Capturing full body motion

Distributed Systems Seminar 2013

Antoine Kaufmann
Student ETH Zürich
antoinek@student.ethz.ch

INTRODUCTION

In general the term motion capture refers to the act of recording the movement of objects in varying levels detail [17]. This report discusses the more specific task of capturing and analyzing human full body motion. Note that the term motion capture is used both in the general sense and also when referring only to human motion, especially in relation to character animation for movies and video games.

The idea of capturing the motion of a human actor in order to create more authentic computer animations of characters has been adopted in the late 1970s [14]. And this continues to be one of the most important applications, with many commercial systems being available by companies such as Vicon [3] or Qualisys [4]. Motion capture is used for animating characters in movies with recent examples including Avatar [6] and The Hobbit [13], but also in video games such as Bioshock. There are also other applications such as interaction with video games as enabled by the Xbox Kinect system [2], as well as virtual and augmented reality applications. Motion capture techniques are also used in medicine for example for gait analysis [9] or in sports.

This reports provides a survey of techniques for capturing full body motion, and includes a discussion of three relatively recent research projects in this area.

EXISTING SYSTEMS

Motion capture has been a very active area of research. A survey of just the subset of computer vision based systems by Moeslund et al. [10] lists more than 350 papers published between 2000 and 2006. In order to get an overview of these systems, different criteria are commonly used for classification. One of the most commonly used criteria is based on the physical medium and the types of sensors the respective system is using. Welch and Foxlin published a survey [16] of motion capture systems, where the usual advantages and limitations of different system types are discussed. Another criterion discussed by Welch and Foxlin is the classification into inside-out and outside-in looking systems, depending on whether the sensors are placed on the actor or fixed in the environment. The remainder of this section will discuss the different types of systems.

System types

Mechanical At the same time one of the most direct and first approaches used [14] for motion capture is mechanical sensing. The basic idea is to use mechanical sensors such as potentiometers in combination with an exoskeleton to find joint angles. While providing a way to directly measure the relative position of the limbs without requiring a lot of post-processing, global root motion (basically the movement of an actor through a room) cannot be recovered usually. Also the exoskeleton is uncomfortable to wear and can limit the range of motion. There are commercial systems available such as for example the Gipsy system by Animazoo [1].

Inertial Gyroscopes and accelerometers can be used for capturing motion as well, usually multiple sensors are combined into one inertial measurement unit or IMU. Multiple IMUs are used to track different body parts. The main problem with these systems is *drift*, since the sensors only report relative position changes but not the absolute position. Drift is caused by inaccuracies in the sensors and possibly also during post-processing. Drift can be eliminated by combining different techniques such as IMUs and ultrasound [15]. Commercially available systems include Xsens MVN [5].

Acoustic Sound can also be used as a basis for tracking motion, either by emitting acoustic waves from fixed positions in the environment and placing trackers on the actor, or vice versa, and using time-of-flight to get positions. Problems include reflections which cause a signal to be received repeatedly and in combination with the relatively slow propagation speed limit the number of sounds that can be sent out and tracked per time unit. Other difficulties include ambient noises (can be dealt with by using ultrasound), and the fact that sound waves are affected by wind outside. One major advantage of acoustic system is that sound waves do not require a direct line of sight between the source and the sensor.

Magnetic Another category of systems uses magnetometers or electromagnetic coils to detect the orientation or movement in a magnetic field. The magnetic field used as a basis can either be the natural magnetic field of the earth or artificial. Magnetic fields have the advantage that they pass through the human body without being affected, but on the other hand metallic objects in the field can change the field properties.

