Markerless motion-capture for point-light displays

Thomas F. Shipley* and Jonathan S. Brumberg**

*Temple University

** Boston University

Contact: Thomas F. Shipley
Department of Psychology
Temple University
1701 N. 13th St.
Philadelphia, PA 19122
(215) 204-7890
tshipley@temple.edu

Abstract

A markerless motion capture technique is described for creating point-light displays. The technique has a number of advantages over other methods for creating point-light displays, it does not require specialized equipment and can be used with human and animal actions. The method uses public domain and open source software to record the x-y coordinates of the visible joints in each frame of a video taped action. The coordinate files allow individual presentation of each element for psychophysical research on biological motion. A corpus of over 90 actions collected with this method is described.

In the area of event perception Johansson's (1973) point-light display, in which human action is reduced to a few moving points of light, have generated particular interest. In such displays there may be no static form information, yet complex forms will be seen when the points move. For example, if the major joints of a human (e.g., head, shoulders, elbows, wrists, hips, knees, and ankles) are visible as discrete points, the motions of these thirteen dots are sufficient to see a human engaged in an action. Indeed, the complex kinematic patterns of motion, generated by biological organisms, provide a wealth of information about the organism. Observers can use the point-light motion to identify the gender (Kozlowski & Cutting, 1977; Mather & Murdock, 1994; Troje 2002), identity (Cutting & Kozlowski, 1977; Stevenage, Nixon & Vince, 1999) and emotion (Dittrich, Troscianko, Lea & Morgan, 1996) of the point-light actor. The motion of the dots can also provide information about the action (Bingham, 1987; Johansson, 1973; Runeson & Frykholm, 1981). For example, subjects can identify the weight lifted in a point-light display, even if the actor is pretending to lift a different weight (Runeson & Frykholm, 1981).

The wealth of the information in these displays makes them of particular interest to researchers trying to understand how we perceive complex properties of the world in general, and events in particular. The complexity of the motions also makes them very hard to model, and despite extensive research in the field of computer graphics, few compelling examples of artificial human motions have been generated. When human action is needed for a graphics application it is almost always based on a recorded human action.

The theoretical and practical interest in generating point-light displays has led to the development of a variety of methods. These methods may be classified into three general categories. Here we review the advantages and disadvantages of each category, and present a relatively simple technique that takes advantage of current digital video technology to create point-light displays of any recordable action.

The three methods for producing point-light displays that exist today are: Image based recording with lights or retroreflective material on the actor, marker based motion tracking, and computer synthesis.

Image-based recording

Perhaps the simplest way to create a point-light display is to directly record a sequence of images that is the point-light display. To do this one must create recording conditions so that nothing is visible other than points of light on the joints of an actor. Johansson (1973) first did this by filming actors dressed in dark suits with small lights attached to their joints. By using low lighting conditions, only the lights are visible in the resulting films. One obvious drawback to using lights is that a light will not necessarily be visible when the corresponding joint is visible (e.g., for a human viewed from the side, a light on the outside of the more distant elbow would not be visible). This problem can be avoided by wrapping retroreflective material around the areas to be marked (Johansson, 1973); when video taped with contrast set high and brightness low, the actors is not visible, only the reflective patches. However, even this solution has its problems, such a marker will change size and orientation as the actor or camera moves (Dekeyser, Verfaillie & Vanrie, 2002). A solution to these problems is to identify the location of each joint in a frame by hand. Mather and West (1993) generated animations, frame by frame, where the position of dots in each frame was based on photographs of biological motion. Obviously, such an animation technique is fairly labor intensive, however it allowed Mather and West to generate point-light displays of animals actions. Alternatively, it is possible to produce point-light films without special lighting conditions using post-production software filters. If dots are painted directly on an actor, one can use video processing software to remove all color, except the dots, from the display (Thomas & Jordan, 2001). One advantage of this technique is that point-light displays can be compared with full illumination conditions by reversing the filtering process and removing the dots, leaving the actor and scene visible.

The image-based methods are simple and require little specialized equipment. However, they are limited in several critical ways. First, they generally require a willing actor (someone to wear the lights or reflective material), allow only a single view point, require special environmental conditions (e.g., low illumination levels or dark backgrounds), and for some uses the images need to be filtered to remove artifacts, such as light patches that do not correspond to joints. Finally, these point-light displays are integral; there is no simple way to alter only a portion of the display. For example, removing the upper body lights leaving only the leg lights visible would require redoing the entire recording procedure (be it filming or animation of drawings). Similarly, creating a different view of a point-light action requires a new recording from a new viewing angle because the record is 2D.

Marker-based motion capture

Many of the constraints of image-based point-light displays do not apply to marker-based motion capture. Combining markers (electromagnetic or optical) with an array of detectors allows collection of precise joint locations in 3D (for a review see Dekeyser, Verfaillie & Vanrie, 2002). The collected joint locations can be used by display software to present the action without artifacts, from any desired viewpoint, and the individual motions can be used independently. This is useful for employing a variety of psychophysical methods. For example, spatially scrambled elements of a walker could serve as the noise in a point-light walker signal detection task. Because the set of coordinates needed to create point-light displays can be stored, it is easy to share the files with other researchers.

