72.b Documentation with motion capture

1. Introduction

Modern tracking technology for full body tracking, also referred to as *motion capturing*, can be used as an alternative or a complementary method to video recordings (see also **Chapter 72.a on documenting with data gloves** for a discussion). It offers a high precision of position and orientation information regarding specific points of reference, which can be chosen by the experimenter. Motion tracking can be used to collect movement trajectories, posture data, speed profiles and other performance indices. Disadvantages are: the employment of obtrusive technology that has to be applied to the body of the target, such as reflective markers used by optical tracking systems; data which describes rather the postures and movements of a stick-like figure than that of a body with certain masses – therefor motion capturing should almost always go together with video recordings; and last but not least the costs of the installation.

The following text provides a short introduction into the state of the art of tracking technology that is available for the use in laboratories. While there are some companies that offer high-quality motion capturing services for the film and gaming industries, making use of their services is probably beyond scope of any modest research project. The following presentation is therefore focused on the technology that can be found in the research laboratories at the present time.

2. Tracking technologies

There are two general types of tracking technologies available: *marker-based* systems and *marker-less* systems. Both types have their advantages, especially in the context of authentic empirical research of natural human communication.

2.1. Marker-based tracking systems

Marker-based tracking systems are the most common type. When speaking of motion capturing, the picture of a guy in a black suite (a *motion-capturing suite*) which is systematically sprinkled with bright markers might come into our minds. Such systems are commonly used in film and game development to control computer graphical animations; a very famous example is the animation of Gollum in Lord of the Rings by Andy Serkis (who actually wore a blue dress with small black patches and bright markers).

As the name suggests, marker-based tracking systems rely on the detection of specific markers. The systems differ in the types of markers they use: some are based on passive reflective markers or colored patches others use active infrared-LEDs. Also combinations of different types of markers are used. The markers can be single objects, most frequently spheres that support a tracking of their position. As the position is specified in three coordinates (X/Y/Z), these markers provide three *degrees of freedom* (3 *DoF*). Other markers, such as the ones presented in Figure 1, have a more complex 3D structure. These markers also support a tracking of their orientation and thus have six degrees of freedom (6 DoF, 3 DoF for the position and 3 DoF for the orientation). Markers with 6 DoF can also be uniquely identified, which is not easily possible with sphere-like 3-DoF markers.

