Using Motion Tracking Data to Augment Video Recordings in Experimental Social Science Research

Jeremy P. Birnholtz[1], Abhishek Ranjan[2], Ravin Balakrishnan[2]

[1]Department of Communication, Cornell University, Ithaca, NY

[2]Department of Computer Science, University of Toronto, Toronto, Ontario

jpb277@cornell.edu

Abstract. While video can be useful as a method for gathering and analyzing social science data, it also has many drawbacks. In particular, recording is constrained by the field of view of a camera, and coding to isolate behaviors of interest can be slow and imprecise. For certain situations, the use of optical motion capture technologies can overcome these obstacles by tracking human behavior in larger spaces and providing systematic, quantitative log data of interesting behaviors. We describe the use of an optical motion capture system in conducting a laboratory experiment using a video-mediated communication system. The system allowed for rapid and accurate analysis of data, and findings that would otherwise have been very difficult to achieve.

Introduction

One goal of the e-science and cyberinfrastructure programs in the social and behavioral sciences is the desire to improve researchers' capacity to interact with and understand large behavioral data sets and new forms of digital records (Berman & Brady, 2005). Data gathering methods have multiplied rapidly in recent years and it is not uncommon to have several types of data describing a single event. For example, one might have video and audio recordings, field notes, and logs of participant behavior.

While presenting many opportunities for detailed analysis and novel discoveries, this veritable smorgasbord of data also raises two critical challenges for researchers. First is the challenge of managing and integrating the data – that is, finding ways to view data from multiple sources in useful and meaningful ways (e.g., seeing simultaneously recorded data at the same time, as opposed to in separate files) that allow for the discovery of novel patterns and relationships. This challenge is the primary goal of projects such as French, et al.'s (2006) "Playback" system that allows for the synchronous playback and manipulation of time-based data streams.

In addition to managing and displaying data, there is the significant challenge of analysis. More data can often mean more time spent on analysis, but additional time does not always result in useful findings. Thus, an increase in the amount of data available and the number of ways of gathering those data increases the imperative to carefully select methods of data gathering and analysis that make good use of available technologies, and can be utilized both effectively and efficiently.

This challenge is particularly prevalent in the gathering of observational behavioral data. As compared with field notes or audio recording, video recording of behavior has vastly improved our capacity to capture and analyze behavior in very detailed and nuanced ways (Heath & Luff, 2000). At the same time, however, video has significant drawbacks. It is limited in scope to the field of view of a camera or cameras; it can be difficult to integrate multiple video sources to capture a larger space; and recordings must generally be viewed and coded in real time by human coders, thereby pinning the reliability of analyses on the subjective perceptions of these coders.

In this paper, we will describe our experience using optical motion tracking technologies to augment video recording in conducting a laboratory experiment on video-mediated communication. This system allowed us to easily understand and visualize detailed behavior patterns that led to novel results that would have been substantially more difficult without the method we describe. Our contributions to the emerging e-social science literature are twofold. We first present a discussion of the tradeoffs involved in assessing methods for the capture and analysis of observational data. Second, we describe the use of a novel method in carrying out such research, and contend that optical motion capture systems are a potentially significant asset to the e-science community.

Background and Related Work

Conducting Observational Research Using Video

The increasing availability of cameras and recording devices has had a substantial impact on the conduct of qualitative social science research (Heath & Luff, 2000). Recordings of video and audio, sometimes using cameras and microphones in locations not suitable for human observers, have made it possible to analyze human behavior at a level of detail not previously possible. While video recordings are useful in that they visually capture all human activity within the field of view of a camera, this can also be problematic in several respects.

First, a video camera only captures what is in its field of view. If it is aimed or focused improperly, or if something surprising occurs outside the range of view, that activity is not captured. Moreover, adding additional cameras to a setting does not necessarily result in a larger field of view, because it can be difficult to understand how different views fit together, and how to accurately measure movement from one camera's visual field to another's (Gaver, Sellen, Heath, & Luff, 1993).

Second, video cameras capture a visual record of activity at a single level of detail. If a higher resolution view of a particular component of a shot is desired, this cannot be achieved retrospectively. It bears mentioning, of course, that high definition video recording makes it possible to capture much more detail, such that one might later zoom in on an area of interest. Nonetheless, the resolution remains finite and constrained by field of view.

