Multimodal Gestural Interaction in Performance

Mary Pietrowicz
National Center for Supercomputing Applications,
Institute for Advanced Computing Applications and Technologies,
University of Illinois
at Urbana-Champaign
maryp@ncsa.illinois.edu

Robert McGrath
National Center for Supercomputing Applications,
University of Illinois
at Urbana-Champaign
remcgrat@illinois.edu

Guy Garnett
School of Music,
Institute for Advanced Computing Applications and Technologies
University of Illinois at Urbana-Champaign
garnett@illinois.edu

John Toenjes
Department of Dance,
Institute for Advanced Computing Applications and Technologies,
University of Illinois at Urbana-Champaign
jtoenjes@illinois.edu

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract

Fully embodied interaction employs the complete range of gestural modalities available to humans: from finger, hand, and arm movement, to whole body movement, to vocalizations and other sound production, to the use of tools and instruments that augment the natural range of human expression, and beyond. We endeavor to facilitate exploration of new paradigms for such interaction by translating sensed input data into abstract control streams, providing a network of signal processing and analytic tools for gestural interpretation, and facilitating the translation of the processed control streams into arbitrary interactions. We have experimented with fully-embodied interactions that control parameterized 3D geometric models projected onto 2D or 3D surfaces, and, in parallel, control algorithms that produce sound. Our overarching goal is to enable the fully embodied human to control any and all aspects of a virtual and/or augmented reality environment. This paper presents aspects of the mWorlds [1,2,10] framework that enable the exploration of multimodal gestural controls for interactive environments, and describes our application of it to a series of performances.

Author Keywords

Interactive performance, virtual world, machine learning, multimodal gesture.

ACM Classification Keywords

H5. [Information interfaces and presentation]: Multimedia Information Systems (Artificial, augmented, and virtual realities), Sound and Music Computing.

Introduction

Current approaches to human-computer interaction fail to exploit the complete range of human multimodal interaction with the world.   Many approaches restrict human-computer interaction to single modalities, or replicate the limited interaction vocabulary of a mouse and keyboard. Everyday interaction with the physical world, including with other humans, however, uses a wider range of modalities such as sound (including spoken language, vocal gesture, music and other instruments), physical gesture (including bodily movements, face, and eye movement, and object manipulation), and visuals (image creation, and writing). We propose a flexible framework for exploring this wide range of multimodal gestural interactions in immersive environments.

Our work is the result of the collaboration of technologists, dancers and musicians. We believe the refined and highly practiced skills of creative performers can help us understand, develop and design new interfaces—perhaps entirely new approaches—to interacting with complex software systems. Furthermore, the performing arts community contains a wealth of formal and informal knowledge about movement and control that has not adequately been utilized in computing systems—this community is both underserved and underutilized. The benefit of our approach is, therefore, symbiotic: 1) the performing arts community provides exciting opportunities for insight into human-computer interaction, and 2) a dynamic framework for multimodal gesture interaction creates new opportunities for creative works in the performing arts.

Framework Overview

Our current work with human-computer interaction in mWorlds divides the problem into three stages: 1) Multimodal Gesture Input, 2) Multimodal Gesture Recognition, and 3) System Action/Response. In the realm of multimodal input, our goal is enabling a diverse, flexible, dynamic, and concurrent set of streaming inputs, such as sensor network data, live music, live video, voice, and more. In the analysis phase, the goal is enabling diverse, flexible, dynamic, and concurrent analysis. After the data has been made available to the system, the multimodal gesture processing and recognition layer interprets the data streams and produces control streams describing the gestures present in the data. The analysis layer may recognize a variety of gestures, including musical, motion, vocal inflexion, or composite combination gestures. The resultant gesture control streams may characterize a single input channel, a performer, a group of performers, or the entire performance. Furthermore, gesture analysis occurs in real time, over varying time scales, with varying inputs. Finally, the system action/response layer interprets the processed control streams, and responds, according to the application, data, state, and available output channels. Framework design issues for multimodal interaction are explored further in [9].

