reprinted from *Proceedings of the 1997 IEEE International Conference on Systems, Man and Cybernetics, October 1997*

Foot-Mounted Gesture Detection and its Application in Virtual Environments

Insook Choi* and Carlos Ricci†
Beckman Institute, 405 N. Mathews, Urbana, IL 61801, USA
email: ichoi@ncsa.uiuc.edu

ABSTRACT

As a computer technology develops, the high-end computing environment such as virtual reality no longer limits its applications to data display. We focus on simulations and real-time explorations of the simulated models, that lead to the situation that we no longer can rely on simple mechanical devices such as the keyboard, mouse, and joystick that are commonly used in Human-Computer Interaction. The foot-mounted input gesture detection is a spin-off project from the demand we have encountered in working on creative projects in a VR environment. The objective was to develop an interface that accounts for the one of the most basic human movements, a natural stance and bipedal locomotion.

Unlike previous walking interfaces such as the sensor tiles, treadmill, and stepper, our device is not limited to a fixed position since it is wearable in free motion. Further, the multiplicity of pressure signals from the foot provides a high-dimensional control source inherent to the design while the modularity of the signals provides a means for differentiating human-determined motion patterns. Pattern recognition was implemented using rule-based inferences based on fuzzy logic.

1. INTRODUCTION

This paper presents a wearable interface device in a virtual environment application. As of today the most commonly used interface devices in virtual environment are pointing devices for tasks such as menu choice, directional navigation, and zooming in and out of displayed objects. Our project was motivated from the fact that gestures other than pointing are prevalent. For reference we also want to state that the gesture of finger pointing is culture-specific and can be considered unacceptable in some cultures. We wish to bring alternative interfaces to mobilize natural components of human body movement in a virtual environment. Piaget has emphasized the importance of movement when learning takes place [6]. Among several human movements we have explored, our first project was to prototype a device that interfaces natural walking motion to a computing environment.

Walking is such a basic locomotion we engage in daily life yet it was not immediately clear how we would apply such motion for exploring and learning in virtual space. The current project studies free motion, unconstrained stance and bipedal balance of a performer, as measured through the forces applied by the foot. The variety and flexibility of mobilized human presence in a computer interface leads to an inquiry of new display and feedback techniques that have been previously overlooked in systems adhering to limited kinesthetic assumptions concerning observers.

2. DESIGN CRITERIA: SYSTEM HARDWARE AND ERGONOMICS

The hardware and software were to support the following criteria: 1) the device should be wearable with minimum obtrusiveness, 2) the device should incorporate multiple-gesture sensitivity by mounting optimal number of sensors for each foot, and the signal flows among them should be continuous, 3) generalization at the software level should support symbolic interpretation of the continuous signals. The design and construction of the system hardware was guided by certain physical constraints, perhaps the most important being those related to the foot forces themselves. The nature of the force interaction between the foot and sensor system was seen to determine the effective choice and placement of the force sensors, and to ultimately determine the “feel” of the system.

2.1 Wearability and construction

In a virtual reality system such as the CAVE [3] the positional data is obtained by a head-tracking mechanism by which the point of view is constantly updated wherever an observer stands. Thus it is desirable to allow a free motion as the observer walks around the space, which suggests the physical mounting of sensors and electronics to the observer. Our general design objective was that the foot sensor system would be easily mounted by the user, and once in use, would be as unobtrusive as possible. The benchmark for this objective would be the ability for a performer to don the hardware as part of an actual performance without significantly altering the course of the performance. These constraints led to the design of the sensor system as integrated pieces, or “inserts”, which encapsulate the force sensors and are fitted beneath the soles of each of the user’s shoes.

The inserts are constructed as a laminate, cut to fit the nominal shape of the sole of the user’s shoe. Varying shoe sizes could be accommodated simply by cutting other substrates within the approximate range of the respective size. The force sensors themselves attach to a substrate of hard vinyl, and a layer of soft vinyl covers the sensors and sensor wires. The wires are drawn to the center region of the insert beneath the arch of the foot, where the foot pressure is typically the least, and a cable is terminated at that point. The inserts were initially attached to the shoes with straps, but in practice this proved too cumbersome. In the current configuration the inserts are placed onto the inside sole of “booties”, of the type used in clean rooms. The booties, which are easily slipped over the shoes and then snapped tight to the legs, not only provide ergonomic

* Human-Computer Intelligent Interaction, Beckman Institute, and NCSA
† National Center for Supercomputing Applications

convenience, but also serve to protect the inserts and provide a means to neatly guide the cables upward to a small interface box which is worn on the waist and houses the interface electronics. The adoption of the boots led to the system being called "CyberBoots".

