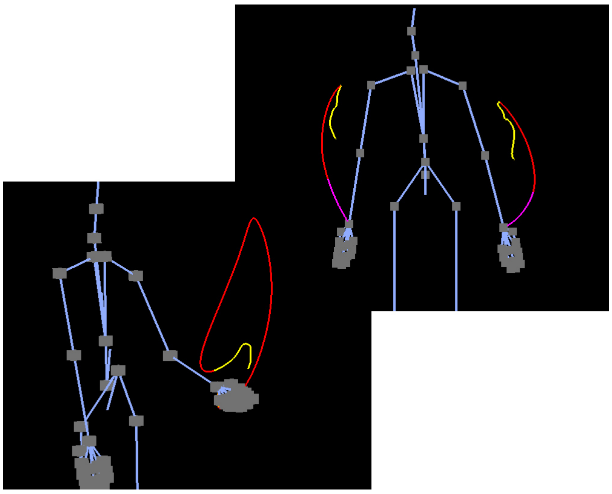

Figure 2: Two examples of gesture trails. Yellow indicates the preparation phase, red the stroke and magenta the retraction.

analysis and called a *phrase* in the literature. It is useful to break these phases out both for analysis and generation. From the perspective of analysis, the *stroke* phase is considered the meaning carrying portion of the gesture, so it is helpful to separate it from the total gestural movement. Another important phase is the *independent hold* if a gesture has no movement at all (e.g. the proverbial raised index finger). Both strokes and independent holds are called the *expressive phase* of a gesture. Other factors like the occurrence and length of holds can help define a particular individual's gesture style. From the perspective of generation, the phase structure provides a convenient framework for specifying animation. A system can solve for the poses at the phase boundaries and interpolate in between to create continuous gesture animation.

We manually annotate gesture phases using the tool ANVIL which has recently been extended to visualize motion capture data using a 3D skeleton (Kipp, 2010b). ANVIL allows users to view synchroized video, 3D skeleton data and time-aligned annotations. Additionally, ANVIL can visualize the position, velocity and acceleration of the hands as curves (either x, y, z separately or as total value) on separate tracks (Figure 3).

3.1. Automated Handedness Analysis

Since motion capture data offers more information than plain video, providing nearly continuous 3D data, it offers increased potential for automatically deriving meaningful descriptions of the movement. In manual annotation, it can be a challenge to arrive at high inter-coder agreement for phenomena like gesture phase annotation, since this can be quite subjective, especially for spontaneous gestures. If more of these tasks can be successfully automated (even if partially so, combined with human corrections), it will increase inter-coder agreement and reduce coding effort. Detecting the hand used for a gesture (LH, RH, 2H) is one annotation task that lends itself to automation.

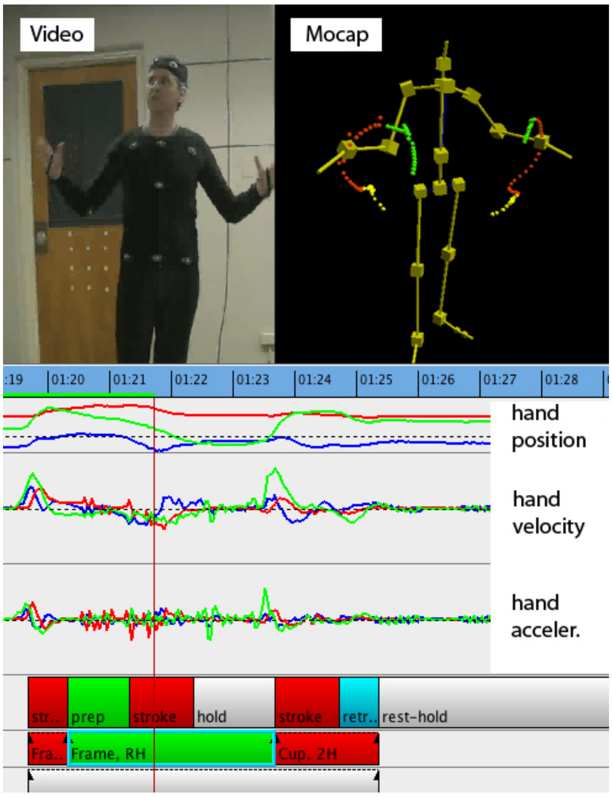

Figure 3: In the ANVIL tool, movement is usually encoded in terms of timeline-based annotations (bottom: colored boxes) and video. Mocap data allows for the display of hand position, velocity and acceleration curves. Motion trails visualize the hands' path in 3D, viewable from all angles.

To detect handedness on the phrase level (i.e. for a whole gesture) we first find the corresponding *expressive phase* on the phase track. The expressive phase is either a stroke or an independent hold (Kita et al., 1998). This phase is marked red in Fig. 2, note how the difference in length gives a clear cue of handedness. Therefore, we take the length of the path travelled by left hand L_{RH} and right hand L_{LH} respectively during this expressive phase (in meters), and normalize it by the duration d of the phase (in seconds). If the normalized difference $\frac{|L_{RH}-L_{LH}|}{d}$ is below the threshold of $0.12\frac{m}{s}$, we label it a bihanded gesture (2H), otherwise we label it right-handed if $L_{RH} > L_{LH}$, or left-handed (LH) if $L_{RH} < L_{LH}$. On an annotated corpus of 269 phrases, we achieved 83% correct annotations with this algorithm.