Dynamic, Diverse Inputs

Our approach to enabling a diverse set of embodied, multimodal inputs has been to leverage and extend existing capabilities and to create a framework for various services to work together. For example, we use the extensive capabilities of Max/MSP [11] for capturing data from many sources, including audio, serial inputs, and Bluetooth devices. The input client will manage driver-level interaction with the data source, and may do some preprocessing to create a coherent output stream of data that analysis components downstream can analyze. We connect this low-level input client functionality by providing a communication path with an event service that provides metadata about the connection and data content, optionally registers an event handler, and begins sending and receiving data. Generally, we have used Open Sound Control (OSC) [14] to send low-level sensor data from the input client to the consuming event service. The input client may do some preprocessing and analysis (as needed) before sending the data to the event service, but it should, in general, delegate the analysis processing.

Figure 1: Blue Lights in the Basement
April 20th, 2009
Ben Smith, violin;
Mary Pietrowicz, flute;
John Toenjes, percussion

Dynamic, Diverse Analysis

At the analysis stage, we see a significant need for dynamic processing. Inputs can drop in and out, and networks of analyzers must be able to adapt in real time. In our current and proposed work, we extend the approach we used in processing inputs to analytics, via registration and description of analytic components, which may include networks of processing, heuristic processing, and machine learning-enabled components.

Dynamic, Diverse Actions and Responses

The flexible description and dynamic handling provided in the input and upstream analysis processing enable flexible, dynamic mapping of actions, depending on the system state and the available presentation channels. To realize these goals, we are developing a framework that supports action description and the discovery of action services. We also propose the description, registration, and discovery of mapping objects that describe the desired system response, given system state and event input. This approach enables flexible, dynamic mapping of events and state to actions, where the desired action can evolve over time. In this way, the interface can grow with the desires of the users, fit each application, adapt to different situations and states, and support individual preferences.

Performances and Workshops

We presented our approach in public performances, including a distributed, live music performance, a workshop for the analysis and creative application of Laban Effort Actions, and a restaging of the Trisha Brown dance “Astral Convertible”. We are currently preparing for another performance at HASTAC 2010, which will integrate motion and sound sensing, expand the range of motion sensing for the performers, and explore new ways of analyzing, interpreting, and visualizing the performers’ actions.

HASTAC III Concert

As part of the HASTAC III conference [4], we presented an improvised performance by three live musicians at two locations, with the performers’ sound driving objects in a virtual world projected to the audience in real time. The performance was realized by software which analyzed multiple sound streams using heuristic analysis, signal processing, and machine learning techniques, and reflected the results of the analysis visually. Different sound qualities mapped onto several different objects in the virtual world in real time: particle system effects, an icosahedron, and an elastic, multi-jointed “bones” object. For example, one network of analysis modules produced a real time stream containing a combined pitch class profile and degree of disjunction present in the sound. The analytic results were transformed into a stream which modulated the behavior of the icosahedron. Each of the twelve vertices on the icosahedron mapped to a pitch class, and each vertex extended and contracted according to the amount of the related pitch class in the sound. In addition, the degree of disjunction present in the sound made the icosahedron appear more spiky or smooth. Figure 1 shows the performance from both locations.

This performance used multiple analytic methods: signal processing techniques to input sound, heuristic techniques to produce a pitch class profile, and machine learning techniques to detect the degree of disjunction in the sound (gesture detection). The results of these analyses were transformed into a new control stream which changed the appearance of the icosahedron (action/response). The streams of data were transported into mWorlds using OSC, and mWorlds reflected the results back to the performers and audience visually.

The added dimension of the virtual world changed the nature of the performance. In effect, the virtual world created a new extended instrument for each performer, and a new combined instrument and immersive environment overall. Performers played not only to create sound and sound combinations, but visual combinations, and visual- sound combinations. The performers' experiences also informed the technology and shaped the next development steps. We observed the need for 1) expanded and more efficient stream handling framework, 2) an expanded set of dynamic, generic 3D representations, 3) flexible, dynamic analysis frameworks, and 4) more flexible, dynamic mapping to actions and responses. The technology created new opportunities for the artists, and the artists' experiences provided feedback that shaped the next software development steps.