This approach does have a few drawbacks, but most are practical. Dekeyser, Verfaillie and Vanrie (2002) outlines some of the technical problems with recording using this method as well as some remedies. One problem is that the recordings tend to include many small errors so the display will seem to bounce around. Unfortunately, most remedies for this problem sacrifice the naturalness of the motion. Also, recordings made using this method contain no information concerning the occlusion of joint locations. Therefore, when rendering a display from a particular viewpoint, it is necessary to compare each dot to a model that estimates the actor’s body position to determine whether that dot would be visible or not.

Despite the appeal of this method, the cost of the equipment presents a practical barrier to its use by many researchers. Motion capture systems (e.g., Flock of birds, Optitrak, or Gypsy exoskeleton), cost in excess of $30,000 to purchase, and can require a full time technician to maintain and operate. Additionally, some of these systems also require a special environment (e.g., a room with calibrated cameras).

Computer synthesis

Computer synthesized actions free the user from the constraints imposed by the marker-based techniques. Many algorithms exist for the creation of artificial human motion (e.g., Brand & Hertzmann, 2000; Hodgins, Wooten, Brogan, & O'Brien, 1995; Li, Wang & Shum, 2002; Liu & Popovic´, 2002). Many of the popular algorithms for the generation of point-light walkers are based on Cutting's (1978) early work on synthesis of human walking. That algorithm was based on elliptical motion of the hip and shoulder, and treating the arms and legs as nested pendulums. Being able to synthesize actions frees the researcher from constraints imposed by physical limitations of human actors. However, most of the models for point-light walkers are simple approximations of the kinematics of human motion. While computer graphics researchers have incorporated dynamic properties and physical models into their motion algorithms, they are still limited; no algorithm accurately simulates natural human motion (Runeson, 1994).

In sum, each technique for generating point-light displays has its own set of advantages and disadvantages. The selection of which technique to use involves a tradeoffs of cost, precision, display quality, and control. Image-based methods are easy and cheap, but lack control. Greater control is available with the marker-based capture method, which has high initial costs, and computer-synthesis, which cannot yet accurately model complex natural motion.

Markerless motion capture

The alternative approach described here was developed as an inexpensive alternative to marker based motion capture that retains many of the important advantages. The use of markers is replaced by a human operator who identifies key locations in each frame of a video sequence. While automation has been lost, this technique has an important advantage over marker based motion capture; the actor does not need to be human or even particularly cooperative.

This approach to the acquisition of point-light displays uses video recordings as a basis for the point-light displays, and creates digital data files containing the 2D coordinates of every visible point in the display (similar files were used by Ahlström, Blake, and Ahlström, 1997). The coordinate data file can then be used by another program to manipulate and display the point-light sequence.

The basic procedure, in brief, is to record an action with a digital video camera, load the digital file into a computer, use any consumer grade video editing program (the versions that come with current operating systems should suffice, e.g., iMovie for Mac and Moviemaker for Windows) to step through the movie one frame at a time, and for each frame record the X-Y coordinates of each joint by clicking the mouse at each of the visible joints. The recording requires only a simple program (source code for a coordinate recording program that runs on a Macintosh is available at http://astro.temple.edu/~tshipley/mocap.html) running in the background, recording the X-Y coordinates of each mouse click. Conceptually this is similar to the way Mather and West (1993) generated their point-light animations; however, the use of a coordinate file allows manipulation of the elements, and sharing the data with others.

The program that records the screen coordinates of the mouse cursor whenever the mouse is clicked needs only a few features. In addition to recording the X-Y coordinates of each click and saving them to a file, it needs to know how many points to record per frame, and how to handle missing points. When the program is started, the user specifies the number of points per frame. Indicating an occluded point is achieved by clicking in a distinct unused area of the screen (e.g., the bottom right), this communicates that a particular point is missing without requiring special input, such as pressing a “missing” key, and thus avoids conflicts with the video display software over keyboard input. The program then records a unique value (for example, coordinates of 0,0) to indicate the point was not visible.

Auditory signals can be used by the background software to provide feedback to the operator without modifying the contents of the screen. For example, a single beep could indicate each time the mouse is clicked. A double beep could indicate that the operator has clicked in the area designated for occluded points. Finally, when the operator clicks the last point in a given frame, the program could beep to indicate it anticipates beginning a new frame.

The use of background software, which does not allow much user input requires some consideration of how to handle operator errors. A simple solution is to have a way to indicate that a frame should be ignored. One way would be to have the operator enter all the points as missing to indicate that the last frame had an error and was going to be redone. An editing program could then filter the raw (with errors) output of the mouse click recorder.

For some research, it is desirable to separately display body parts, e.g., to independently translate arms and legs. To do this, it is critical that the joint locations are recorded in the same sequence for each frame. Consistency can be checked by drawing each frame to screen with the click order next to the visible points. If the right elbow was supposed to be the fifth click, there should always be a five at its location.