Optical In optical systems cameras and other optical sensors are used to track light sources or reflections. The dominant group of systems in this category today are *camera and marker* based systems, where there are cameras fixed in the environment, and the actors are marked with reflective (passive) markers or LEDs (active) markers that are tracked using the camera. Usually multiple high-framerate and high-resolution cameras, very specific lighting conditions, and good contrast between markers and actor is required to achieve high accuracy. Common problems with these systems are occlusion of markers (especially with multiple actors) and marker-swapping, meaning a difficulty of

reacquiring markers when multiple markers are temporarily occluded. This can be addressed by using actively controlled markers that allow the system to identify the markers, but such systems are usually limited in the number of markers that can be tracked. Occlusion is usually addressed by using more cameras. Note that also other optical devices can be used such as e.g. rotating laser beams and photosensors. Many Commercially available optical motion capture systems exist by companies such as Vicon [3] or Qualisys [4]. Two of the systems introduced later in this report fall into this category.

Electromagnetic The final category discussed by Welch and Foxlin is systems based on radio or microwaves and electromagnetic waves in general. A well known example on a larger scale is GPS. The main difficulty with using radio waves for accurate motion capture is their very fast propagation, which makes estimating time of flight challenging. But these systems again have the advantage of not suffering from occlusion. The last system introduced in this report is based on electromagnetic waves.

PRAKASH

In contrast to most optical systems, the system developed by Raskar et al. [11] does not rely on cameras observing a scene. The components used are the two most basic optic building blocks, LEDs as light sources and photosensors to receive light. LEDs are combined with grey code slides to build projectors are used for *space-labelling* and are placed fixed in the scene. The tags to be tracked on the other hand are built from photosensors and a micro controller that is able to determine the position of the tag using the pattern emitted by the projectors. The resulting system allows for a high frame rate and accuracy while keeping the cost much lower than comparable camera and marker based systems, also issues such as sensitivity to ambient light or marker swapping can be avoided. Another benefit of the system is that more information is recovered for each tag than just the position, it also allows determination of the orientation and illumination of the tag.

Approach

As mentioned above, the instrumentation required for the system consists of two parts: projectors and tags. A projector consists of multiple beamers which in turn consist of an LED, an assigned gray code slide and some optics. Infrared LEDs are used so the light is invisible, also the LEDs are modulated at a high frequency to allow the tags to differentiate between light from the projectors and ambient light. A tag consists in its most basic form just of a photosensor and a micro controller analysing the data it gets plus a transmitter that can send the location of the tag to a central system. Both a projector and a tag are shown in figure 1. Note that the tag shown there consists of multiple photosensors besides just the location sensor, which will be discussed below.

Location tracking The basic idea in this system is space-labelling which is basically doing binary search in space using the gray codes (also shown in figure 1), by turning on one LED (beamer) after another. This allows the tag to deduce its position using just the photosensor since every LED cuts the space in half where the tag can be located. A single projector

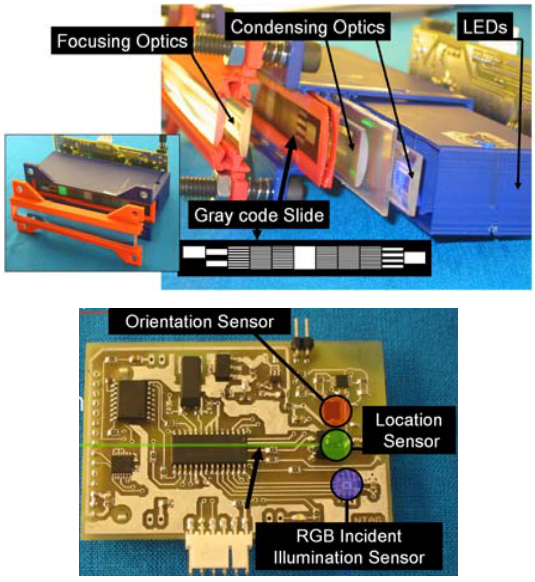

Figure 1: Projector (top) and tag (bottom) used in the Prakash system [11].