Third, video is a visual record of an entire scene, while it is typically only certain behaviors in this scene that are of interest to researchers. Thus, the interesting components of the recording must be identified and annotated in ways that facilitate rigorous analysis. In a small number of cases where conditions are very controlled, this can be done automatically. Quaranta, et al. (2007), for example, were able to automatically identify the angle at which a tightly-confined dog's tail wagged, and Knight, et al. (2006) experimented with extracting specific gestures and facial expressions from recorded videos.

For the most part, however, video recordings must be viewed in real time and coded by human research assistants, who identify specific behaviors and may annotate the recording using analysis tools (e.g., TransAna). If one were interested in the frequency and duration of head turns, for example, as in Kuno, et al.'s study (2007), coders would watch for and identify each turn of the head.

Manual coding of videos is problematic in two respects. First, it is slow and painstaking, since videos must typically be viewed in real time, or even slower if it is detailed behaviors that are of interest (e.g., Heath & Luff, 2000). Second, reliability is called into question both because different coders may differ in their subjective perceptions of phenomena of interest, and because it may be difficult to make accurate and consistent measurements of behavior, given the camera angle and resolution (e.g., "How do we know how far the head turned?").

Behavioral analyses of this nature can be vastly simplified and improved through the use of optical motion capture technologies that allow for automated (or near-automated) analysis of data from a much wider range of settings than traditional video alone.

In particular, such technologies are designed for the detailed tracking and logging of human activity. Thus, the capacity for easy gathering and analysis of motion capture data represent a potentially valuable component of cyberinfrastructure for the social sciences. Such a resource would vastly simplify some existing research methods and enable new forms of research.

Motion Capture and Computer Vision Technologies

Given their capacity to speed up and increase the accuracy of the qualitative research process, motion capture and computer vision technologies are already of clear interest to some in the social science research community. In particular, such techniques have been applied to the analysis of video recordings to detect certain types of head movements (Knight et al., 2006), and to detect certain types of hand gestures for real-time interaction using gesture-based interfaces (Kang, Lee, & Jung, 2004).

One problem with using computer vision to automatically parse video recordings is that elaborate algorithms are required to parse video images and identify scene elements that are of interest (e.g., heads, hands). Moreover, analysis of recordings is constrained by the same problems of field of view and resolution that were outlined above.

The use of optical motion capture technologies with passive markers can overcome many of these hurdles. Optical motion capture systems utilize an array of three or more infrared cameras that sense the presence of reflective plastic markers within their field of view. These systems have been used to capture and understand the details of individual human motion in such fields as animation (Sifakis, Neverov, & Fedkiw, 2005) and human kinesiology (Salarian et al., 2004). In these cases, an individual might wears a suit covered with passive markers, and the details of certain movements or expressions can be captured for analysis or to increase the apparent realism of an animated character's movements. Our approach differs from these in that we track the motion of certain body parts relative to physical objects and, potentially, other individuals in the space. We also use motion capture data together with video data in our analyses.

As we shall describe in more detail below, passive markers vary in size, can be attached to most any physical object, and require no power or other wires (unlike "active marker" systems that require markers to be physically wired). The cameras are carefully calibrated

such that the location of a marker can be identified extremely precisely within the three-dimensional space defined by the overlap of the cameras' views.

Compared with video cameras, optical motion capture systems are substantially less constrained by camera field of view in that the field of view of the entire system can be easily expanded by adding cameras. Because the cameras are calibrated relative to each other and because marker location is determined using simultaneous data from multiple cameras, adding more cameras increases the size of the overall space that is being monitored. In other words, the system calibration means that the entire array of cameras is analogous to one very accurate video camera in that their fields of view combine to form a unified space. The advantage here, however, is that the size of that field of view can be increased by adding, and calibrating, additional cameras.

Further, optical motion capture systems capture data about marker location at an extremely fine level of resolution. The location of a marker can typically be pinpointed with accuracy of one millimeter or less. While it is certainly possible to look at aggregate patterns that do not require this level of detail, one always has the high-resolution data to go back to if necessary.

Finally, optical motion capture systems provide a detailed quantitative description of marker location and movement over time. Quantitative descriptions allow for easy isolation and accurate measurement of behaviors of interest, without the need for video viewing or manual coding.

We will now turn to a detailed description of our specific technique for using an optical motion capture system in conducting a laboratory experiment.

