Optical Tracking for Music and Dance Performance

J. A. Paradiso and F. Sparacino
Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
USA

Abstract

This paper describes three different types of real-time optical tracking systems developed at the MIT Media Laboratory for use as expressive human-computer interfaces in music, dance, and interactive multimedia performances. Two of these, a multimodal conducing baton and a scanning laser rangefinder, are essentially hardware-based, while the third is a computer vision system that can identify and track different segments of the performer's body. We discuss the technical concepts behind these devices and outline their applications in music and dance environments.

1. Introduction

Computers are playing an increasingly important role in musical performance. Although modern digital synthesizers and synthesis algorithms give composers access to an essentially unlimited palette of timbres and dynamics, performers using standard synthesizer keyboard interfaces are severely limited in evoking such levels of expression. There are two approaches to solving this problem. One is to develop new kinds of dexterous musical interfaces that measure several different modes of performer control, thereby opening additional expressive channels. The other is to embed some degree of intelligence in the interface itself, allowing simple musical gestures to trigger and shape complex musical phrases at a higher level of abstraction [1]. Both of these approaches converge in real-time systems for interactive music and dance, where many aspects of a performer's kinematical state are measured and translated into a complex musical and graphical response.

Several different sensing technologies have been exploited in designing interfaces for computer music [2]. Optical tracking is well-suited to many of these applications, and the decreasing cost of the required components, coupled with the ever increasing capability of computers and faster image processing algorithms often make it a practical choice. This paper describes three different interfaces that we have developed and applied in musical applications, all exploiting real-time optical tracking at different levels of sophistication: a simple, fast point tracker embodied as a digital conducting baton, a scanning laser rangefinder to measure limb position in an active plane, and a 3D vision system to track detailed body motion.

Presented at the Fourth Conference on Optical 3D Measurement Techniques, ETH, Zurich, September 1997

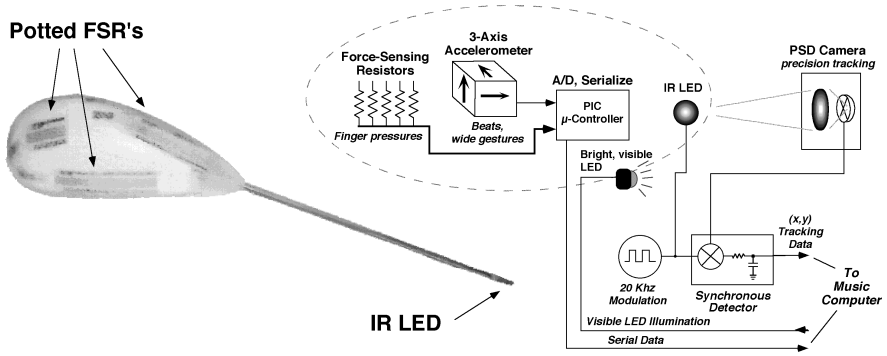

Figure 1: The Digital Baton; Photograph (left) and Schematic (right)

2. The Digital Baton

The conducting baton holds a vaunted spot in the music world, where for centuries a simple stick, appropriately maneuvered, has been able to shape and direct the sound of an entire orchestra. Since its role is such a natural fit to computer performance, as outlined above, several different types of baton interfaces have been constructed, many using optical tracking. Most of these are based on a CCD camera looking at a simple IR light source [3,4], and others use a segmented photodiode detector [5].

Our baton system [6], shown in Fig. 1, is a multimodal, handheld input device that measures three types of user activity. The position of a wide viewing-angle infrared LED at the baton's tip is precisely tracked, allowing it to be used in a conventional "conducting" role. Two additional sensing channels, however, expand its application as an expressive performance instrument well beyond that of a traditional baton. An array of five force-sensitive resistor strips [7], mounted along the performer's grip, measure continuous pressure of the thumb, index finger, middle finger, combined last two fingers, and palm. A set of three orthogonal micromechanical accelerometers measure directional velocity changes and beats, plus, on slower timescales, provide an indication of the baton's orientation. Figure 2 shows data taken from these baton sensor subsystems; i.e., tracker coordinates from sweeping a circle and a cross, accelerometer response to beats, and finger presses.