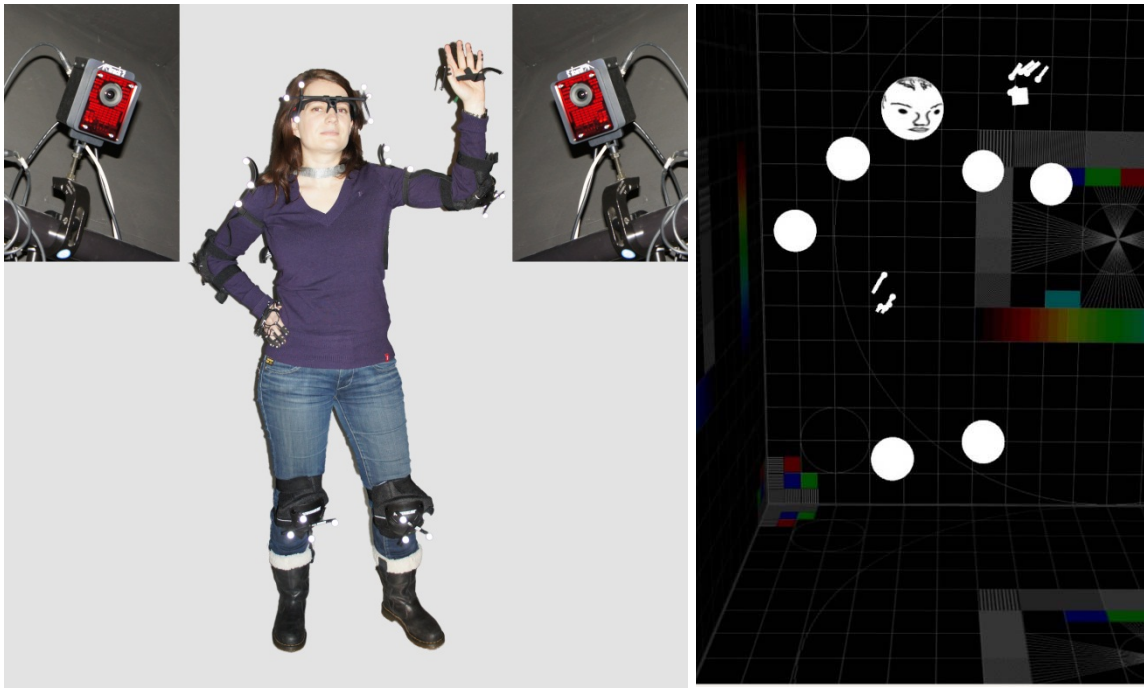

Figure 1: *Left:* The optical tracking system from Advanced Realtime Tracking GmbH uses infrared cameras (upper left and upper right corner) to measure the position and orientation of specifically designed targets. The targets use reflective spherical markers in unique 3D configurations, which enables the system to uniquely identify each marker. The person at the center wears markers at distinct positions relevant for a study on body movements related to verb productions. *Right:* The picture on the right shows a screenshot of the software visualizing the tracking data recorded using the setup on the left.

The systems with motion-capturing suites operate in the so-called *outside-in* mode: the markers are attached to the object of interest and the motion of the markers is observed by appropriate devices, such as high-speed infrared cameras. The markers have to be carefully arranged on the target, so that all relevant movements are captured and the structure of the body can be reconstructed from the data. Figure 1 shows a participant of one of our studies with a small set of markers attached to knees, shoulders, elbows and the head. The hands are tracked with the special finger-tracking system offered by the Advanced Realtime Tracking GmbH (ART 1999). Often, the body is represented by a skeletal stick figure. The movements of the markers are then translated into the movements of the skeletal representation of the body (see Figure 1, right). As the markers are attached to the soft skin of the body and not to the bones themselves, a mapping from the markers to the skeletal figure has to be defined, e.g. in an initial calibration procedure. This type of tracking is the standard tracking system for human body movements today. Prominent examples of such tracking systems are the systems from Advanced Realtime Tracking GmbH (ART 1999), Vicon Motion Systems (VMS 1984) and Motion Analysis (MA 1982).

The advantages of the outside-in tracking are that the markers which are attached to the tracked target are very small, lightweight and cheap. If the full body is to be tracked, many

markers need to be attached to get a good approximation of the body's posture and thus size, weight and costs of markers are important factors.

A major disadvantage is that the devices used to detect the markers and their movements need to be set up with a good view on the target object. Thus outside-in tracking faces a similar problem as standard video recordings do. The resolution and, e.g., the focal distances of the cameras are also restrictive factors which put hard constrains on the operational area in which movement can be tracked, the interaction space of the tracking system. A typical setup for the tracking of dyadic communication has an interaction space of 3 m x 3 m. Such a setup would already require eight or more cameras, which have to be placed carefully in the surrounding of the interaction space and which have to be thoroughly calibrated to construct a common coordinate system as frame of reference. An advantage of the motion tracking approach in contrast to normal video recordings is that whereas the video from each camera has to be annotated separately, the motion-capturing system provides one integrated data point for each marker. Thus increasing the number of cameras in an outside-in tracking system will only increase the accuracy and range of the system and have no effects on the effort to invest for the analysis of the data.