Our Method

What we were interested in

We were conducting an experiment to investigate differences in participants' behavior in performing a video-mediated "remote repair" task across three experimental conditions in which the style of camera operation was varied.

The details of this experiment are described elsewhere and are not the focus of this paper (Ranjan, Birnholtz, & Balakrishnan, 2006), but we will provide a brief summary here. Pairs of participants were brought into the lab and randomly assigned to the roles of "helper" and "worker." The pairs could communicate only by two-way audio and one-way video conferencing (i.e., the helper had a view of the worker's space, but the worker could not see the helper's). The task was to build several objects using Lego bricks. In each case, the worker had only the pieces to use in the construction, and the helper had only the instructions. The task was completed in three experimental conditions, where the type of camera shot available to the helper was the independent factor being varied: 1) a fixed wide shot of the workspace, 2) a helper-controlled pan/tilt/zoom (PTZ) camera, and 3) a PTZ camera controlled by a dedicated operator, intended to simulate automated camera control.

We were particularly interested in evidence of adaptation to the different styles of camera operation, as evidenced, in part, by differences in participants' physical movements and usage of the space. Seeing how our participants utilized the space differently would allow us to better understand both the deficiencies in the different camera operation styles, and as well as how and when people adapt to different technologies.

Such an understanding would help us toward our longer-term goal of developing an automated camera control system based on real-time motion capture data (Ranjan, Birnholtz, & Balakrishnan, 2007), so were also interested in discovering relationships between participant motion and desirable camera shots.

What we did

Our laboratory is equipped with an array of 5 Vicon infrared motion capture cameras with an approximate frame rate of 122 frames per second. These Vicon cameras combine to create a 1.7m x 3.0m x 2.3m space in which all activity can be tracked extremely precisely (with sub-millimeter accuracy) through the use of passive reflective markers. Reflective markers range in size from 5 mm to 20 mm, and can be attached to almost any object (see Figure 1). A marker is visible to a Vicon camera if there is no other object present on the imaginary line between the camera and the marker. A marker can be tracked in 3D if it is simultaneously visible to at least three of the five Vicon cameras.

One constraint resulting from the use of passive markers is that markers cannot be distinguished from each other, except by location. In other words, consider a simple scenario in which we have two markers, A and B, with the x, y, z coordinates at time T_0 of 3, 3, 5 and 5, 5, 3, respectively. If we exchange the locations of the two markers at time T_1, there will be no difference in system state from T_0 to T_1, because the system cannot tell one marker from another and "sees" one at each of those locations at each time. Understanding this attribute is important for our purposes here, because our interest is in tracking the location and movement of physical objects and human body parts to which markers can be attached. In these cases, tracking requires the ability to uniquely identify such objects.

We accomplished this through the combination of multiple markers (at least three markers) in uniquely identifiable patterns. In our experiments, for example, our participants wore gloves with markers attached to them. Each glove has a unique arrangement of markers (see Figure 1), and we have programmed the system to identify and log the location of this marker arrangement. In this way, we are able to track the exact location of the two hands. Finger-level detail was not of interest to us at this phase, but could easily be achieved through the addition of more markers.

Figure 1. Participant's right hand wearing a glove with reflective markers attached in a unique arrangement.

To log this tracking information for analysis, we updated a text file once per second. Each second, the following information was written to the file:

- **Absolute timestamp –** The time stamp allowed us to know exactly when data were recorded, and facilitated synchronization with other data sources.

- **Right and left hand locations –** The location of the worker's right and left hands allowed us to determine which region the hands were in, and whether or not they were in the PTZ camera shot. Data were recorded in terms of x, y and z coordinates.

- **PTZ Camera Values –** The pan-tilt-zoom settings of the camera was recorded so that we could determine (using the method below) whether or not the worker's hand(s) were in the camera shot at any given moment. If we assume that, most of the time, the helper would want to see what the worker is doing, whether or not the workers hands are in the shot becomes a rudimentary metric of camera operator effectiveness.

In addition, we video taped all of the sessions. To allow synchronization of the video data with the motion capture data, we have developed a "low-tech," but effective, solution. Just after the logging and video recording have started, we tell the worker we are about to turn off the lights and instruct them raise their hand as soon as it is dark. We then count to three and quickly switch the lights off and then on again. The dark video frame can then be matched to the raised hand in the motion capture data.