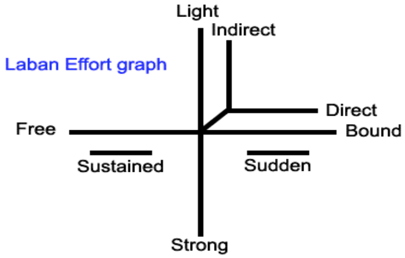

Figure 2: Combined Graphical Representation of Laban Efforts (from [12])

Laban Gesture Analysis Workshop

In September 2009, we hosted a workshop [7], led by Canadian artist Thecla Schiphorst, that explored the sensing and interpretation of Laban Effort Actions. Laban Movement Analysis (LMA) provides distinct and definable delineations of movement that we discovered are a useful starting paradigm for gesture recognition applicable to artistic ends [13]. In this interactive demonstration, we instrumented each dancer with a wristband containing a small hardware package consisting of a controller board, motion sensor, RF transmitter, and battery. The eight Laban Efforts direct/indirect Space, strong/light Weight, quick/sustained Time, and bound/free Flow—and their combinatorial Actions—Float, Punch (Thrust), Glide, Slash, Dab, Wring, Flick, and Press—were analyzed in real time, and were reinforced visually by the appearance on a projection screen of their associated graphical notation, and in sound through sample playback of dancer's pre-recorded vocalizations of the movement.

The Laban Efforts were detected via supervised learning algorithms (using Weka [5]). Certified LMA experts trained the system to recognize the different Effort Action types, and then the machine learning components were directed to classify motion according to how much of each kind of Laban Effort Action was detected in the movements of the dancers. The results of the classification were presented visually in a projection of graphics based on the standard notation of the Laban Efforts (Figure 2 shows a map of the Laban efforts, which defines a graphical markup of the Effort Actions).

For this workshop, we used only one sensor per dancer in order to eliminate unnecessary complexity (for the scope of the workshop). The single sensor was all that was necessary to prove concept and train the system to recognize ideal Laban Effort Action Types. With the simple system, we were better able to evaluate the overall approach, and gain understanding of the efficacy of our handling of technical details such as sampling rate, time window sizes, heuristic analysis vs. machine learning techniques, stream management, preprocessing (what derived data would be most useful and efficient for the analytics), distribution of processing across a network, and overall performance. We were able to isolate and improve handling of these essential details, which made it easier to scale the system out to multiple sensors per performer in subsequent iterations.

The workshop provided valuable experience and feedback about our approach. Participating dancers experienced visualizations of their motions, which in turn, affected how they performed, and envisioned new opportunities for training, rehearsal, and artistic expression (as well as more therapeutic or scientific uses for this technology). The visualizations enabled the workshop sessions to be highly interactive. In some sessions, the "audience" actively coached dancers using the visualization.

From the workshop, we discovered that we needed more frequent samples over shorter time windows to detect quick, light gestures, and developed simultaneous, weighted analysis over multiple time scales to optimize recognition of different classes of motion. Although we could do significant work with single-sensor processing per performer, we observed the need to expand to multiple sensors and sensor types for more sophisticated processing; and we observed the need to improve processing efficiency of real-time sensor streams. To improve performance, we distributed processing across network resources.

We also discovered that preprocessing the data to compute changes in acceleration (as opposed to acceleration values read directly from the sensors) improved accuracy and guarded against variation in the calibration of the accelerometer hardware. We derived values such as change in acceleration, rate of change of acceleration, change averages across multiple time scales, ratios of instantaneous change vs. average change over multiple time scales, average number of direction changes, and relative stillness. These values were useful as inputs for machine learning models, heuristic analysis components, and "edge" detectors for changes in motion type. We continue to use these values in future iterations of our work.

Conceptually, we were forced to reconsider our presentation of movement analysis, and our machine learning training methods, because during our experimentation we were confronted with questions such as: "How do you perform 10% of a gesture? Does that not constitute a different gesture altogether?" "When does one gesture end, and another begin?" Answering these questions will inform subsequent approaches to analysis and presentation as we continue with this work. Once again, performers and technologists informed each other.

Restaging of "Astral Convertible" Dance

We contributed to a restaging of Trisha Brown's dance, "Astral Convertible" [3, 7], which was presented at the Krannert Center for the Performing Arts February 4-6, 2010. In this performance, we used the dancers' gestures to modulate live sound and visuals. Dancers wore sensor hardware, consisting of 5DOF accelerometers, Funnel I/O boards, XBee RF communication modules, and batteries, which fit easily into small pockets in the costumes. The technology was selected with the following criteria in mind:

1) small, light, and easy for dancers to wear,
2) low cost,
3) long battery life,
4) long communication range,
5) safety,
6) providing sufficient data for accurate recognition of choreographed motions, and
7) providing a platform for future work and experimentation with a range of sensor inputs.