In order to provide reliable, fast response in theater performances with a large, unpredictable background from incandescent stage lighting, we chose not to employ video tracking, but instead built a synchronously-demodulating tracker based around a 2D position-sensitive photodiode (PSD). This PSD camera is positioned several meters away from the performer (generally out in the audience), in order to view the entire performance volume. The baton LED is modulated at 20 kHz, and synchronously detected in the camera electronics. An IR filter placed over the PSD camera, together with a narrow demodulation filter bandwidth, completely attenuates disturbance from static and dynamic stage lighting, while retaining an adequately prompt response (better than 20 msec) to fast baton dynamics. One PSD camera provides relatively precise ($\sigma \approx 1$ cm w. camera at 8 meters) 2D tracking of the baton tip; these coordinates are directly provided by the analog baton electronics, without the need for digital image

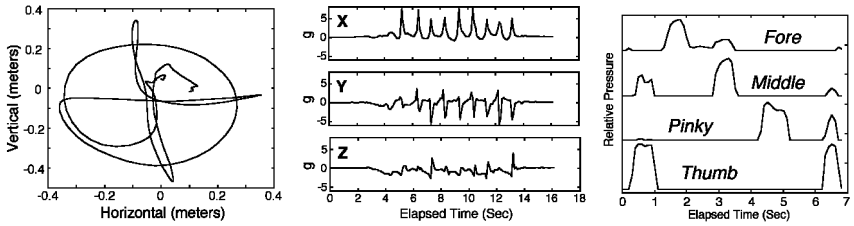

Figure 2: Tracker (left), accelerometer (middle) and finger pressure (right) data

processing. The 3D tip position can be inferred to some extent from the detected LED intensity and the accelerometer data, or precisely determined by a stereo pair of PSD cameras. The "Visible LED" of Fig. 1 illuminates the baton via computer control; it is for performance aesthetics only, and has nothing to do with the tracking system.

This baton has been used both in research projects and actual stage performances, the latter for conducting synthesizer and computer-driven music for the Brain Opera [8], a large musical installation that has appeared at several locations around the world. The baton is a rich source of gestural data that is applied to control several musical parameters in different ways. Often, the precise tracking and finger pressure data are used to select a particular sound and set of sound parameters, which are then varied as the baton tip is moved. Accelerometer data is used to vary current sounds or fire new sounds upon detected beats.

3. Scanning Laser Rangefinders as Gestural Input Devices

The current baton system uses a thin, multiwire cable to pass power, data, and LED modulation. Although the cable can be dispensed via a simple wireless link, the baton is still, by nature, a handheld interface. While this works well under the assumption of a conductor or musical instrument player, we have often encountered situations where it is less practical to assume that performers can hold or carry particular hardware devices (as in dance applications or interactive installations open to the public). These situations call for untethered sensor systems that respond appropriately to hand and body gesture, without requiring the player to be carrying special electronics. We have built several such systems using capacitive or electric field sensing techniques [9], extending technologies that date back to the Theremin. While these sensors respond readily to gestural input, they are also sensitive to body mass and general body position; the capacitive hand trackers that we have designed require fairly strict postural constraints and calibration for each performer in order to obtain results with any degree of accuracy and repeatability.

We have thus developed very simple and inexpensive scanning laser rangefinder systems for use as a precise gestural interface in musical performance. Our current device works by triangulation, scanning a 3 mW diode laser beam across a 90° sensitive region at up to 30 Hz and synchronously detecting (at 250 kHz) the range offset in a 1D PSD camera. An AGC compensates the received gain for variations in reflection strength, and the demodulation filter allows finger-sized objects to be detected beyond a meter, while adequately suppressing noise and background light, yielding useful ranging data at up to a 2 meter radius. A 68HC11 microprocessor digitizes the intensity