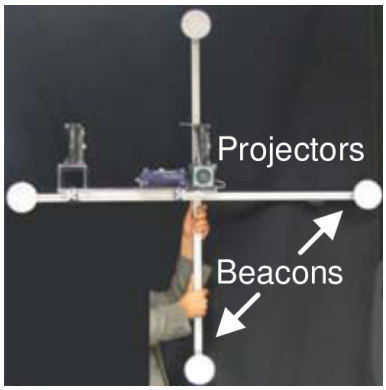

Figure 2: Arrangement of beacons and projectors used to recover 3D position and orientation [11].

only allows for determining the location in one dimension, thus in order to recover a full 3-dimensional position at least 3 projectors are required to allow triangulation (one possible arrangement of projectors is shown in figure 2).

Note that the projectors are not actively controlled by the tags, and each tag can get its position independently of other tags, thus allowing for an arbitrary number of tags, without increase in latency. Another consequence of the tags finding their own position is that there are no problems with tag reacquisition when multiple tags vanish behind an object and reappear, since every tag has a unique identifier.

Orientation and illumination In addition to the position the tags are able to determine their orientation and the RGB illumination information. The orientation is determined using a number of bright IR beacons located in the scene (figure 2) with known positions and a flat photosensor without a lens on each tag. The beacons are again turned on one after another, and the tag will compare the measured brightness of

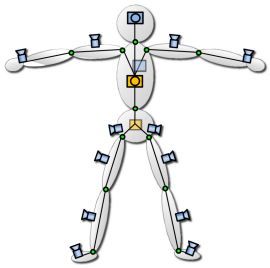

Figure 3: Body-mounted cameras and skeleton [12]

each beacon which enables it to determine the position using cosine-falloff. Note that a rotation around the optical axis cannot be detected. RGB illumination is detected by another group of 3 photosensors on the tag with different color filters, which allow recovery of the light intensity and color.

Summary

The Prakash system allows for motion capture at a high frame rate (500Hz for the prototype mentioned in the paper) and very low latency (below 1ms) using cheap hardware. Raskar et al. demonstrated the applicability in different scenarios that are at least challenging if not impossible for traditional optical motion capture systems, such as tracking an actor moving in daylight wearing regular clothes with tags imperceptible to the camera. Another experiment shows the use of orientation and illumination data: an actor is moving an instrumented prop sword through a scene lit by different colors, and the data is used to accurately render a virtual sword with correct lighting and even showing slight rotation across the axis of the sword as performed by the actor.

BODY-MOUNTED CAMERAS

Shiratori et al. [12] tried to avoid the need for instrumentation of the environment, to allow free capture outside, such as for example jogging in a park. Their basic idea is to use a number of cameras that are placed on the actor facing outward and then using computer vision methods to recover the motion of the actor from the footage captured by the cameras. The system is able to provide both the global root motion of the actor and the relative motion of body parts.

Approach

The actor is instrumented using 16 or more wide-angle cameras strapped to different parts of the body. Multiple cameras are used for body parts that are often occluded to increase the chance of getting a usable image due to the bigger field of view. A digital skeleton of the actor (see figure 3) with the cameras is then created by performing a predefined range-of-motion exercise and some manual tweaking if necessary.

Structure-from-motion Figure 4 shows the process used to recover the whole body motion from the footage recorded by the body-mounted cameras. The first step is to use structure-from-motion (SfM) to get a 3D representation of the scene using reference images of the scene that are taken beforehand. Note that SfM could also be applied to the images from the body-mounted cameras directly, but this approach leads to substantial drift.

Camera registration This 3D representation is then combined with the footage from the cameras to calculate the positions of the cameras, in a process called absolute camera registration. Since this won't always yield positions for every camera, for example when the viewpoints of the reference images differ significantly from the viewpoints of the camera in question, another step is used: relative camera registration. The idea behind this step is to use commonalities between the unregistered cameras i.e. the cameras with unknown poses, and the registered cameras in order to estimate a pose. Depending on the number of unregistered cameras this has to be iterated multiple times. Note that in case of occlusion or motion blur there might still be unregistered cameras after relative camera registration.

Global optimization The purpose of global optimization is to make sure that the constraints imposed by the skeleton are respected and the motion is temporally smooth. This is expressed as an optimization problem of minimizing the sum of the reprojection error (calculated using the new camera poses and the 3D structure) and the smoothness of the global root motion and joint angles (roughly the differences of the angles and positions between frames). A comparison of the results with and without global optimization shows that it significantly reduces noise in the output and also improves the accuracy.