Inside-out tracking reverses the arrangement of tracking devices and markers: the tracking devices are attached to the target and the markers are distributed over the environment, e.g. the ceiling. This kind of setup can reduce costs if only few positions need to be tracked. It also enables tracking in large interaction spaces which are too large to be covered with an outside-in tracking system. The tracking devices, however, are larger and less lightweight than the corresponding markers and they will be more expensive and more fragile. This will increase the obtrusion of the tracking technology experienced by the tracked participant and thus might negatively influence the performance to be observed. Well-known examples for inside-out tracking are the GPS system used for navigation or the Wii-Remote (Wii 2006) developed by Nintendo for their Wii gaming console. In case of the Wii-Remote, however, the rather large sensor has to be held in the hand during the recordings.

Figure 2: Ascension's Flock of Bird is the most prominent magnetic tracking solution. The black box in the back generates a magnetic field in which the sensor presented in the front can determine its position and orientation.

A special case of marker-based inside-out tracking are magnetic tracking systems, such as the Ascension Flock of Birds (ATC 1986, see Figure 2). These systems have sensor devices, which are attached to the target, but they do not use discrete markers in the environment. Instead, they have an outside unit which produces a magnetic field that covers the interaction space (3 m x 4.5 m) and in which the sensors can measure their position and orientation. The advantage of these systems is that they have no need of a clear line of sight to markers and provide a high update rate greater than 100 Hz. Current systems, such as Ascension MotionStar, can operate wired or wireless and provide up to 20 sensor positions per target. The accuracy of these systems depends

on the structure of the magnetic field and is typically better, the closer the sensor is to the field generating unit.

2.2. Marker-less tracking systems

Marker-less tracking systems only require the sensor devices and no artificial enhancements of the environment or the target. They also come as inside-out or outside-in systems. Examples for inside-out systems are *inertial trackers*, which measure relative accelerations from which changes in translation and orientation can be derived (or to be more precise: integrated). The disadvantage of these systems is that they only measure relative movements and are subject to drift errors, which need to be accounted for.

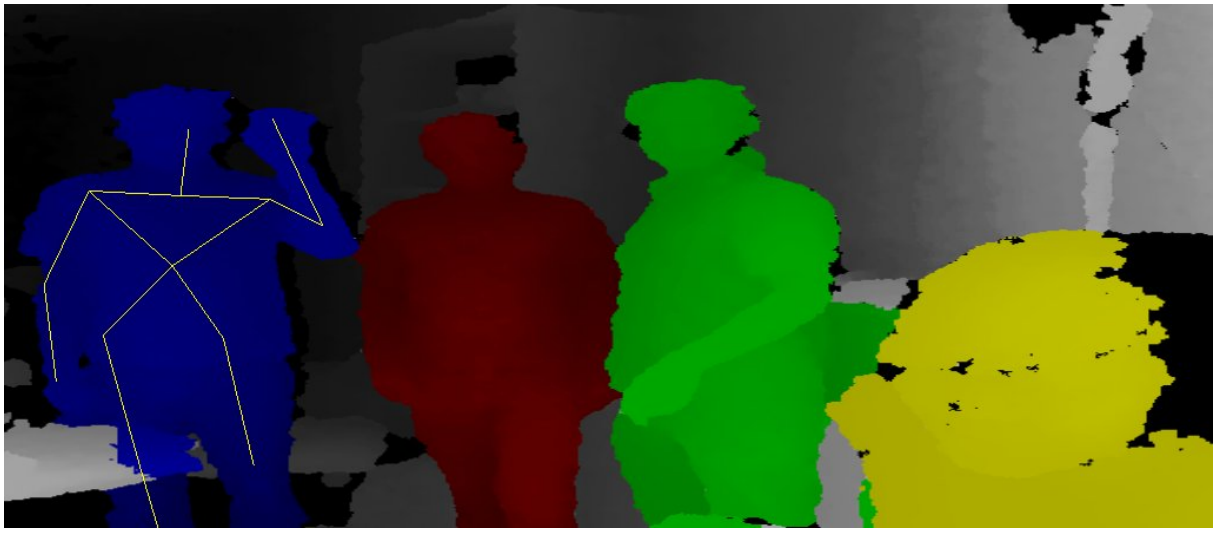

Figure 3: Example of a segmentation of four persons based on the depth image provided by a Microsoft Kinect using marker-less tracking. The basic depth image data is visualized as a grey-scale image and brighter colors represent areas closer to the Kinect. The segmentation and detection of persons, here overlaid using colored regions, was computed by OpenNI (ONI 2011). The person to the left also had OpenNI's skeleton tracking activated.