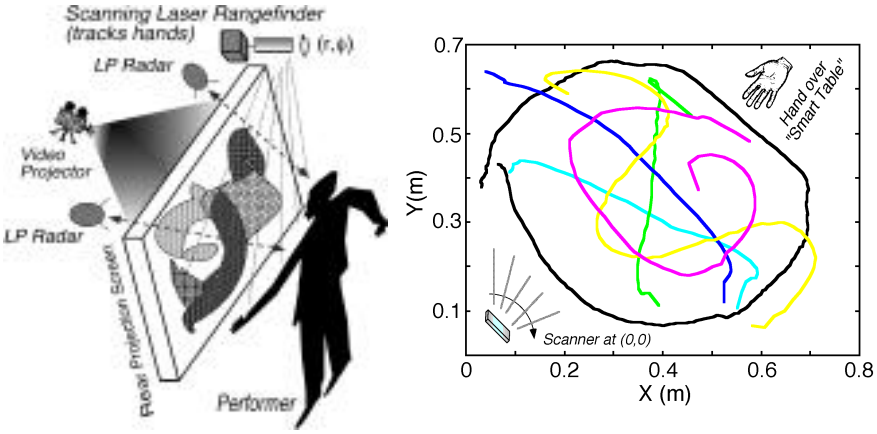

Figure 3: "Smart Wall" and hand trajectory measured by triangulation rangefinder

and range data over every scan, locates peaks in the intensity data corresponding to different objects, and outputs the angle and range information for each one over a serial connection. In order to increase the sensitive range and resolution to enable room-scale coverage, we are now building a rangefinder based on quadrature phase detection. As the performance requirements for our applications are relatively modest compared to the state of the art [10], we expect this device to be similarly inexpensive.

Scanning laser rangers are most often used for applications such as 3D surface measurement [11] and robot collision avoidance [12]. This system was built for a different purpose; a low-cost and robust means of capturing real-time gestural data for expressive human-computer interfaces. The signals reflected from bare hands in the scanned beam are extremely clear, allowing hand position to be precisely determined in the scan plane. Mounting this scanner at the corner of a projection screen enables hand motion to be tracked just above the display surface, as depicted at left in Fig. 3 (the "low power radars" [13] see through the screen to sense a person approaching). The rangefinder allows a user to interact directly with the graphics and display by moving their hands above the screen; no contact is required, and precise, repeatable coordinates are produced. Mounting this device atop a floor tracks foot position within the scanned area, useful for determining performer locations for dance and public installations. We are now exploring these systems for several applications that control musical and graphical output for interactive performance. Fig. 3 shows sample hand trajectories above such a "smart surface", plotting data from our current triangulation rangefinder.

4. An Interactive "DanceSpace" based on Computer Vision

While generally somewhat slower in response and more sensitive to lighting changes and clutter, computer vision techniques can be used to track many more details of human body position and motion. The third system described is a real-time computer vision "person finder", abbreviated as "Pfinder" [14], a system for tracking the body and interpreting the movement of a single performer. It uses only one wide-angle camera pointed towards the stage and a standard Silicon Graphics Indy computer. The

Figure 4: Pfinder-driven "blob" avatar responding to body state

system employs a multi-class statistical model of color and shape to segment a person from a background scene, then find and track people's body parts in a wide range of viewing conditions. It adopts a "Maximum *A Posteriori*" Probability (MAP) approach to body detection and tracking, using simple 2.5 dimensional models. It incorporates *a priori* knowledge about people, primarily to bootstrap itself and to recover from errors. We have used Pfinder to create an interactive performance space, called DanceSpace, where both professional and non-professional dancers can generate music and graphics through their body movements.

Pfinder builds the scene model by first observing the scene without people in it. When a human enters, a large change is detected in the scene, which cues Pfinder to begin constructing a model of that person, built up over time as a dynamic "multi-blob" structure. The model-building process is driven by the distribution of color on the person's body, with blobs being added to account for each differently colored region. Separate blobs are generated for the person's hands, head, feet, shirt, and pants. The process of building a blob-model is guided by a 2D contour shape analysis that recognizes silhouettes in which body parts can be reliably labeled.

The computer vision system is composed of several layers. The lowest layer uses adaptive models to segment the user from the background, enabling the system to track users without the need for chromakey backgrounds or special garments, while identifying color segments within the user's silhouette. This allows the system to track important features (hands) even when these features aren't discernible from the figure-background segmentation. This added information makes it possible to deduce the general 3D structure of the user, producing better gesture tracking at the next layer, which uses the information from segmentation and blob classification to identify interesting features: bounding box, head, hands, feet, and centroid. These features can be recognized by their characteristic impact on the silhouette (high edge curvature, occlusion) and (*a priori*) knowledge about people (heads are usually on top). The highest layer then uses these features, combined with knowledge of the human body, to detect significant gestures and movements. If Pfinder is given a camera model, it also

back-projects the 2D image information to produce 3D position estimates using the assumption that a planar user is standing perpendicular to a planar floor.