As mentioned above, we were also interested in whether or not tracked objects were in the view of the pan-tilt-zoom (PTZ) conventional video camera. To determine whether or not an object was in the shot required several steps.

The first step was to manually calibrate the PTZ camera. The camera was mounted on a stationary tripod pointing at the participant's workspace. The workspace was divided into two main regions (pieces region: 25 cm wide, work region: 60 cm wide) and region boundaries were noted. The worker used the work region for constructing Lego objects, and the pieces region was used to place unattached Lego bricks prior to their use in construction. In our analysis, we noticed that participants primarily worked in the center of the work region, so we referred to the periphery of this region as the "intermediate region," and its boundaries were noted as well. For each region, a close-up shot was framed by manually adjusting pan, tilt, and zoom values of the camera. Thus the calibration information consisted of a table of region names and corresponding pan-tilt-zoom values for the camera.

In order to determine if an object is in the shot or not, we read the pan-tilt-zoom value of the camera and the corresponding "shot region" (say, Region 1) using the calibration table. We then determined the "object region," or region where an object of interest was located, by calculating using preset region boundaries and the object's location. If the "shot region" and "object region" matched, we concluded that the object was in the camera shot, otherwise it was not.

Analyses

Behavioral Adaptation

As mentioned above, we were interested in evidence that participants changed their behavior under different camera control conditions, which would suggest that human effort on the part of both helper and worker might be reduced if camera control were automated. Our motion capture data allowed us to explore how participants used the available space, by looking at scatter plots of worker hand position on the x, y and z planes.

To create these scatter plots, we created three aggregate data files, with one for each experimental condition. Each file contained all of the time stamped hand location data from all of the participants in one of the conditions. Each unique hand position (measured once per second) was treated as a single "case" (i.e., row in the data file). The aggregate set of hand positions for all participants and at all measurement points could then be plotted for any plane by using the two axes of that plane in the 3-dimensional space[1].

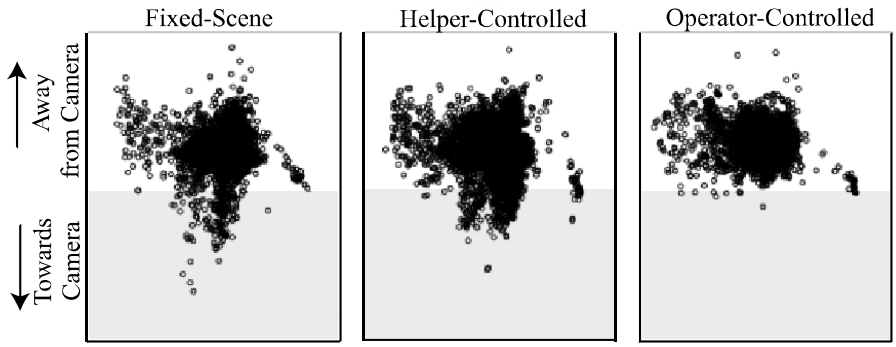

Figure 2. Top view of participant hand locations for all participants in three experimental conditions. Each dot represents a hand location at one point in time (figure reproduced from Ranjan et al., 2006).

Figure 2, for example, shows a top view of all participant hand positions plotted on the x and y dimensions. The participant assigned to the worker role was sitting approximately where the labels are at the top of each plot, facing the camera which was located approximately at the bottom of the plot. What can be seen in this plot is that participants moved their hands in the direction of the camera more in the fixed-scene camera control condition than in the operator-controlled camera condition. This suggested to us that since the camera was not zooming in on their hands to highlight particular objects for the helper to see, workers had to do this on their own – by holding objects up to the camera themselves. This was confirmed by analysis of the data using the vertical plane (x and y axes), in which (as is shown in more detail in (Ranjan et al., 2006)) it is clear that participants were holding their hands at almost the exact height of the camera.

While it would certainly be possible to plot and analyze data in three dimensional space, we did not use this technique in our analyses for this particular experiment.

Camera Control Effectiveness

In order to get a sense of how effective our camera operator was, we wanted to know how often the worker's hand was in the PTZ camera shot, and about how often the camera followed hand movement from one area of the workspace to another.