Marker-less outside-in tracking systems observe the movement of the body from a distance. Most of the available systems operate in the visual domain. For an overview of purely *vision-based techniques*, see Wang and Singh (2003), Moeslund, Hilton and Krüger (2006), Poppe (2007) or Poppe (2010).

A recent prominent example of an outside-in tracking system is the Microsoft Kinect (MSDN 2010), an interaction device based on a *depth camera* produced by PrimeSense (PS 2005), which actually uses a kind of marker, a pattern of structured light which is projected onto the target and whose distortions are measured to extract depth information. The Kinect does not require any attachments to the target. As a first result, the Kinect provides a depth image that is represented as a greyscale image where the individual intensities encode the depth of the first object hit by the light (see Figure 3). The provided software frameworks, Microsoft Kinect SDK (MSDN 2010) for Windows or OpenNI (ONI 2011) for Windows and other platforms, analyze the depth image and extract skeletal information in a second step (see Figure 3, person to the left). This skeleton model is still rather coarse, as can be seen in Figure 3, and does not contain hands, fingers or the orientation of the head. This technology is rather new, so more precise versions of Kinect-like systems and better software frameworks for skeleton extraction can be expected.

3. Representation formats and tools for analysis

Besides device-specific raw data formats, a small collection of standard formats exists. One of these formats is the Biovision Hierarchy (BVH) format. This format has been originally developed by the motion tracking experts at Biovision for their data exchange and has been

widely adapted. The file format is split into two sections. A short example for the left arm is given below in Listing 1.

The first section (HIERARCHY) is used to specify a skeleton based on a hierarchy of joints (JOINT, see example below), starting from a common root (ROOT). Each joint has a certain fixed position (OFFSET) which can be specified as offset relative to the parent joint. Besides the fixed offset, channels (CHANNELS) with dynamic data can be defined as well. The end of such a chain of joints is marked by an end position (End Site) which can have a last offset.
After the specification of the skeleton, the captured data is specified as a table in the section (MOTION). First, however, the number of frames and the frame duration is specified. Within each row, the data is arranged in columns with a layout corresponding to the channels specified in the hierarchy. In the example, the correspondences have been specified using subscripts – which are not part of the plain ASCII-file format but only used here for educational purposes.

Listing 1: Example of a Biovision Hierarchy file

```
HIERARCHY
ROOT Hips
{
  OFFSET Xhip Yhip Zhip
  CHANNELS 6 Xpositionhip Ypositionhip Zpositionhip Zrotationhip Xrotationhip
Yrotationhip
  JOINT Chest
  {
    OFFSET Xchest Ychest Zchest
    CHANNELS 3 Zrotationchest Xrotationchest Yrotationchest
    JOINT LeftCollar
    {
      OFFSET XLeftColloar YLeftCollar ZLeftCollar
      CHANNELS 3 ZrotationLeftCollar XrotationLeftCollar YrotationLeftCollar
      JOINT LeftUpArm
      {
        OFFSET XLeftUpArm YLeftUpArm ZLeftUpArm
        CHANNELS 3 ZrotationLeftUpArm XrotationLeftUpArm YrotationLeftUpArm
        JOINT LeftLowArm
        {
          OFFSET XLeftLowArm YLeftLowArm ZLeftLowArm
          CHANNELS 3 ZrotationLeftLowArm XrotationLeftLowArm YrotationLeftLowArm
          JOINT LeftHand
          {
            OFFSET XLeftHand YLeftHand ZLeftHand
            CHANNELS 3 ZrotationLeftHand XrotationLeftHand YrotationLeftHand
            End Site
            {
              OFFSET XEnd YEnd ZEnd
}}}}}}
MOTION
Frames:     30
Frame Time: 0.033333
Xpositionhip Ypositionhip Zpositionhip Zrotationhip Xrotationhip Yrotationhip
Zrotationchest Xrotationchest Yrotationchest ZrotationLeftCollar XrotationLeftCollar
YrotationLeftCollar ZrotationLeftUpArm XrotationLeftUpArm YrotationLeftUpArm
ZrotationLeftLowArm XrotationLeftLowArm YrotationLeftLowArm ZrotationLeftHand
XrotationLeftHand YrotationLeftHand
```

