Transforming Human Interaction with Virtual Worlds

Mary Pietrowicz, Robert E. McGrath
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
1205 W. Clark St.
Urbana, IL 61801
maryp,mcgrath@ncsa.uiuc.edu

Ben Smith, Guy Garnett
School of Music
University of Illinois at Urbana-Champaign
1114 W. Nevada St
Urbana, IL 61801
bdsmith3, garnett@illinois.edu

ABSTRACT

Virtual worlds, 3D simulations of real or imagined worlds, are far richer and more dynamic than standard 2D computer applications. Extended realities, which integrate an experience between both the physical and virtual worlds, provide even more possibilities. We believe this richness cries out for a more expressive, more powerful, more dynamic human control paradigm. To effect a paradigm change toward traditional human-computer interaction, we are investigating high performance interfaces modeled after the techniques of musicians and other performing artists. We approach the problem by extracting structured information from the actions of performing artists, translating that information into an appropriate control language and applying it to high-performance interactions with virtual worlds. These developments employ automated learning and data mining techniques to extract features and relationships from multiple streams of data (audio, motion capture, etc.), to discover meaningful performative "gestures", and to provide mappings between multiple semantic domains.

Author Keywords

Performing arts, virtual worlds.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Most human interfaces to computers are low bandwidth and are designed for the lowest common denominator. We are no longer in the age where computers are novel to most users and we believe that enormous leaps forward in human computer interaction will result from changing assumptions: if the payoff is great, users will be willing to practice to achieve mastery of new interfaces much like musicians must practice to master the violin. There is no longer a need to make the interfaces as easy as possible at the expense of their full power. We ask the question: What would an interface be like if we wanted to make use of highly skilled and practiced users?

We are investigating a music performance model because we believe the refined and highly practiced skills of these creative performers can help us understand, develop and design new transformative interfaces—perhaps entirely new approaches—to control the complex software systems associated with virtual worlds. Studying musicians will help us bring about a paradigm shift in computer control and give us new insights into designing high performance computing systems. Musicians spend a lifetime mastering their instrument and utilizing its full expressive potential. In the hands of a musician, an extended virtual world becomes an extension of the performer, the instrument, the music, and the performance itself. We believe for at least certain computer users, especially in the realm of virtual worlds, effort spent mastering new technology ought to be rewarded with transformative new capabilities. While we focus on performing musicians, we believe our results will generalize to other trained movement experts including dancers and actors, but also athletes and astronauts.

Achieving these goals requires advances and entirely new approaches to navigating, controlling, manipulating, and interacting within multiuser virtual worlds. This project builds on our existing and ongoing work (mWorlds [1]) to develop and extend its control interfaces to incorporate more fully embodied human gestures through movement and sound allowing us to create a multimodal interface driven—uniquely—by a performing musician model. The performing arts community contains a wealth of formal and informal knowledge about movement and control that has not adequately been utilized in high performance computing systems—this community is both underserved and underutilized.

CHI 2009, April 4–9, 2009, Boston, MA, USA.

OVERVIEW OF CURRENT WORK

Our research group has been developing mWorlds [1], an open framework for collaborative, distributed creation and use of such virtual worlds. This has been used experimentally and elements of it have been used in public musical performances (e.g., [3]) and in a museum installation [4].

We are developing and exploring new paradigms for user interaction and control of such environments within a specific context of performing musicians.

As a first test case, our goal is to create a system that will make it relatively easy for a violinist to generate nuanced, high bandwidth, multimodal data (including the audio output of the instrument and the movements of the performer) and to be able to use that data to control navigation, object creation and object manipulation in an aural and visual virtual world or virtual environment. To the violinist/user, the computer system should be as transparent as possible: they should be thinking of how they are playing the violin and they should see and hear the results without having to worry very much about data acquisition, processing, control mappings, etc. By "virtual world" and "virtual environment" we mean a computer simulation of a physical or pseudo-physical system that is represented as a 3D graphic and sonic world. Virtual Reality systems ([2]) are thus virtual worlds, as are many video and computer games.

To achieve this goal, we are following two tracks: 1) prototyping exploratory, one-off experiments in violin-based control; and 2) creating a framework for generalizing what we learn and for facilitating new developments.