To do this, we divided the workspace into two large regions which we named the "pieces area" (because it was where the Lego pieces were placed at the start of the task) and the "work area" (because it was the area directly in front of the worker where construction of the

[1] Note that by "plane" we mean the two dimensional space that is defined when all values on the third, unused dimension are considered to be 0. In other words, Figure 2 shows the horizontal plane (x and z axes) in which all values on the vertical dimension (y axis) are considered to be 0.

objects took place). Based on log data about the position of the PTZ camera at each measurement point, we could then easily explore both of these questions.

With regard to the first question about how often the hand was in the camera shot, we simply calculated the fraction of the time that the coordinates of the worker's hand position were within the workspace region associated with the camera shot.

For the second question, we used the logged camera position data and the logged hand position data to create a second data file that, for each measurement point, indicated whether the hand and camera were in the work area, the pieces area, or neither of these. This data file allowed us to create a plot similar to Figure 3, in which it can be seen that the camera followed the worker's hand some of the time, but occasionally moved too late. Sometimes the hand movement was so brief that the camera did not move at all (as in cases labeled 'a').

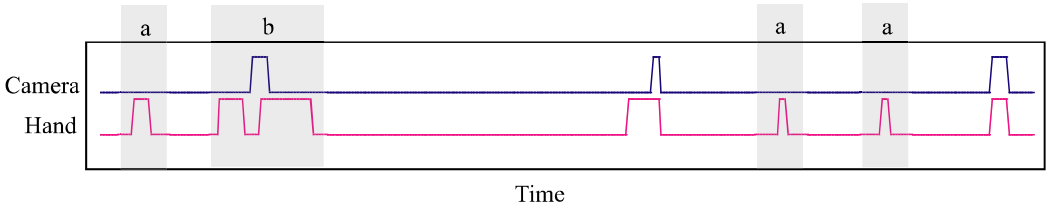

Figure 3. A 500 second snapshot of hand and camera movement in the Operator-Controlled Camera condition. A rise in the plot indicates move to the pieces area and a drop indicates move to the work area. Letters 'a' and 'b' indicate mismatches in camera and hand position (reproduced from Ranjan et al., 2006).

Difficulties

This technique, which we found to be quite valuable, is not without difficulties. One particular problem we encountered was that the Vicon cameras in our laboratory are situated above the participants. Thus, markers must be visible from above if they are to be tracked. This became somewhat problematic when participants rotated their hands such that the markers faced the table. Thus, there were some brief periods of time where we did not have complete tracking data. In subsequent experiments, we have devised tasks in which participants would be less likely to rotate their hands in this way.

Another difficulty with the technology is that Vicon cameras detect markers using the infrared light reflected from the markers. Thus, any other object that reflects infrared could also be detected as a marker, resulting in erroneous tracking. This requires the laboratory space to be free of stray retro-reflective objects, such as watch faces, plastic shrinkwrap, and others. While both of these difficulties ccould be eliminated by using non-optical (e.g., magnetic) motion tracking systems, such systems often use active hardwired markers and are susceptible to interference from electrical sources. In some cases, they are also less accurate.

Our technique also requires that objects be identified by unique marker patterns, which could become difficult if large numbers of objects must be tracked. While active marker systems can help overcome this difficulty, they typically require that markers be attached to wires, which could become quite cumbersome and restrictive in a laboratory environment.

Discussion and Future Possibilities

We have described our experience using motion tracking data in conjunction with video recordings for behavioral analysis of observational data. While we contend that the method itself is useful and has valuable applications, we further argue that it raises an interesting set of issues about the nature and analysis of digital data for the e-social science and cyberinfrastructure communities.

First among these issues is the question of what video really provides to those who use observational data. While video provides a visual record of human behavior, it does so in a way that can be difficult to scale (i.e., to cover a larger area) and parse (i.e., to code and annotate behaviors of interest) in automatic ways. Motion capture technologies such as those described here have the capacity to sidestep these troubles with video by quantitatively capturing only the components of the visual scene that are of interest. This can allow for more complete and accurate capture, which can be augmented by video recording as detailed below.

To be sure, we are not suggesting that video recording and analysis should be eliminated as a social science research method. Rather, we argue that the visual properties of video are maximally exploited when it is used to explore and augment.

By explore, we mean that video is unique in that it provides a comprehensive view of a specific scene, even if this view might lack rigor in certain ways, as described earlier. Where it is unclear what specific behaviors are of interest, such a visual record can be extremely useful. Watching video to spot interesting or unusual phenomena in a pilot study, for example, could allow researchers to determine what movements to capture.