There are several alternatives to the BVH file format, such as the Hierarchical Translation Rotation (HTR) and the hierarchy-less Global Translation Rotation (GTR) formats used by the company Motion Analysis (MA 1982). A more recent format is GMS (Luciani et. al. 2006) which is more compact as it uses a binary representation and is also more flexible than BVH. It is, however, less widespread and thus has not gained a similar support as BVH yet.

The National Institute of Health defined a format mainly targeted at biomechanical research which is called Coordinate 3D (C3D 1987). It is a binary format and supports a large variety of data well beyond the pure 3D position and orientation data described in the BVH format. A C3D file can include data from EMG, force plates, patient information, analysis results, such as gait timing, and is extensible to support new data.

Formats that are quite popular are the commercial Autodesk FBX format (FBX 2012), which is a binary format that can be accessed and manipulated using the FBX SDK, or the Collada format (COLLADA 2012), which is an open format based on XML (Bray et al. 2008). The gaming company Acclaim developed their own motion capturing system and defined two file formats, Acclaim Skeleton File (ASF) and Acclaim Motion Capture data (ACM), to store the recorded data (Schafer 1994). These file formats has been adopted by Oxford Metrics (OMG 1984) for their Vicon system (VMS 1984). An advantage of the ASF/ACM files is, that they are text-based ASCII files and thus human readable.

Gestures are typically analyzed by annotating a visualization of the data recordings. Please view Section 4 in **Chapter 72a** on documentation with data gloves for a description of software tools that support the annotation of gesture recordings using motion capture.

4. Examples of mocap studies

One relatively well document example of the use of motion capturing in linguistic research is the project "CAREER: Learning to Generate American Sign Language Animation through Motion-Capture and Participation of Native ASL Signers" by Matt Huenerfauth (for more examples on machine learning on ASL see Loeding, Sarkar, Parashar and Karshmer 2004). Huenerfauth and Lu (2010) collected a corpus on American Sign Language (ASL) for improving software for the generation of ASL. They were especially interested in spatial reference points created by the speakers in ASL. Spatial reference points are locations in 3D signing space that are associated with entities under discussion and that serve as reference points later.

For their investigations, they combined motion capture recordings with other technologies. They used two Immersion CyberGloves for recording the hand and finger movements, an ASL H6 eye-tracking system, an Intersense IS-900 (ultrasonic inside-out tracking) for absolute head-tracking and an Animazoo IGS-190 bodysuit for relative motion tracking of the body (interial/magnetic inside-out tracking). In previous work Huenerfauth (2006) recorded motion capture for ASL in a study and achieved only poor quality, so that ASL readers could barely understand the gestures. Apparently, reasons for this were dropped connections, noise and poor calibration. In their most recent work, they thus emphasize the need for a proper selection, setup and calibration of the equipment. One way to achieve this is by elaborating and standardizing the procedures used in this phase of a study as thoroughly as possible. For example, Lu and Huenerfauth (2009) report on a special protocol they developed for calibrating the data-gloves (Immersion CyberGloves) used in their studies.