To summarize the general framework briefly: conceptually, there are four distinct components of the system we propose:

1. Data input, capture, and control system.

   Our current implementations use a violinist generating audio signals and movement data as the control input.

2. Analysis, transformation, and interpretation of the multimodal, unstructured input data, and its transformation into knowledge representations.

   Signal processing, pattern matching, machine learning algorithms, etc.), tools for parsing, analyzing, and synchronizing input data across modal boundaries.

3. Operations on the distributed virtual environment itself.

   A multiuser, distributed, collaborative 3D virtual world with many high-level tools for manipulating the virtual objects and virtual environment.

4. Operations generated by the virtual environment on the physical environment.

   High-level tools that allow state and events in the virtual world to manipulate the physical environment (such as lighting, displays, audio speakers, etc.).

An Example Scenario: A "Virtual Concert Performance"

To help motivate and explain the problem, consider the following imagined scenario.

Scenario: Virtual worlds, such as Second Life, are already used for "virtual concerts" and other performances. The artists and audience are represented by avatars in a scene, and the music and words are multicast to the audience. Current environments are very limited: the music is delivered as a stream, and an avatar's movement is only crudely controllable. For example, a violinist is presented as an avatar with a model of a violin, which produces a stream of music. However, the motions of the user/player are not tracked or reflected by the virtual world, and it is difficult or impossible to generate music from the objects in the world (e.g., from the virtual bow on the virtual strings, rather than a recorded or live stream passed to the clients).

Using our system, on the other hand, a highly trained, professional violinist will be able to interact, navigate and project herself into a shared, virtual environment using her performative gestures, and the resultant audio signals, alone. Through a transparent violin-to-virtual-world interface, the musician will control her virtual presentation, eventually including the design and "editing" of virtual objects, virtual navigation, and all other aspects of the virtual world.

Using a combination of audio spectrum analysis, machine learning and data-mining techniques, video motion-tracking and gesture sensors, a rich picture of the violinist's performance gesture data can be realized. These data streams then function as control inputs to a synthetic world. Through careful shaping of a musical improvisation, or careful "performance" of a particular "score," the violinist can propel a virtual avatar through virtual landscapes. With subtle variations in musical timbre the avatar can be driven to mold virtual terrain, create a virtual sculpture, or even "sing" a tune – coming full circle from acoustical musical input to a new electronic musical output now projected in a shared virtual space.

For example, our prototypes have already explored mapping sets of pitches played on the violin to locations in 3D virtual space: playing notes from pitch set A moves you toward location A′, playing notes from pitch set B moves you toward location B′. Relying on the existing skills of musicians—in this case pitch memory, pitch production, melodic improvisation with a collection of pitches, etc.—allows us immediately to begin addressing the heart of the problem: discovering and implementing dynamic mappings between the signals generated by the music interface and the rich possibilities of the virtual world. These interactions enable not only a new realm of aesthetic expressions and experiences as mediated by technology, but also a new model for navigation, creation, and modification of the properties of the virtual world.

THE ROLE OF AUTOMATED LEARNING

The example scenario suggests the complex challenges presented by these ambitious goals, as well as the eclectic mix of analytics that might be applied. The mWorlds software provides a flexible framework for managing multiple streams of data, as well as dynamic 3D scenes, which enables the use of many different analytic components in an integrated system. Many of the individual components we need to realize the above scenarios have already been developed, some by us and some by other researchers.

Two of the major problems to be solved are:

1. analyzing the input streams to translate unstructured data into a well-understood control stream of derived events, and bounded, queryable representations of system knowledge and interpretive meanings from different perspectives and time frames;
2. transforming this control stream into useful and interesting parametric control of the extended physical-virtual world.

This will require simultaneously solving a number of technical challenges, including:

1. managing and storing high-volume streams;
2. identifying gestures and other features purely in the context of a single data stream (stream history, and what is likely to occur next);
3. attaching meaning to the identified gestures and features, purely in the context of the system (the system state, and simultaneous stream activity);
4. attaching meaning, taking into consideration external references, such as who the performers are, their backgrounds and interaction history;
5. attaching meaning in relation to the specific application domain (e.g., music performance, with later generalizations to other domains).