3.2. Annotation Visualization with Motion Trails

Previous approaches (Neff et al., 2008; Kipp et al., 2007) for annotating gesture have included positional data by estimating the wrist positions at the start and end of a stroke. This provides a sparse description of the gesture sequence. One of the chief advantages of motion capture is that by capturing over 100 samples per second, it approximates a more continuous representation of the motion. This allows us to visualize the overall form of a gesture.

We suggest a new visualization technique that draws the 3D movement of the speaker's hand as a "trail" through space, shown either as a continuous line or by discrete sphere (Figures 2 and 3). This allows one to closely examine the actual path of a gesture from all angles, revealing the smoothness or edginess of the curve and even giving an impression of the velocity profile which is reflected in the spacing of the spheres.

The trail feature has been incorporated into both a standalone animation package and the ANVIL annotation tool. The gesture trails are color coded to indicate the phases of the gesture, as shown in Figure 2. We can play an animation of the trail data with its actual timing, scrub through the trail and also view it from any direction in 3D. This allows for more carefull study of the gesture form and the transitions between the phases, to examine features like the continuity across phases. One insight we gained with respect to phase boundaries is that changes in hand shape may play a significant role in defining these boundaries because judging from the trails alone (no hand shape information visible!) boundaries would often have been placed a little earlier or later.

3.3. Validation by Recreation

In both ANVIL and our standalone animation system, we can simultaneously playback the motion capture data with the gesture trail over top of it. This provides an easy and effective method for validating an annotation. If the animated character performs a gesture, but there is no accompanying gesture trail, this indicates an error in the annotation. This makes it very easy to detect errors such as marking the incorrect hand, missing a gesture, or annotation errors in the timing of the gesture.

We can also use the motion capture data as input to a procedural motion generation system. For instance, the systems presented in (Neff et al., 2008; Heloir and Kipp, 2009) use the positions at the start and end of the stroke in order to generate animation. This position data can be automatically calculated from the motion capture data and then used as input to the procedural systems. We can overlay both the motion capture and generated animations and produce gesture trails for each. This allows for direct comparison between the form of the gesture created by the procedural system and the form of the original gesture. It provides a way to evaluate and improve the procedural generation system so that it can better match the captured data.

4. Summary and Outlook

We have described the recording of a motion capture corpus involving two speakers interacting in various improvisational scenarios. We showed that a standard optical motion capture setup was sufficient to provide a faithful reconstruction of the body and hand motion of both subjects. We found that 7 hand markers, strategically placed, were suffcient to reconstruct hand shape.

We also presented a basic annotation scheme in terms of gesture phases and two visualizaton helpers designed in order to reduce annotation errors and to increase inter-coder agreement. The first one, motion trails, shows the 3D path of the hands colored according to the movement phase annotation. The second one, re-creation, animates a stick figure according to extracted information which allow direct visual feedback concerning the quality of the animation algorithm.

In the future we plan to pursue two major lines of inquiry: first, determining methods to automatically derive movement phases from motion capture data and second, analyzing the particular interactions between two speakers in terms of timing, rhythm and imitation.

Acknowledgements

We would like to thank the performers for contributing to our corpus. Many thanks Quan Nguyen (DFKI) for implementing motion trails in ANVIL. Financial support for this work was provided in part by NSF grant IIS 0845529, by DAAD grant D/08/09945 (PPP USA) and by the DFG-funded Excellence Cluster *Multimodal Computing and Interaction* (MMCI).

5. References

Lillian Y. Chang, Nancy Pollard, Tom Mitchell, and Eric P. Xing. 2007. Feature selection for grasp recognition from optical markers. In *Proc. of the 2007 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS 2007)*, pages 2944 – 2950.

Alexis Heloir and Michael Kipp. 2009. EMBR - a realtime animation engine for interactive embodied agents. In *Proc. of the 9th Int. Conf. on Intelligent Virtual Agents (IVA-09)*.

Alexis Heloir, Sylvie Gibet, Franck Multon, and Nicolas Courty. 2005. Captured motion data processing for real time synthesis of sign language. In *Proc of GW-05*.

Keith Johnstone. 1981. *IMPRO: Improvisation and the Theatre*. Methuen Drama.

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*, 41(3-4):325–339, December.

Michael Kipp. 2001. Anvil – a generic annotation tool for multimodal dialogue. In *Proc. 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.

Sotaro Kita, Ingeborg van Gijn, and Harry van der Hulst. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In Ipke Wachsmuth and Martin Fröhlich, editors, *Proc. of GW-97*, pages 23–35, Berlin. Springer.

Rudolf Laban. 1988. *The Mastery of Movement*. Northcote House, London, fourth edition. Revised by Lisa Ullman.

Anna Majkowska, Victor B. Zordan, and Petros Faloutsos. 2006. Automatic splicing for hand and body animations. In *SCA '06*.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. on Graphics*, 27(1):5:1–5:24, March.