2.2 Force-based multiple-gesture sensitivity

We draw multiple gestures from foot movements derived from bipedal locomotion. Three pattern groups of bipedal locomotion were initially identified and studied from performer's movements: natural walking forward and backward, mime walking forward and backward, and leaning on a plane. The walking patterns were comprised of repeating sequences of rest states and state transitions, the leaning patterns of rest states without transitions. Multiple sensors define these states as combinations of individual sensor signal states. By introducing multiple sensors we allow for a broader repertoire of states by which patterns may be constructed. We identified force as the only means by which movement information would be conveyed. Compared to position measurement, force is underutilized in virtual reality interfaces. At the same time, force and acceleration are more intimately tied to the user's sensation of feedback, whereas position implies a reference frame external to the user.

The forces which were chosen to be sensed were compressive, normal to the plane of the base of the foot. This was considered to provide for more direct, independent measurement of the various sources of pressure along the bottom of the foot, more so than may be inferred from measurements of other types of forces such as shear, bending, or twisting forces. To simplify the electronic hardware, the total number of force sensors in the system was limited to eight, distributed four per foot. Four key pressure points on the base of the foot were identified for the sensor placement: the heel, the inner and outer ball, and the toe tip. These points are considered consistent with the four dominant peaks of distribution of force along the base of the foot and so may be considered in this case to convey the greatest amount of information.

The force sensors themselves were chosen to be implemented using simple devices called Force Sensing Resistors[*] (FSRs). FSRs were chosen because their size and shape allow for multiple, planar sensor mountings per foot. They allow for a relatively simple electronic interface which provides repeatable, linear force responses with a dynamic range which is reasonably suited to the nominal expected range of foot forces. Other benefits of the FSRs are their reliability, commercial availability and relatively low cost. While FSRs are not accurate in an absolute sense, this is not problematic to the current system: the gesture inference processing (see section 4) only requires that the measurements be consistent in a relative sense.

3. SIGNAL FLOW AND PROCESSING

The flow of signals in the foot-mounted gesture recognition system is given as a block diagram in Figure 1. The foot sensor assembly appears to the left of the figure. Four force sensors per foot, represented in the figure by small discs, are mounted to the assembly as shown. By way of a cable harness, the sensors connect to analog interface circuitry where the sensor signals are conditioned and then digitized by a small microcontroller. The analog circuitry and microcontroller comprise a small module which is worn on the waist. The microcontroller translates the data into packets and sends them across a standard serial interface connection to the main graphics computer.

At this point the eight pressure signals are normalized to fall within the range [0,1], where the lower bound corresponds to no pressure (i.e., toe and/or heel completely off of the floor) and the upper bound to pressing reasonably hard on the floor (i.e. standing tip-toe). The mid-value 0.5 is mapped to correspond roughly to standing at rest with the feet flat. For the initial experiments, a fixed normalization was used to accommodate the absolute weight of a single user.

For the investigation of inferring simple walking and leaning gestures, we were only interested in patterns arising from the differentiation of the heel and toe. Thus, the signals from the left and right ball of the foot were combined with that of the toe-tip to generate a composite "toe" signal. Combining the three signals by taking either the maximum or the weighted average produced similar results.

We call these normalized heel and toe signals $H_{l,r}$, $T_{l,r}$ where the subscripts l,r correspond to the left and right feet, respectively. Let us now consider the fuzzy set *P* into which full membership requires a heel or toe being "fully pressed". Thus, we may view the values of *H* and *T* to correspond with partial membership in *P*. In the subsequent rule logic, these signals will be seen to form the *static* or *gating* conditions.