Approach

Our first step is to characterize the input gestures with higher-level semantics. Gestures are identifiable events that can be given meaning. In our case they include physical movements and musical sounds, but can be extended to other temporal phenomena (speech, brain waves, etc.). Gestures may be simple or may be composed of other gestures, but they tend to be interpreted as a single semantic unit. Identifying and summarizing the input as a language of gesture, therefore, will help simplify the representation of stream knowledge and help address the challenges presented by high-volume streams. We hypothesize that this gesture language will define a stream of events that can more easily be interpreted as a control stream for a virtual world. We also hypothesize that input from skilled performers will provide a rich and dynamic, yet controlled and repeatable set of data, which should be well suited for developing these techniques. Finally, we hypothesize that the techniques we develop will translate into other application areas.

We have already started experimenting with audio gesture detection. This requires feature selection and extraction of a hierarchical set of sound properties. We consider fundamental properties (such as loudness, pitch, timbre, and sound attack types) as well as intermediate-level constructs that detect properties over time and compositions of fundamental properties (melody, contour, tempo, pitch centricity, and pitch collections). Higher-level constructs include sound gestures, or shapes that tend to be heard as a single unit by the listener. Some of these include rapid alternation (trill, vibrato), spectral change (timbre), sweeps (rapid runs in a single direction), and compression/expansion.

We have used automated learning techniques to successfully process sound streams to detect transposition within a musical set class. While this is a simple problem, easily achieved by other methods, it allowed us to test our framework and verify results. We used a sliding window technique both for training and processing, and achieved greater than 90-95% accuracy with the "Decorate," "BayesNet," and "LMT" classifiers, and with a "Multilayer Perceptron" algorithm. These classifiers handled transitional passages particularly well and correctly classified the gray areas with reasonable probability of belonging to more than one set class. We also demonstrated our analysis by generating an accompaniment that followed the transposition changes in live sound.

One of the challenges of sound gesture analysis is the inherent fuzziness of gestures in real music. Gestures often combine (sweeps that include trills), or deviate from the ideal (a sweep that is a bit slower than expected), or combine or overlap in ways that make segmentation difficult. Simple heuristics and single analytic models are not well suited for this challenge. We are addressing this challenge by extending our earlier experiments to include dynamic networks of analytic processing. The networked aspect of the solution will enable detection of an arbitrary number of gestures and combinations of gestures. The dynamic aspect of the solution will support changing the processing as the music changes, or as the performer provides feedback to reinforce or discourage the system responses in real time.

To begin exploring this approach, we developed a reference application that processes audio data from a live musical performance and detects the degree of disjunction and separation among the notes played. We refer to this as the degree of "pointillism." Highly pointillistic sound features maximum difference in timbre, loudness, and pitch. The listener may perceive this sound as being "pointy" or "spiky", so, for demonstration purposes, we translated this quality in the degree of spiky effects in the visual, virtual world. As the sound became more pointillistic, the visual object (in this case, an icosahedron) in our virtual world became more “pointy” or “spiky.” In addition, we analyzed the distribution of pitch classes (the twelve notes in the chromatic scale), and mapped each pitch class onto one of the twelve vertices on the icosahedron. When a musician played something, the icosahedron changed shape to reflect both the distribution of pitch classes and the degree of pointillism.

Our reference application included three stages of processing: 1) heuristic analysis components, 2) machine learning modules, and 3) output transformation. The heuristic analysis included components for pitch class analysis, intervallic distance, silence–to-sound ratio, and direction change. This layer of preprocessing extracted quantifiable features from the sound and abstracted away difference caused by individual musical instruments or players. The machine learning classifiers had, at this point, a clean control stream which they used to classify music on a continuum ranging from “very smooth” to “smooth” to “mixed” to “pointillistic” to “very pointillistic”. Then, the output transformation resolved the output into a control stream that the virtual world used to transform the appearance of an icosahedron in real time.

Another challenge appears when we try to analyze the large variety of less well understood signals generated by motion tracking the musicians. It is relatively easy to use the one-to-one correspondence between human and avatar as a stand in for real knowledge—explicitly mapping hand to hand and foot to foot, with implicit embedded knowledge about joint range of motion and related physiology, but extracting meaning from a sequence of gestures is much more difficult. A beginning approach to improving this situation was taken with respect to the specific gestures involved in music conducting by Garnett, et al. [5].