Pfinder provides a modular interface to client applications. Several clients can be serviced in parallel, and clients can attach and detach without affecting the vision routines. Pfinder is a descendant of the vision routines originally developed for the ALIVE system [15], which performed person tracking but had no explicit model of the person and required a controlled background. Pfinder is a more general, and more accurate method for segmentation, tracking and interpretation. Fig. 4 shows the image of a user as seen by the camera, together with the Pfinder blob description (note the tolerance to background clutter).

DanceSpace is an interactive stage that takes full advantage of Pfinder's ability to track the dancer's motion in real time. Different parts of the dancer‘s body (hands/head/feet/torso) can be mapped to different musical instruments that constitute a virtual keyboard. Moreover, the computer can recognize hand and body gestures, which can trigger rhythmic or melodic changes in the music. Additionally, a graphical output is generated from the computer vision estimates.

The computer-generated music consists of a richly-textured melodic base tune, which plays in the background for the duration of the performance. As the dancer enters the space, a number of virtual musical instruments are invisibly attached to her body. The dancer then uses her body movements to “magically” generate an improvisational theme above the background track.

In the current version of DanceSpace, the dancer has a violin in the right hand, a flute on their left hand, and bells and drums attached to her feet. The dancer's head works as the volume knob, bringing down the sound as they move closer to the ground. The height of the dancer's hands is mapped to the pitch of the notes played by the musical instruments attached to them. The musical instruments on both hands are played in a continuous mode (i.e., to get from a lower to a higher note the performer will have to play all the intermediate notes). The bells and drums, on the contrary, are "one shot" musical instruments. More specific gestures of both hands or combinations of hands and feet can generate melodic or rhythmic changes in the ambient melody. The dancer can therefore "tune" the music to her own taste throughout the performance.

As the dancer moves, their body leaves a multicolored trail across the large wall screen that comprises one side of the performance space. The color of the trail can be selectively mapped to the position of the dancer on stage or to more "expressive" motion cues like speed. This multicolored trail is intended to represent a virtual partner that the dancer generates though their performance or a "shadow" that follows the dancer around on stage. If the shadow has a long memory of trails, the dancer can paint more complex abstract figures onto screen. Fig. 5 shows the DanceSpace visuals, from the perspective of the performer on an active stage (left) and the overhead camera used by the vision system (right).

The choreography of the piece can then vary according to which elements of the interactive space the choreographer decides to emphasize. In one case, the dancer might concentrate on generating the desired musical effect; in another moment of the performance, the dancer may want to concentrate on the graphics (i.e., painting with the body). Finally, the dancer might just focus onto dance itself and let DanceSpace generate the accompanying graphics and music.

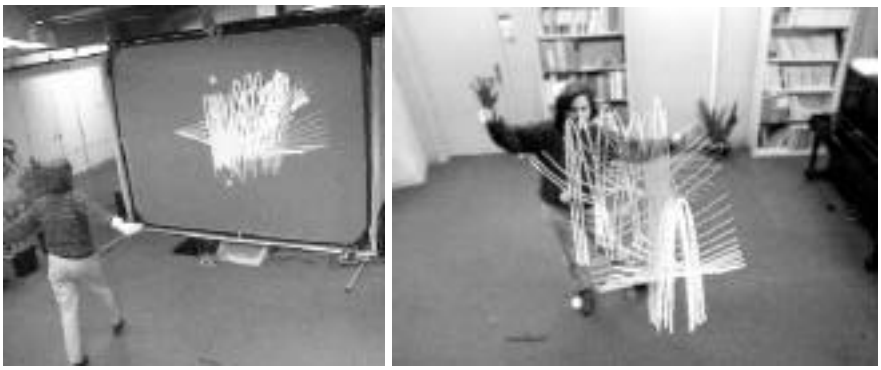

Figure 5: DanceSpace in action; stage (left) and overhead camera (right) views

The philosophy underlying DanceSpace is inspired by Merce Cunningham's approach to dance and choreography [16]. Cunningham believed that dance and movement should be designed independently of music, which is subordinate to the dancing and may be composed later for performance, much as a musical score is in film.

Previous interactive dance systems [17,18] were based on using different sensing schemes to estimate the gross motion of the performer, or required the dancer to wear special clothing, sometimes outfitted with wired sensors or targets. DanceSpace avoids these constraints, achieving much finer real-time kinematic detail by exploiting the Pfinder system. Consequently, the current DanceSpace environment reliably tracks only solo performers, and imposes restrictions on the dynamics of stage lighting.