Also important to the gesture inferencing process are the transitions from one static condition to another. So, the time derivatives of *H* and *T* are estimated using a bandlimited, first-order finite-difference approximation to the continuous time derivative. A signal diagram representing this process appears in Figure 2. For the arbitrary raw gating signal input x_i, a bandlimited signal x is produced along with its partial-membership complement $\overline{x}$, in addition to the linear time derivative estimate $\dot{x}$. The derivative signal passes through a comparison block to produce the outputs dx and $\overline{dx}$ which are "gated" to be positive-going according to

$$dx = \begin{cases} \dot{x} & \dot{x} \geq 0 \\ 0 & \dot{x} < 0 \end{cases}$$

and

$$\overline{dx} = \begin{cases} 0 & \dot{x} \geq 0 \\ -\dot{x} & \dot{x} < 0 \end{cases} \tag{1}$$

[*] Interlink FSR#402, 0.5 inch diameter discs

Figure 1. Signal flow of the foot-mounted gesture detection.

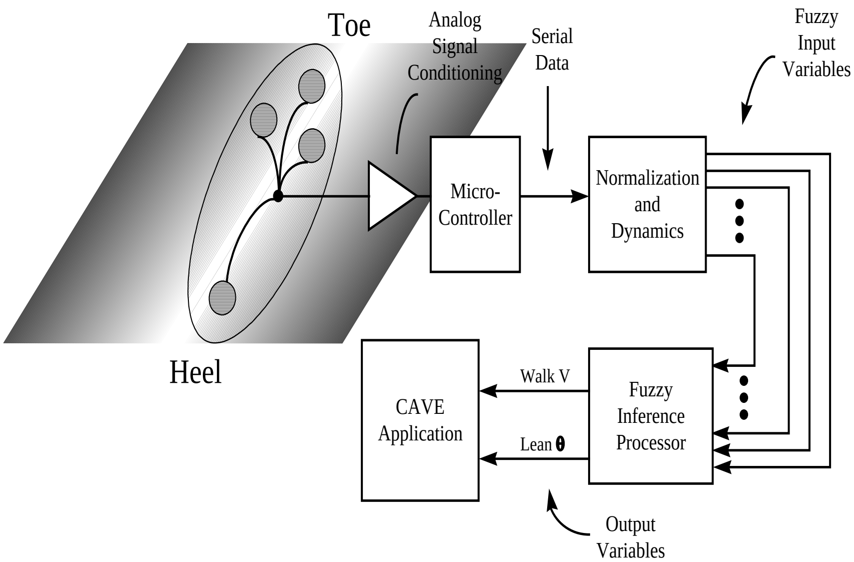
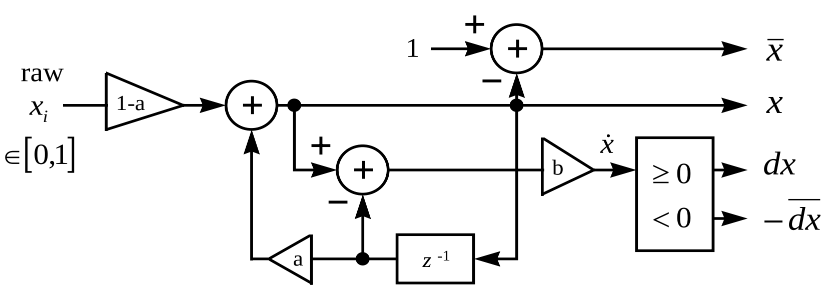


Figure 2. Generation of fuzzy inputs.



Let us consider the fuzzy sets *I* and *D* into which full membership requires that x or $\overline{x}$ be "increasing at a full rate", respectively. Then, given an appropriate scaling of parameter "b", we may say that $dx = 1$ implies full membership into *I* and correspondingly, $\overline{dx} = 1$ implies full membership into *D*. These values will be seen in the subsequent rule logic to form the *dynamic* or *transient* conditions. In practice, parameter "b" is adjusted for a natural "feel" with regard to the rate of pressing or releasing, typically set in the current configuration so that derivative output magnitudes of unity map to a full-scale change of x in 0.5 second or less. The bandlimiting parameter "a" was typically set in the experiments to an effective lowpass time constant of 50 msec. At run time, both "a" and "b" are adjusted dynamically to account for non-deterministic execution times in the main graphics computation loop.

The above mapping may thus be seen to form the so-called "fuzzification" of the analog pressure values, and so may be considered to play the role of the traditional "input membership functions". The collection $x, \overline{x}, dx, \overline{dx}$ therefore comprise the fuzzy input variables to the inferencing process. The collection is repeated for each heel and toe of each foot, for a total of 16 generated fuzzy inputs.