Our current approach is to extract higher-level features, and to combine features from many sources (audio signals, movement tracking, etc.). Thus, we seek to recognize meaningful “gestures” not only in movement, but also in musical performance and other human actions, allowing us to correlate many different gesture modes. Together these form a synthetic composite input, which—because the musicians are highly trained—is repeatable and meaningful to the human users, and tractable for the computer system. We believe that our interpretation of the inputs as gesture language will simplify processing while preserving knowledge.

The second step is to transform the interpreted gesture language (along with system state) into appropriate actions in the extended reality. To do this, we will “annotate” the gesture language with concepts that capture the relationships within a single stream, within the context of the whole system, and with external references related to the performer and the performance environment. We will characterize actions in the virtual world with similar concepts, and select actions that provide a coherent conceptual mapping. We treat the whole system, physical device, mapping software, and virtual world much as a musical instrument creating a strong feedback loop between the performer and the resultant auditory and visual displays that will be responsive to the “human in the loop,” to his practice, training, and learning. Our approach augments this with machine learning and real-time feedback to create dynamic, many-to-many mappings between the various parameters making up the performer’s movement (such as velocity of hands or feet, relative position of limbs, orientation of head, etc.) and a multi-parametric representation of control features within the virtual world.

The element of feedback is crucial—both the machine and the performer will learn to adjust their responses until they arrive at a satisfactory result. This will lead to a deeper level of meaning in human-computer interactions tapping into intuitive yet highly trained and skillful actions of a professional musician.

Instead of relying on just these one-to-one mappings from simple interface devices, we will analyze the relationships among the various streams of data we extract from the performer over time and apply state-of-the-art machine learning techniques to make adaptive multi-layered mappings. The quality of the motion, and the amount and kind of aural and visual and kinesthetic feedback being more important than merely measuring quantities of change on a few axes of motion, this approach will allow us to interact with the world on a deeper level.

While it is not likely there is a fully general solution that will fit all possible applications or artistic visions, we need a flexible system containing all of these components (sensor inputs, real-time data analysis tools, and virtual world creation tools) to begin a fuller exploration of the possibilities.

Our initial explorations have uncovered a set of novel approaches based on the extreme flexibility, nearly infinite variability, and extraordinary control and repeatability that performing musicians have achieved in a lifetime of practice and study. As a first approach, we are categorizing some of these new possibilities and exploring systems that allow for flexible, user-definable, mappings between the complex domain of music instrument performance “gesture” and the resultant virtual world. We will use machine learning techniques to formalize and quantify these “gestures”.

CONCLUSION

The mWorlds project is developing and exploring new paradigms for user interaction and control of virtual environments within a specific context of performing musicians, to develop and extend its control interfaces to incorporate more fully embodied human gestures through movement and sound, and allowing us to create a multimodal interface driven—uniquely—by a performing musician model.

Automated learning and data mining techniques are a critical technique for dynamic analysis of sensor data to understand the users behavior, and to map between the real world and the virtual world.

Whether or not we can achieve breakthroughs in designing and conceptualizing user interfaces to virtual worlds, we are certain to open up entirely new areas for enhanced creative exploration.

ACKNOWLEDGMENTS

Special thanks to the Institute for Advanced Computing Applications and Technologies (IACAT) and the National Center for Supercomputing Applications (NCSA) for supporting our collaborative work.

REFERENCE

1. Garnett, G., McGrath, R., Campbell, R.: Virtual Worlds: Infrastructure for Large-scale Collaboration. (2007)
2. Sherman, W.R., Craig, A.B.: Understanding Virtual Reality: Interface Application and Design. Morgan Kaufmann, San Francisco (2003)
3. Smith, B., Garnett, G.: MusiVerse. International Computer Music Conference, Copenhagen (2007)
4. Smith, B.D.: Musiverse. CANVAS (Collaborative Advanced Navigation Virtual Art Studio). Urbana (2008)
5. Garnett, G., Jonnalagadda, M., Elezovic, I., Johnson, T., Small, K.: Technological Advances for Conducting a Virtual Ensemble. International Computer Music Conference (2001) 167– 169