Summary

In a comparison with an industry standard Vicon camera and marker based system Shiratori et al. show that their system comes reasonably close while being much more portable. Demonstrations include motions such as running outside or swinging on monkey bars in a park, both of which are hard to capture using existing optical motion capture systems. One drawback of the system is the amount of calculation that is required for processing the raw footage and getting the output, they report that a minute of capture requires about a day of processing with their prototype system. Also their approach will become more useful as cameras get smaller and cheaper.

HUMANTENNA

Cohn et al. [8] considered a totally different approach also with different applications in mind. The goal of their system is to allow gesture recognition for applications in a home, where it is desirable to avoid the need for instrumentation of every room, and a heavy instrumentation of the user should also be avoided for everyday use. To achieve these goals Cohn et al. have extended the idea of using the human body as an antenna for electromagnetic noise present in the environment which they explored in previous work for sensing touch gestures on walls [7].

Approach

The goal is to recognize the 12 predefined gestures shown in figure 5, also the system should be able to determine which room the gesture was performed in. The required instrumentation consists of an electrode to be placed on the neck and a small device that takes the measurements and transmits them. This allows the system to use the human body as an antenna to pick up EM noise from the environment emitted from power lines and electrical appliances. If the actor performs a gesture, the antenna characteristics and thereby the

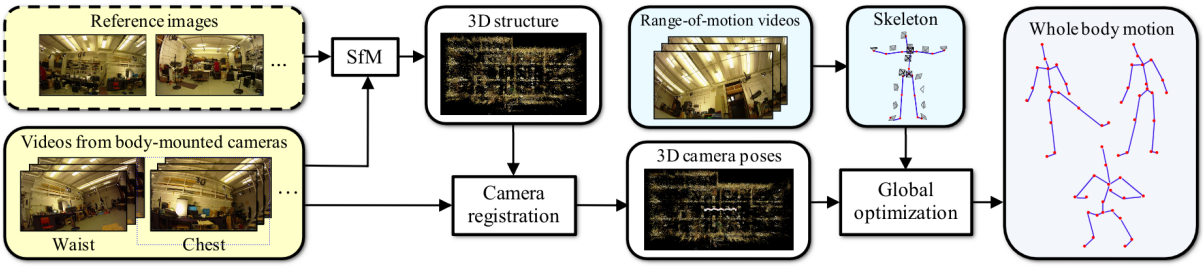

Figure 4: Process used to process the raw video footage from multiple cameras into whole body motion [12].

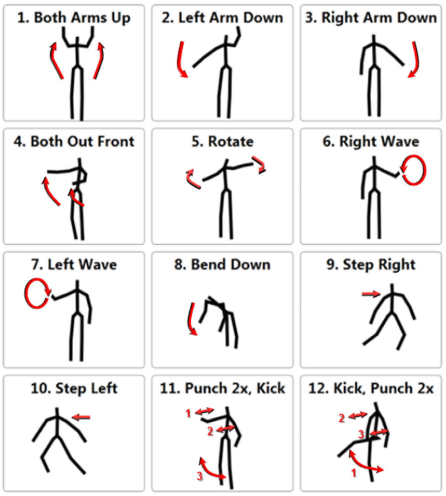

Figure 5: Gestures to be recognized by the Humantenna system [8].

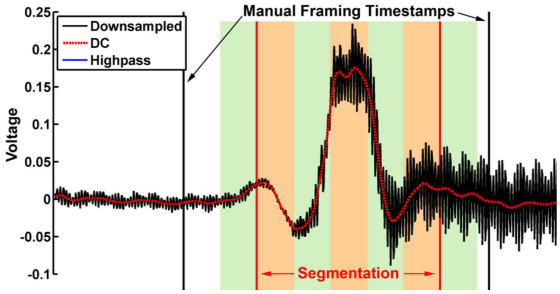

Figure 6: Measured signal for a rotate gesture [8].

measured signal change significantly, also these changes depend on the type of gesture, thus making it possible to use this signal to recognize gestures.