As indicated in Figure 1, these fuzzy inputs are passed on to the fuzzy inference engine. There, the gesture inferencing is executed using predefined rule sets to produce multiple "crisp" outputs which are then passed to the virtual reality application.

4. INFERENCE PROCESSING

The inferencing of both walking and leaning gestures is based on the process of executing sets of pre-defined rules in a rule base. The rule execution or "firing" occurs entirely in response to the fuzzy inputs comprising the antecedents of the rules. The consequents of these rules, also known as fuzzy outputs, are then applied as weights to corresponding output membership functions. All output membership functions associated with a particular output variable are then linearly combined, or averaged, to produce a final output value. This operation is known as "defuzzification" since through it any property of "fuzziness" in the final output values is considered to be combined and/or averaged out. The outputs are correspondingly referred to as "crisp" values and may be applied back to the "real-world" plant or system.

While many generalizations to the rule-based method of fuzzy inferencing exist [4], we hold that for the current system the rule base methodology provides a structured framework and language for development of the inferencing system design. The aspect of rule language has played a particularly important role in the current development of the rule base for walking gestures.

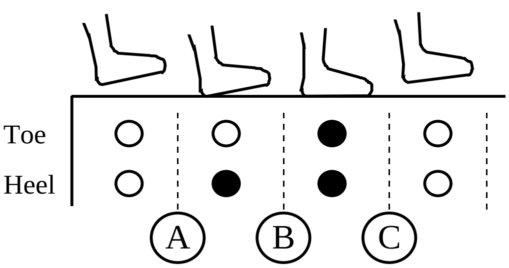
4.1. Walking Gestures

Pervasive throughout the design of the walking gesture recognition is the notion that a "walk" is in essence a time-indexed pattern or sequence of events, or states. If a means is first developed to describe these events, then a rule base is readily established as a natural extension of this event description. We will use as an example here one of the simplest sequences to study, namely, that arising from the basic, or "natural" pattern casually employed by most humans as they walk. The method employed in the current work analyzes the walk pattern from the perspective of the sensors, or more specifically, the static conditions set up through fuzzy input variables *H* and *T*. By considering the bounding (Boolean) values of these variables as states, one may break the walking pattern down into a sequence of such states. This is consistent with the traditional description of rule bases in hard Boolean terms, while the underlying AND, OR operations are actually fuzzy operations.

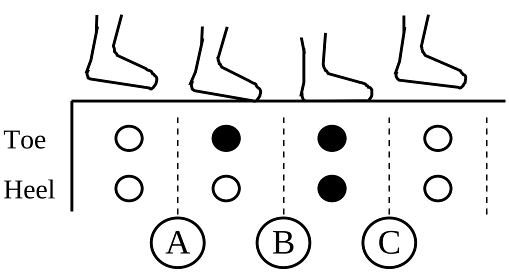
For simplicity in the example, we will look at the pattern of only one foot. Note that for walking patterns that feel “regular” or “smooth”, the pattern will typically be found to also exhibit symmetry; i.e., both feet will typically be found to exhibit the *same* pattern, except staggered from one foot to the next (see section 5). The basic walk pattern is diagrammed in Figure 3 in the form of states progressing forward in time from left to right. The forward walking pattern in Figure 3a begins with both the toe and the heel off of the floor. The associated state is defined by T=0 and H=0. At the next defined state, the heel is on the floor, but the toe is off of the floor, so that T=0 and H=1. Next, the toe comes down and T=1, H=1. Finally, both the toe and heel lift and the sequence repeats.

A fourth state, where the heel lifts but the toe is still on the floor, does exist in some walks, particularly if the pattern is stopped in mid-walk. This state was found to be very short in duration relative to the whole sequence, and was ignored here. Note that the fuzzy processing allowed this omission to take place with negligible consequences. In contrast, a recognizer based on a “hard” Boolean state machine would demand strict adherence to a pattern or otherwise would reject that state transition entirely.