In our own study on manual pointing (Kranstedt et al. 2006) we used an optical tracking system from ART (ART 1999) to record the pointing movements of speakers when referring to objects in a naming game task. These references to the environment bring in another quality of research that can be done with motion capture: by creating an abstract 3D model of the experiment setup, we were able to relate the pointing gestures to the target objects. This way we could measure the precision of a pointing gesture automatically and evaluate different models in data-driven simulations to derive a model of the intended pointing direction from the recorded gesture trajectory and hand shape (Pfeiffer 2011).

Other well documented studies using motion capturing have been recorded by the Spontal project (Beskow et al. 2009) and by the POETICON project (Pastra et al. 2010).

5. Best practices for study design

Common configurations of tracking systems include an outside-in marker-based tracking for full body movements and data-gloves if the fine grained movements of the fingers are to be observed (see **Chapter 72.a**). Only few systems support both body and finger movements with the same tracking technology, thus the body tracking is often done visually and finger tracking using data-gloves measuring changes in flexible cables. If only few data points are required, such as position and orientation of the hands, 6 DoF markers are of great use.

Every tracking system has a *coordinate system* as reference. If several tracking systems are to be combined, it is essential to have a reliable mapping from one coordinate system to the other. Also, the order, orientation, and scaling of the axes might be different from one system to the other. The position and orientation of the origins of the coordinate systems might be located depending on the research question. If, e.g., participants are interacting with the environment, the coordinate system might best be located at an absolute position within the interaction space (with reference to the room). In other situations, e.g. when co-verbal gestures of a speaker are of interest, the origin of the coordinate system might best be located at a specific point of the participant's body, such as close to the pelvis bone or the center of mass on the level of the chair if the participant is seating (with reference to the body). With the latter choice, issues of laterality can be analyzed more easily (e.g. data from left-handed participants could be flipped by negating one axis) and data from different participants can be compared more easily. It could also be interesting to normalize gestures, by taking the length of the arms and the height of the body into account and scaling the recorded data accordingly.

At the beginning of a recording session, it is helpful to generate some artifacts that are recorded by all employed recording technologies to provide grounds for the *synchronization* of the multimedia material. One way of achieving this is by asking the participants to make certain exaggerated movements (Lu and Huenerfauth 2010). Alternatively, a special device, such as the clapperboard used in film-making, can be used, to which markers from the tracking system are attached. In some conditions it is also extremely important to get a precise measurement of the different body parts of the participants. This is necessary whenever skeletons have to be fitted onto the recorded point clouds.

Most motion capture recordings will include some *noise*, i.e. missing marker positions, orientations or little jumps of marker data. This noise is the result of partial occlusions or of targets leaving the tracking range of the sensors. Depending on the later usage of the data, it might be relevant to add a post-processing data cleanup procedure to identify and correct the data, if possible. A thorough preparation and conduction of a motion capturing study should thus be combined with a control of the quality of the actually recorded data. While problems in audio or video recordings are relatively easy to identify, the evaluation of motion capture data is more difficult. One straight-forward way is to generate statistics about dropped markers and general noise (e.g. jumps). For most applications, a visual review of the recorded data based, e.g., on a data-driven animation of a stick figure or an animated human-like virtual avatar will provide best results. The animation of a virtual avatar, however, will be rather costly to realize in terms of effort and time. In Pfeiffer, Kranstedt and Lücking (2006) we describe an approach where we visualized the recorded data in an immersive Virtual Reality environment that allowed us to evaluate a life-sized 3D multimodal visualization of all data recorded in our study on pointing gestures simultaneously in an integrated view. This helped us to assess the quality of the different data channels (motion data, video, audio, speech transcription, and gesture annotation) in one place and led to an improved quality of the created corpus.

6. Conclusion

Motion capturing technology paves the way for large corpora of quantitative data on gestures at high spatial and temporal resolutions. The temporal and spatial precision of today's motion capturing systems is higher than anything that has been archived based on the annotation of video recordings, and systems are still improving.