The obvious advantages of this method include its cost-effectiveness; a unique point-light display can be created with a digital video camera, a computer, free video editing software, and two small open source programs (the recorder and editor). This method accurately records human movement and does not vary the size or intensity of the point. Most importantly, this method does not require actors to wear any kind of marker to indicate their joints. Although this technique can be used with actors wearing some sort of marker, the error in recording points without markers is relatively small. When a 60 frame markerless video sequences of walking was redone, the mean difference in point location was 3.5 pixels (the display resolution was 1024 by 768 pixels and the walker was 308 pixels tall). Markerless motion capture frees the user from the restrictions of special environments with calibrated sensors. Furthermore, point-light display can be created from any mobile object with visually unique points. The raw material for biological motion displays is available at the nearest zoo, gymnasium, or even in home movies of a toddler’s first steps.

The disadvantages of this method are familiar. The data is 2D, so creating a new view of an action would require re-filming and re-processing that action. The precision of joint localization in markerless motion capture is lower than that of commercial marker based motion capture systems. This method also requires a human operator, so it can be labor intensive, however with modest training the short video sequences (2-3 seconds) used in much biological motion research can be processed in 20 to 30 minutes. While the loss of 3D information means that some analytic tools, such as those developed by Troje (2002), for analyzing biological motion can not be used, the freedom from the constraints imposed by markers should make this a useful technique for many point-light researchers.

Markerless motion capture corpus

Using the method outlined here, we have created a corpus of over ninety point-light displays. The actions include: Orangutans brachiating, dogs walking, seals lumbering on land, bats and owls flying, humans crawling, walking, and running at different speeds, various gymnastics (including cartwheels, jumping jacks, swinging on monkey-bars, and handsprings), throwing, kicking and hitting various balls, and a variety of karate kicks executed by actors with different levels of training.

The point-light actions are generally readily recognizable. For research that requires displays of established quality, we have collected basic data from 16 subjects on the appearance of 14 of these actions. The phenomenal report was coded for accuracy (most displays were recognized by at least 90% of the subjects). We also collected ratings of how well each of the 14 point-light displays matched a single verb, or phrase (on average, all displays were rated 7.35 or higher on a ten point scale), but there should be sufficient variation across displays for regression analyses. The X-Y coordinate files and a java applet for viewing them, and the normative data are all available at http://astro.temple.edu/~tshipley/mocap.html.

Conclusion

Numerous techniques for capturing biological motion exist. Commercially available methods for 3D motion capture are expensive and can be limited by the need for the actor performing the actions to be cooperative. This restriction prevents creating point-light displays with some classes of actors (i.e. animals and infants), and the equipment can place the actor in a situation where they may not act in the same was as in their natural environment. The ability to record without markers presents the opportunity to capture the movement of humans and animals in their natural environments. With the markerless motion capture described here there is a virtually limitless supply of actors waiting to be transformed into point-light displays.

References

Ahlström, V., Blake, R., & Ahlström, U. (1997). Perception of biological motion. *Perception*, **26**, 1539-1548.

Bingham, G. P. (1987). Scaling and kinematic form: Further Investigations on the visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception & Performance*, **13**, 155-177.

Brand, M., & Hertzmann. (2000). Style machines. *Proceedings of ACM SIGGRAPH 2000, Annual Conference Series*, 183-192.

Cutting, J. E. (1978). A program to generate synthetic walkers as dynamic point-light displays. *Behavior Research Methods & Instrumentation*, **10**, 91-94.

Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, **9**, 353-356.

Dekeyser, M., Verfaillie, K., & Vanrie J. (2002). Creating stimuli for the study of biological-motion perception. *Behavior Research Methods, Instruments, & Computers*, **34**, 375-382.

Dittrich, W. H., Troscianko, T., Lea, S. E. G., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception*, **25**, 727-738.

Hodgins, J. K., Wooten, W. L., Brogan, D. C., & O'Brien, J. F. (1995). Animating human athletics. *Proceedings of ACM SIGGRAPH 95, Annual Conference Series*, 71-78.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, **14**, 201-211.

Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, **21**, 575-580.

Li, Y., Wang, T., & Shum, H. (2002). Motion texture: a two-level statistical model for character motion synthesis. *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 465-472.

Liu, C. K., & Popovi´c, Z. (2002). Synthesis of complex dynamic character motion from simple animations. *Proceedings of ACM SIGGRAPH 2002, Annual Conference Series,* 408-416.

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London: Series B*, **258**, 273-279.

Mather, G., & West, S. (1993). Recognition of animal locomotion from dynamic point-light displays. *Perception*, **22**, 759-766.

Runeson, S. (1994). Perception of biological motion: The KSD-principle and the implications of a distal versus proximal approach. In G. Jansson, S. S.

Bergström, & W. Epstein (Eds.), *Perceiving events and objects* (pp. 383-405). Hillsdale, NJ: Erlbaum.

Runeson, S., & Frykholm, G. (1981). Visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 733-740.

Stevenage, S. V., Nixon, M. S., & Vince, K. (1999). Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, **13**, 513-526.

Thomas, S. M., & Jordan, T. R. (2001). Techniques for the productions of point-light and fully illuminated video displays from identical recordings. *Behavior Research Methods, Instruments, & Computers*, **33**, 59-64.

Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, **2**, 371-387.