Gesture recognition Figure 6 shows the measured signal for a rotate gesture. A machine learning approach is used to classify the gestures. The classification algorithm consists of 3 phases: segmentation, feature extraction and classification.

For the offline version of the algorithm described here, explicit notifications where given to the system when the gesture starts and ends, but these are not accurate enough. So the segmentation phase determines where a gesture actually begins and ends. This can be determined by looking at the DC waveform of the signal, which is obtained by applying a low-pass filter to the signal. Afterwards the signal is divided into fixed sized windows and for every window a metric is calculated determining if the window is active i.e. the DC signal shows significant changes. The first and last active window are used as start and end of the gesture.

Feature extraction calculates a number of numerical values that will be used to classify the gesture. To that end the gesture is divided into a fixed number of windows. One set of features describes the DC waveform that is consistent across multiple repetitions of the same gesture. Another set of features describes the amplitude of the AC wave that is clearly also changing, and the last set of features describes frequency domain features and is obtained by calculating a fast Fourier transform and dividing the frequencies up into buckets.

In the last step a support vector machine (SVM) is used to classify the gestures using the features calculated in the previous step. Experimental results show accuracies in the range of 90% across multiple homes and participants. Confusion mainly occurs between the symmetric right/left wave gestures. Cohn et al. also extended the offline algorithm described above to work online in real-time with a latency of around 0.4s, allowing an actor to use the system to interact with an application.

Location classification Also using a similar machine learning approach it is possible to determine in which room a gesture was performed. Here it is not necessary to find out when a gesture starts, it is sufficient to take the first 0.5s of a gesture and calculate frequency domain features. Experimentation showed that the most important part for classifying the location are higher frequencies, since these are mostly emitted by appliances. The evaluation showed an accuracy of over 99%.

Summary

Clearly this is a significantly different approach than the ones previously discussed. The goal is not to catch whole body motion as accurately as possible, but to recognize a fixed set of gestures. Also note that the system has to be trained since it is based on machine learning, although the authors suggest that some models could be used across actors reducing the training time. On the positive side this system only requires minimal instrumentation of the user and no instrumentation of the environment.

DISCUSSION

These three systems are all significantly different from each other and also based on different ideas than existing motion capture system. Both the Prakash and the body-mounted camera (BMC) system are optical systems, but other than most optical systems they can be classified as inside-looking-out since both of them use sensors placed on the actor looking at the scene. The Humantenna system is also inside-looking-out and falls into the category of the electromagnetic wave based systems.

Generality While the Prakash system is intended to just track tags on any kind of object, the Humantenna and BMC systems are specifically tailored to tracking human full body motion relying on properties of the human body such as the characteristics as an antenna or biomechanical structure. Humantenna differs significantly from the other systems in the respect that it does not try to capture the exact motion of the actor but only recognizes predefined gestures. Prakash allows almost arbitrary precision when tracking tags.

Instrumentation One aspect that is shared by both BMC and Humantenna is the fact that they avoid the use of any instrumentation of the environment with signal (or light) sources. On the other hand Humantenna uses much lighter instrumentation of the actor. Prakash instruments both the environment and the actor, but the tags can be placed (almost) imperceptibly in the clothing.

Latency In terms of latency, that is the time from when the actor performs some movement until the system has captured and processed the input, there are also significant differences. Prakash is the fastest system of the three with latency below 1ms, and should also scale well to a higher number of tags. Also operating fast enough for real-time interaction is Humantenna, that recognizes gestures with a delay of around 0.4s. The BMC system is in a whole other league in this regard with processing times of about a day for one minute of footage, due to the complex analysis and processing.