By augment we mean that video can be used to help explain surprising or unexpected events in data from other sources. Using motion tracking technologies, for example, one might see a lot of movement in a particular region of the participants' space, but have no idea what was physically in that region that warranted this movement. Video recordings can help clarify this sort of uncertainty. We also used the video recordings in our study to determine which Lego pieces were being used in some cases where the motion capture data suggested participants were having difficulties.

All of this increases the imperative for novel ways of integrating, reviewing and analyzing multiple forms of time-based data, as described in (French et al., 2006). In the simplest case, one can imagine viewing a video recording and a transcript of a scene along with motion capture data from the same scene overlaid on the video. Once these streams are effectively integrated, however, there are many more possibilities. One could imagine a player application, for example, that includes facilities for space-based playback and search, in addition to time. With such a system one could search the video stream for all scenes where a participant's hand rises above a certain threshold. Or define a region on screen and search for all instances where a tracked object was inside that region. This, of course, raises its own set of issues about visual representation of the motion capture space and how to align this with the video view, but the possible applications are nonetheless quite valuable.

Other possibilities abound as more data gathering technologies become widely available. One could also imagine integrating portable eye tracking and motion tracking data into a single 3-dimensional representation of a space. In that case, these data could be integrated with video to allow for analysis and viewing of what participants were looking at.

Acknowledgments

We wish to thank John Hancock and Clarissa Mak for their assistance with this research, and the anonymous reviewers for their feedback.

References

Berman, F., & Brady, H. (2005). *Workshop on cyberinfrastructure for the social and behavioral sciences*.

French, A., Greenhalgh, C., Crabtree, A., WRight, M., Brundell, P., Hampshire, A., et al. (2006). Software replay tools for time-based social science data. *In Proceedings of the Second International Conference on e-Social Science*, Manchester, UK (pp.

Gaver, W., Sellen, A., Heath, C., & Luff, P. (1993). One is not enough: multiple views in a media space. *In Proceedings of the InterCHI Conference*, New York (pp. 335-341). ACM Press.

Heath, C., & Luff, P. (2000). *Technology in Action*. Cambridge, UK: Cambridge University Press.

Kang, H., Lee, C. W., & Jung, K. (2004). Recognition-based gesture spotting in video games. *Pattern Recognition Letters, 25*(15), 1701-1714.

Knight, D., Bayoumi, S., Mills, S., Crabtree, A., Adolphs, S., Pridmore, T., et al. (2006). Beyond the text: construction and analysis of multi-modal linguistic corpora. *In Proceedings of the Second International Conference on e-Social Science*, Manchester, UK (pp.

Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A., & Kuzuoka, H. (2007, April 28 - May 3). Museum Guide Robot Based on Sociological Interaction Analysis. *In Proceedings of the CHI 2007*, San Jose, CA (pp. 1191-1194). ACM Press.

Quaranta, A., Siniscalchi, M., & Vallortigara, G. (2007). Asymmetric tail-wagging responses by dogs to different emotive stimuli. *Current Biology, 17*(6), R199-R201.

Ranjan, A., Birnholtz, J., & Balakrishnan, R. (2007). Dynamic shared visual spaces: Experimenting with automatic camera control in a remote repair task. *In Proceedings of the ACM CHI*, San Jose, CA (pp. 1177-1186).

Ranjan, A., Birnholtz, J. P., & Balakrishnan, R. (2006, November 4-8). An exploratory analysis of partner action and camera control in a video-mediated collaborative task. *In Proceedings of the ACM Conference on Computer Supported Cooperative Work*, Banff, AB (pp. 403-412).

Salarian, A., Russmann, H., Vingerhoets, F. J. G., Dehollain, C. B., Y., Burkhard, P. R., & Aminian, K. (2004). Gait assessment in Parkinson's disease: toward an ambulatory system for long-term monitoring. *IEEE Transactions on Biomedical Engineering, 51*(8), 1434-1443.

Sifakis, E., Neverov, I., & Fedkiw, R. (2005). Automatic determination of facial muscle activations from sparse motion capture marker data. *In Proceedings of the ACM SIGGRAPH*, Los Angeles, CA (pp. 417-425).