These advantages, however, come at a cost: tracking equipment is still expensive and it requires some expertise to be set up and used. Also, pure tracking data is not as easily evaluated as video recordings. With some experience, these investments are compensated by a greatly reduced time required for the analysis of the data, as the manual annotation of the recordings is reduced to a minimum, e.g. to identify and label the time intervals of interest.

The increased interest in game production and consumer electronics has also led to a forthcoming development. Basic tracking systems already have a four-digit price tag and more advancement can be expected. Consumer tracking systems, such as the Microsoft Kinect, allow for a tracking of simple skeleton models in a restricted interaction volume. Solutions that combine several of such systems to extend precision and interaction volume are also available. In the near future, we can expect high quality tracking systems to be found in nearly every household. As consumers are also not too enthusiastic in attaching markers, the most successful systems will be based on unobtrusive marker-less tracking technology. And, maybe even more important, we will become attuned to being tracked by such systems through everyday exposure – maybe even more than we are used to being videotaped – and thus we will be less affected by the use of motion capturing in experimental settings.

Motion capturing, however, is not the panacea for linguistic research, as the following last example underlines. Martell and Kroll (2007) used a machine-learning approach to identify gesture phases in a pre-annotated video-based corpus. Their annotation of the gestures is based on FORM. In a first study, they pre-annotated the positions of the end-effector (the hand) within a 5x5x5 grid manually and were able to train a Hidden Markov Model to detect gesture phases – but only with moderate results. In a second study, they compared the use of motion-capture data with the manual annotations for the same problem. They expected that the fidelity of the motion capture data would lead to more successful classifications. However, the opposite was the case and manual annotations performed better than motion capture as basis for the machine learning process. They explained this by a smoothing of the data done by the human annotators that did not happen with the raw motion capture data. In addition, the annotation scheme based on the discrete 5x5x5 grid abstracted away from detailed trajectories and thus the annotation of many different gesture paths look the same. This more coarse representation, a result of the preprocessing by human annotators, could have been easier to learn by the machine learning algorithm. The punchline is that one has to be careful when optimizing away human involvement from an analytical process, as we might not yet have penetrated the topic deep enough to mold our knowledge into an algorithm and leave the machine alone to handle the data.

7. Literature

ART 1999 *Advanced Realtime Tracking GmbH*, online: http://www.ar-tracking.de (last access January 2012)

ATC 1986 *Ascension Technology Corporation*, online: http://www.ascension-tech.com/ (last access February 2012)

Beskow, Jonas, Jens Edlund, Kjell Elenius, Kahl Hellmer, David House, and Sofia Strömbergsson 2009 Project presentation: Spontal – multimodal database of spontaneous dialog. In *Proceedings of FONETIK 2009* (pp. 190-193). Stockholm: Stockholm University.

Bray, Tim, Jean Paoli , C. M. Sperberg-McQueen, Eve Maler, and François Yergeau 2008 *Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation*, 26 November 2008. W3C.

C3D 1987 *The 3D Biomechanics Data Standard.* Online: http://www.c3d.org/ (last access: January 2012).

COLLADA 2012 *Khronos Group, COLLAborative Design Activity*, online: https://collada.org (last access January 2012)

Edlund, Jens and Jonas Beskow 2010 Capturing massively multimodal dialogues: affordable synchronization and visualization. In *Proceedings of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)* (pp. 160 - 161). Valetta, Malta.

Edlund, Jens, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson, and David House 2010 Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (pp. 2992 - 2995). Valetta, Malta.

FBX 2012 *Autodesk FBX*, online: http://www.autodesk.com/fbx (last access January 2012)

Huenerfauth, Matt 2006 Generating American sign language classifier predicates for English-to-ASL machine translation. Ph. D. dissertation, University of Pennsylvania.

Huenerfauth, Matt and Pengfei Lu 2010 Eliciting spatial reference for a motion-capture corpus of American Sign Language discourse. In *Workshop on the Representation and Processing of Signed Languages (LREC 2010)*. 121-124.