Cost All three systems are on the low end of the scale in regards to cost, at least compared to existing commercial optical system, especially if fabricated using commercial processes. Humantenna would probably be the cheapest system, since it basically only requires an electrode and some electronics to amplify and digitize the measured signal. Prakash is also fairly cost-effective since it is based on cheap components such as LEDs and photosensors, but it also requires a larger set of components with projectors and tags that need to be installed. The most expensive system of the three is probably BMC due to the large number of cameras that are required. Cost can be reduced by using cheaper cameras, but this will also be reflected in the quality of the output.

CONCLUSION

Looking at these fairly recent research projects (2007-2012) and their features it seems that the observation made by Welch and Foxlin and used as a paper title in 2002 [16] "Motion Tracking: No Silver Bullet but a Respectable Arsenal" still holds. All three systems focus on different tasks and have different weaknesses. There is no general purpose motion capture system that can be applied in any situation to capture full body motion, but there are many systems that address different challenges, and more research can be expected to come in this area.

REFERENCES

1. Gypsy 7 electro-mechanical motion capture system. `http://www.metamotion.com/gypsy/gypsy-motion-capture-system.htm`. [Online; accessed 29-April-2013].

2. Kinect - xbox.com. `http://www.xbox.com/en-US/kinect`. [Online; accessed 28-April-2013].

3. Motion capture systems from vicon. `http://www.vicon.com/`. [Online; accessed 28-April-2013].

4. Motion capture mocap qualisys motion capture systems. `http://www.qualisys.com/`. [Online; accessed 28-April-2013].

5. Xsens mvn : Inertial motion capture - xsens. `http://www.xsens.com/en/general/mvn`. [Online; accessed 29-April-2013].

6. Behind the scenes look at the motion capture technology used in avatar. `http://avatarblog.typepad.com/avatar-blog/2010/05/behind-the-scenes-look-at-the-motion-capture-technology-used-in-avatar.html`, 2010. [Online; accessed 28-April-2013].

7. Gabe Cohn, Daniel Morris, Shwetak Patel, and Desney Tan. Your noise is my command: sensing gestures using the body as an antenna. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 791–800, New York, NY, USA, 2011. ACM.

8. Gabe Cohn, Daniel Morris, Shwetak Patel, and Desney Tan. Humantenna: using the body as an antenna for real-time whole-body interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1901–1910, New York, NY, USA, 2012. ACM.

9. C Frigo, M Rabuffetti, DC Kerrigan, LC Deming, and A Pedotti. Functionally oriented and clinically feasible quantitative gait analysis method. *Medical and Biological Engineering and Computing*, 36(2):179–185, 1998.

10. Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126, November 2006.

11. Ramesh Raskar, Hideaki Nii, Bert deDecker, Yuki Hashimoto, Jay Summet, Dylan Moore, Yong Zhao, Jonathan Westhues, Paul Dietz, John Barnwell, Shree Nayar, Masahiko Inami, Philippe Bekaert, Michael Noland, Vlad Branzoi, and Erich Bruns. Prakash: lighting aware motion capture using photosensing markers and multiplexed illuminators. In *ACM SIGGRAPH 2007 papers*, SIGGRAPH '07, New York, NY, USA, 2007. ACM.

12. Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K. Hodgins. Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*, SIGGRAPH '11, pages 31:1–31:10, New York, NY, USA, 2011. ACM.

13. Katy Steinmetz. Gollums getup: How the hobbits groundbreaking technology works. `http://entertainment.time.com/2012/12/05/gollums-getup-how-the-hobbits-groundbreaking-technology-works/`, 2012. [Online; accessed 28-April-2013].

14. David J Sturman. A brief history of motion capture for computer character animation. *SIGGRAPH 94, Character Motion Systems, Course notes*, 1, 1994.

15. Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. Practical motion capture in everyday surroundings. In *ACM SIGGRAPH 2007 papers*, SIGGRAPH '07, New York, NY, USA, 2007. ACM.

16. Greg Welch and Eric Foxlin. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Comput. Graph. Appl.*, 22(6):24–38, November 2002.

17. Wikipedia. Motion capture — wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Motion_capture&oldid=551557698`, 2013. [Online; accessed 28-April-2013].