Figure 3. State and transition definitions for the “Natural Walking” pattern. a) Forward. b) Backward.



a)



b)

Now, since “walking velocity” is reasonably nonzero only while state transitions are occurring, we choose to define the pattern logic at the transitions between the states. Hence, to complete the rule base we must apply to the above static definitions the dynamic conditions set forth by the fuzzy input variables dH and dT. Referring again to Figure 3a we see that the state transitions are denoted by the circled letters A, B, and C. Let us consider the state transition A. We see that the toe remains in the air so that T=0 throughout the transition. However, the heel makes contact with the floor, so that we may define the dynamic bounding condition dH=1 for the transition. Thus, the transition is fully defined by T=0 AND dH=1. Similar combinations of static and dynamic conditions may be set up for the remaining transitions, so that we may describe a corresponding set of rules according to

$$\begin{aligned}
&\text{A:} \quad && \overline{T}\cdot dH \Rightarrow B_F \\
&\text{B:} \quad && dT\cdot H \Rightarrow B_F \\
&\text{C:} \quad && \overline{dT}\cdot\overline{dH} \Rightarrow B_F
\end{aligned} \tag{2}$$

where the term B_F is the fuzzy output variable excited by the firing of rules in the basic, forward walk. The fuzzy AND operator $(\cdot)$ takes the form of multiplication in the current experiments; the more traditional minimum operator may instead be used but is expected to produce similar results.

This method of specification may be seen to form a kind of graphical language for walking or more general patterns. It may be readily applied to more complex walking patterns involving longer sequences and/or more sensor values. One easily accommodated extension involves conditions set up on *both* the feet, such as those encountered in certain dance steps.

In similar fashion we may define the rule set corresponding to the backward walk sequence of Figure 3b according to

$$\begin{aligned}
&\text{A:} \quad && dT\cdot\overline{H} \Rightarrow B_B \\
&\text{B:} \quad && T\cdot dH \Rightarrow B_B \\
&\text{C:} \quad && \overline{dT}\cdot\overline{dH} \Rightarrow B_B
\end{aligned} \tag{3}$$

resulting in excitation of the backward-walk fuzzy output variable B_B.

Note that because of the time-dependent behavior of the dynamic conditions, which are themselves time derivatives of the gating conditions, the fuzzy outputs B_F and B_B tend to behave like narrow pulses along the time axis. (For a natural walking pace, the pulses are typically confined to around 100-300 msec in width.) These pulses are in direct response to fuzzy rule firings and so are indexed by the same time variable which indexes the walking sequence itself. We observed that these pulses could in fact be interpreted as a type of output membership function, only indexed by time rather than by output value as with the more formal definition. Just as in the formal case, these alternative output membership functions are weighted directly and smoothly by the values of the fuzzy antecedents. The difference occurs in that, where traditional output membership functions act as densities along the output value and hence carry their information by their shape, this time-based type of membership function is fixed in shape, at least for individual non-overlapping pulses, and carries its information in the height and relative frequency of those pulses.

In order to determine a meaningful defuzzification for such an output membership function, an analogy was drawn to traditional random processes wherein the mean value of an ergodic process can be found by the time average as well as the statistical average. For such processes the time average serves as a powerful estimate of the mean value, particularly when only time-indexed samples of the process are available and when the underlying probability density of the process is unknown. The statistically-based mean value, being an average along the variable weighted by the probability density, is directly analogous to the traditional defuzzification. The time average employed here takes the form of a classic, first-order autoregressive estimate, i.e. a first-order lowpass filter. We call this filter a "defuzzifying filter". The filter time constant was adjusted arbitrarily so that the real-time performance of the system was not hindered by excessive time lag while generating the equivalent of a "statistically significant" estimate. In practice a time constant of roughly 300 msec has produced favorable results.

Applying this linear lowpass filter to either fuzzy output B_F or B_B serves to produce an adequate "crisp" output representing the inferred walking velocity, at least unipolar in one of the two directions. However, the current graphical application also required a single crisp velocity parameter V which was positive for forward walking and negative for backward walking. This parameter was created by applying (B_F -B_B) to the input of the defuzzifying filter, analogous to placing two singletons (point-mass output membership functions) at 1 and -1. Note, however, from (2) and (3) that this causes an ambiguity for state transition C, where contributions from B_F and B_B cancel. This was addressed by adding a non-linear gate to the input of the defuzzifying filter which favors B_F when the output of the filter is positive and B_B when the output is negative. This gated filter takes advantage of the fact that when walking one tends to slow down before reversing direction, so that in practice the behavior of the input gate is not objectionable. The state C ambiguity could also be addressed by adding the fourth state mentioned previously, along with its associated rules.