Kranstedt, Alfred, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth 2006 Deictic object reference in task-oriented dialogue. In *Situated Communication*. Berlin: Mouton de Gruyter. 155-207.

Loeding, Barbara L., Sudeep Sarkar, Ayush Parashar, and Arthur I. Karshmer 2004 Progress in automated computer recognition of sign language. In *Computers Helping People with Special Needs,* 3118, 624-624. Berlin/Heidelberg: Springer Verlag.

Lu, Pengfei and Matt Huenerfauth 2009 Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility,* 83-90. ACM.

Lu, Pengfei and Matt Huenerfauth 2010 Collecting a motion-capture corpus of American Sign Language for data-driven generation research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies,* 89-97.

Luciani, Annie, Matthieu Evrard, Damien Couroussé, Nicolas Castagné, Claude Cadoz, and Jean-Loup Florens 2006 A basic gesture and motion format for virtual reality multisensory applications. In *Proceedings of the 1st International Conference on Computer Graphics Theory and Applications*. Setubal.

MA 1982 *Motion Analysis*, online: http://www.motionanalysis.com/ (last access January 2012)

Martell, Craig and Joshua Kroll 2007 Corpus-based gesture analysis: an extension of the form dataset for the automatic detection of phases in a gesture. In *International Journal of Semantic Computing,* 1, 521. World Scientific Publishing Co.

Moeslund, Thomas B., Adrian Hilton, and Volker Krüger 2006 A survey of advances in vision-based human motion capture and analysis. In *Computer Vision and Image Understanding,* 104, 90-126. Elsevier.

MSDN 2010 *Microsoft Kinect SDK*, online: http://www.microsoft.com/en-us/kinectforwindows/ (last access January 2012)

Müller, Meinard, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber 2007 *Documentation Mocap Database HDM05*. Online: http://www.mpi-inf.mpg.de/resources/HDM05/ (last access January 2012). Universität Bonn.

Schafer, M. 1994 *ASF Acclaim Skeleton File Format.* online: http://mocap.co.nz/downloads/ASF_spec_v1.html (last access January 2012)

OMG 1984 *Oxford Metrics Group*, online: http://www.omg3d.com (last access January 2012), Oxford, UK

ONI 2011 *OpenNI*, online: http://www.openni.org/ (last access January 2012)

Pastra, Katerina, Christian Wallraven, Michael Schultze, Argyro Vataki and Kathrin Kaulard 2010 The POETICON corpus: capturing language use and sensorimotor experience in everyday interaction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA). online: http://poeticoncorpus.kyb.mpg.de/ (last access January 2012).

Pfeiffer, Thies 2011 *Understanding Multimodal Deixis with Gaze and Gesture in Conversational Interfaces*. Aachen, Germany: Shaker Verlag.

Pfeiffer, Thies, Alfred Kranstedt, and Andy Lücking 2006 Sprach-Gestik Experimente mit IADE, dem Interactive Augmented Data Explorer. In *Dritter Workshop Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR,* 61-72. Aachen, Germany: Shaker Verlag.

Poppe, Ronald 2007 Vision-based human motion analysis: an overview. In *Computer Vision and Image Understanding*, 108, 4 – 18.

Poppe, Ronald 2010 A survey on vision-based human action recognition. In *Image and Vision Computing*. 28, 976 – 990.

PS 2005 *PrimeSense Ltd.*, online: http://www.primesense.com/ (last access January 2012)

VMS 1984 *Vicon Motion Systems*, online: http://www.vicon.com/ (last access January 2012), Oxford, UK

Wang, Jessica JunLin and Sameer Singh 2003 Video analysis of human dynamics – a survey. In *Real-time Imaging,* 9, 321-346. Elsevier.

Wii 2006 *Nintendo Wii Remote Gaming Controller*, online: http://wii.com/ or http://en.wikipedia.org/wiki/Wii_Remote (last access February 2012)