We envision DanceSpace as an installation for indoor public spaces, where people usually spend long hours waiting (i.e., airports, lobbies), interactive museums and galleries, or in a performance space, allowing a dancer to play with their own virtual shadow and generate customized music for every new performance.

A large number of users and performers of all ages have tried DanceSpace in several demonstrations at the MIT Media Lab, and we have worked with various choreographers [19] to produce DanceSpace pieces for interactive stage settings. We are now expanding DanceSpace to allow for a greater variety of musical mappings and different graphical representations of the dancers.

5. Conclusions

Optical tracking will play an important role in sensing user gesture for interactive environments. In dedicated applications, where a limited amount of precise tracking information is needed with high reliability under variable lighting and clutter background, the simple, hardware-based trackers (e.g., digital baton, laser rangefinder) performed very well. When more detailed information is desired with somewhat relaxed performance requirements, machine vision schemes, such as Pfinder and DanceSpace, are now able to approach the requirements needed for real-time performance. As these continue to improve with increasing computing and algorithm capability, they will play a prime role in making the interactive entertainment spaces of tomorrow into highly responsive environments.

6. Acknowledgments

The authors acknowledge their many colleagues and collaborators at the MIT Media Lab who contributed to these projects. For the Digital Baton, we thank Ed Hammond, Tod Machover, Theresa Marrin, Maggie Orth and Chris Verplaetse. Josh Strickon is thanked for his contributions to the laser ranger. We acknowledge Chris Wren for Pfinder design, our DanceSpace collaborators Akira Kotani and Chloe Chao for music and graphics software, and thank our Media Lab colleagues Glorianna Davenport and Sandy Pentland for many stimulating discussions.

7. References

[1] Machover, T., 1991. Hyperinstruments: A Composer's Approach to the Evolution of Intelligent Musical Instruments. In *Cyberarts*, William Freeman, San Francisco, pp. 67-76.

[2] Roads, C., 1996. *The Computer Music Tutorial*. MIT Press, Cambridge, MA.

[3] Morita, H., *et. al.* 1991. A Computer Music System that Follows a Human Conductor. Computer, Vol. 24(7), pp. 44-53.

[4] Bertini, G., Carosi, P. 1993. Light Baton System: A System for Conducting Computer Music Performance. Interface, Vol. 22(3), pp. 243-257.

[5] Rich, R. 1991. Buchla Lightning MIDI Controller. Electronic Musician, 7(10), pp. 102-108.

[6] Marrin, T., Paradiso, J. 1997. The Digital Baton: a Versatile Performance Instrument. To appear in Proc. of the 1997 Intl. Computer Music Conf.

[7] See: *http://www.interlinkelec.com/.*

[8] Paradiso, J. New Instruments and Gestural Sensors for Musical Interaction and Performance. To be published.

[9] Paradiso, J., Gershenfeld, N. 1997. Musical Applications of Electric Field Sensing. Computer Music Journal, Vol. 21, No. 3.

[10] Rueger, J.M. 1990. *Electronic Distance Measurement*. Springer-Verlag, Berlin.

[11] Petrov, M., *et. al.* Optical 3D Digitizers: Bringing Life to the Virtual World. Submitted to IEEE Computer Graphics and Applications, 1997.

[12] Everett, H.R. 1995. *Sensors for Mobile Robots: Theory and Application*. A.K. Peters, Wellesley, MA.

[13] Paradiso, J. *et. al.* 1997. The Magic Carpet: Physical Sensing for Immersive Environments. Proc. of the CHI '97 Conf. on Human Factors in Computing Systems, Extended Abstracts. ACM Press, NY, pp. 277-278.

[14] Wren, C., *et. al.* 1997. Pfinder: Real-Time Tracking of the Human Body. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19(7), pp. 780-785.

[15] Maes, P. *et. al.* 1997. The ALIVE System: Full-body interaction with animated autonomous agents. ACM Multimedia Systems, Vol. 5, pp. 105-112.

[16] Klosty, J. 1975. *Merce Cunningham: Dancing in Space and Time*. Saturday Review Press, NY.

[17] Gehlhaar, R. 1991. SOUND = SPACE: an Interactive Musical Environment. Contemporary Music Review, Vol. 6(1), pp. 59-72.

[18] Marion, A., 1982. Images of Human Motion: Changing Representation of Human Identity. MS Thesis, Massachusetts Institute of Technology, Cambridge MA.

[19] Erica Drew (Boston Conservatory) and Claire Mallardi (Radcliffe College).