4.2. Leaning Gestures

The inferencing of leaning gestures takes a more traditional approach. In the current implementation we only make use of the static condition fuzzy inputs. The direction of leaning is inferred as if the user is standing at the origin of the (*x*,*y*) plane. A ray extends away from the user along the plane. The ray points in the direction in which the user is leaning, and the magnitude of the ray is directly related to the amount by which the user is leaning.

The rule base is a direct map into four unit vectors, two along *x* and two along *y*, conditioned on bounding toe and heel values. Specifically, we have

$$\begin{aligned} T_l \cdot T_r &\Rightarrow y = 1 \\ H_l \cdot H_r &\Rightarrow y = -1 \\ T_r \cdot H_r &\Rightarrow x = 1 \\ T_l \cdot H_l &\Rightarrow x = -1 \end{aligned} \qquad (4)$$

where again the product was used for the AND operation. The rule base is simplified by keeping *x* and *y* independent. Two singletons at 1 and -1 on each axis are weighted by the fuzzy outputs produced by each corresponding rule. The centroid along each axis is then found; for this special case this reduces to taking the average of the two corresponding values. This results in the "crisp" estimates for *x* and *y*, each of which are bounded between -1 and 1, so that the vector result falls somewhere on the unit square. The magnitude and angle versions of this estimate are then found using ordinary rectangular-to-polar conversion.

5. DOMAIN OF SENSITIVITY AND RANGE OF APPLICATIONS

By mounting devices exclusively on a performer we predispose the nature of the information available. Orientation is entirely to the performer's limbs and body angle, without reference to external coordinates. The system in this sense operates in parallel to the weight and motion orientation of the limbs and body. Body-centric cues are complementary to world-centric positional cues from the performer's eyes and gravity-centric balance cues from the inner ear.

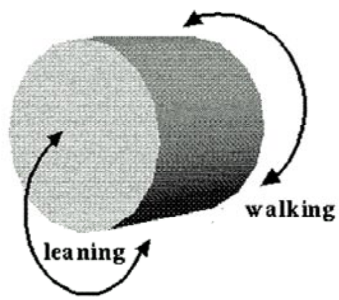
The foot-mounted sensors do not return planar nor polar coordinates fixed to an absolute or world-centric reference. They assume relative foot positions and provide relational information which corresponds to a performer's sensations of weight and weight transfer. The value of these measurements is in the nature of the information that a performer experiences in non-visual sensations of self-directed motion. This information is difficult to measure accurately and inefficient to represent, using externalized spatio-temporal metrics such as geometric coordinates or visual analysis. For example, a person learning to dance must guide their own movements relative to another performer's movements. This is a process of translating sensory-based description of self-motion from visual-based description of external motion. The difficulty of this translation will be familiar to persons who have tried to learn a dance from a series of illustrations of foot-positions. One does not learn to dance by looking at one's feet. The same may be said of athletic performance; we can take measurements including visual analysis and direct physical strength. However most metrics do not provide a commutative function between a geometric or visual value and an interpretation of a performer's internal description of limb motion and feedback from weight orientation.

The foot-mounted sensors are performer-oriented. For example, a world-centric description such as "walk cycle" has a corresponding performer-centric description: (1) the distinction of unidirectional and bidirectional transitions, and (2) the distinction of repeatable and nonrepeatable unidirectional transitions. Walking is continuously repeatable; leaning is non-repeatable without first leaning in the other direction or transferring weight elsewhere, such as from foot to foot, which invokes another form of walking. Thus locally non-repeatable transition sequences may be primitives in larger periodic or quasi-periodic structures. The sensitivity of the foot sensor system is adaptable to larger patterns.

5.1 A virtual reality performance

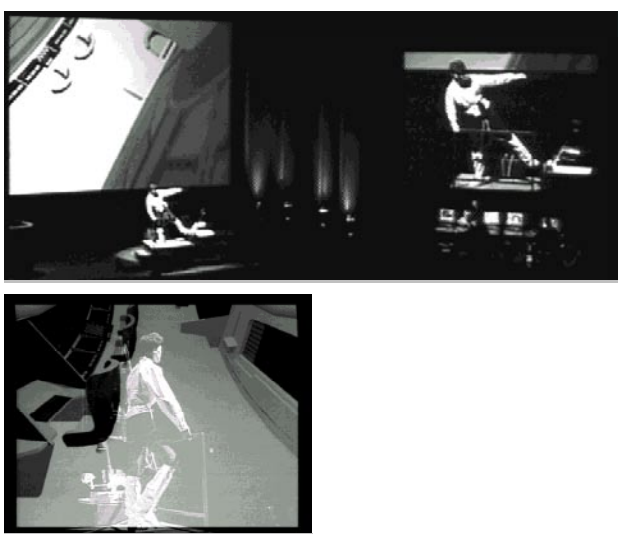
The initial application of CyberBoots was intended to demonstrate base-level functionality by controlling the rate of transformation of geometric objects projected in a graphical scene. In the scene observers are situated inside a large cylindrical space and the performer's walking causes the cylinder to rotate around the observers at a rate corresponding to the walking tempo. Leaning causes the cylinder to rotate on its longitudinal axis (see Figure 4).

Figure 4. Schematic of cylinder rotations and foot actions.



The context for presenting this scene is a virtual reality composition intended for live performance [2]. The theme of the performance piece is a comparison between contemporary intelligent computing technology and fabled advancements in artificial intelligence as depicted in the film *2001: A Space Odyssey* [5]. The cylindrical object is a detailed model of a gravity-generating centrifuge in a space ship. Applying walking motion to rotate this set serves as a reference to specific well-known action performed in the film. A model of the movie set was

Figure 5. CyberBoots in performance.



created in a computer graphics modeling and animation environment* and imported into the virtual environment for motion control (see Figure 5). We apply a damping coefficient for smooth acceleration and deceleration of the rotation. Audiences become involved when they can make a strong visual association between the CyberBoot performance movement and the graphical movement controlled by the CyberBoots driving an underlying computational model. The performer executes a mime-walk pattern in order convey locomotion to the audience, while remaining located in a position centered in front of the projection screen. From this experimental application we can determine that walking patterns are an effective naturalistic method to propel oneself along paths in a virtual scene.

* Alias PowerAnimator version 7.

5.2 Application to movement training

In movement-based activities such as dance or sports, a performer's self-described movement orientation is closely related to his or her level of performance achievement. As a practical application of this technology, we foresee devices that provide feedback closely corresponding to internal sensations of movement, in order to assist a performer to evaluate and modify movements. Sports training and physical therapy are areas where a performer is engaged regularly in movement-based self-evaluation and movement modification. By tracking both locomotion and weight distribution we can search for combinations of transitions that might correspond to a particular movement performance that requires corrective attention. We envision a performer exercising a repertoire of movements while attending to a visual or auditory display controlled by those movements. A performer could listen to musical sequences and fine-tune the sounds by refining his or her corresponding movements [1]. We foresee this technology will have an application as an enhancement of existing devices for measuring physical performance.

6. CONCLUDING REMARKS

The foot sensing and gesture inferencing technology is still in an experimental stage. The continuous kinesthetic presence of a human in a computing interface is a powerful idea. For the intended context, the results obtained were very effective. We were able to intuitively control geometric transformations in a continuous fashion. Observers were able to appreciate and understand the relationship between the actions of the performer and the corresponding motions of virtual objects. These results have enlightened us to the fact that we are not yet accustomed to such a high-bandwidth coupling between human and machine. We are currently studying more complex patterns and the basic properties of patterns, both in the method of their description by humans and in the construction of rules for recognition by machines.

7. BIBLIOGRAPHY

[1] Brady, R., Bargar, R., Choi, I. and Reitzer, J. (1996). "Auditory Bread Crumbs for Navigating Volumetric Data". *Proceedings of IEEE Visualization '96*. San Francisco: IEEE.
[2] Choi, I. *Machine Child*. Virtual Reality composition. Premiere performance, Cyberfest '97, University of Illinois at Urbana-Champaign. March, 1997.
[3] Defanti, T., Cruz-Neira, C., Sandin, D., Kenyon, R., and Hart, J., "The CAVE," *Communications of the ACM*, V. 35, No. 6, 1996.
[4] Klir and Folger, *Fuzzy Sets, Uncertainty and Information*, New York: Prentice-Hall (1988).
[5] Kubrick, S. and Clarke, A., *2001: A Space Odyssey* (theatrical release film). Los Angeles: MGM (1968).

[6] Piaget, J. *Science of Education and the Psychology of the Child*. Paris: Editions Denoel, 1969, trans. D. Coltman. New York: Viking Press, 1971.