The software running on-board the controller sent the accelerometer data continuously over the XBee RF interface, along with header information that identified the dancer, controller, and individual sensor. Each dancer wore two sensors (5DOF accelerometers), one on the chest, and another on the right arm. We selected the number and position of the sensors based on the choreography, the gestures we wanted to detect, budget, and complexity.

An XBee RF coordinator module received the sensor data at 100 samples per sec. per sensor, and transmitted it to a Max/MSP ([11]) application, which routed the data to the correct channel for analysis. Each analysis channel evaluated the data stream over three window sizes (0.5, 1.0, and 2.0 sec.) simultaneously to generate 45 heuristic measurements that characterize motion within each time window. Window sizes were easily configurable, which enabled tuning and experimentation.

Next, the preprocessed data flowed into machine learning algorithms, which (aided by heuristic analysis informed by our experience at the Laban workshop) had been trained to classify motion into gestures relevant to the choreography. The outputs of the classifiers were weighted according to the gesture and window size to produce a single motion analysis summary for each sensor set. These results were combined to produce a single motion analysis summary for each performer. Finally, the individual performer results were combined to produce a single motion classifier stream for the entire cast of dancers as a whole. The system sent per-dancer and performance-wide motion classification profiles via OSC to downstream applications that responded to the data with light, sound, and projected visuals. See Figure 3 for a system overview.

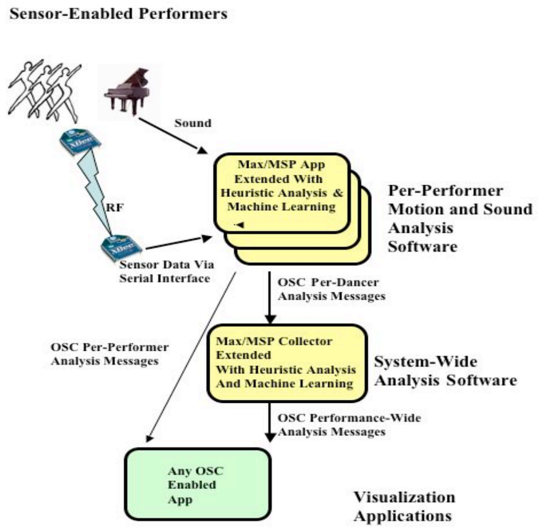
Sensor-Enabled Performers
Sound
RF
Sensor Data Via Serial Interface
Max/MSP App Extended With Heuristic Analysis & Machine Learning
Per-Performer Motion and Sound Analysis Software
OSC Per-Dancer Analysis Messages
OSC Per-Performer Analysis Messages
Max/MSP Collector Extended With Heuristic Analysis And Machine Learning
System-Wide Analysis Software
OSC Performance-Wide Analysis Messages
Any OSC Enabled App
Visualization Applications

Figure 3: Astral Convertible Hardware/Software System Overview Figure

HASTAC 2010 Performance

We are in the process of developing a multimodal performance for the HASTAC 2010 conference [6] which will include live music and dance; sound, motion, and composite gesture analysis; interactive visuals projected onto 2D and 3D surfaces; processed sound; and control of external devices such as lighting. The challenges for this iteration are 1) integrated analysis of multimodal gesture (such as dancers' motion, musicians' sound, and musicians' performative gestures), and 2) aesthetically coherent reflection of the analytic results in the sound processing, lighting, and visual presentation. This approach extends the musician's instrument and the dancer's body. No longer will they be controlling only their own motion and sound, but they will also control – collectively – a virtual world and the physical environment around them.

In an effort to contextualize data in emotional and humanistic terms, we may also interpret performance data in terms of "community." That is, as the system attempts to detect unisons, canons, and other groupings in motion, these determinations are transferred to concepts such as amount of "togetherness" or "factionalism" or other group dynamics. This analysis begins to bring pure data into the realm of human behavior, which is largely what theatrical performers are concerned with, facilitating interaction among engineers and artists.

Conclusions

Our collaborations with performers and movement specialists are leading us to explore new modalities for human-computer interaction. Their refined and highly practiced skills provide insight into new approaches and focus our work on the core technical problems to solve. Performers, in turn, gain new avenues for creative expression and the opportunity to explore new techniques, and investigate new audience experiences. Future performances will explore a wide range of combined sensor inputs, sound, visuals, and gestures.

We anticipate challenges in developing dynamic analytic and control mechanisms, creating analytics that detect a hierarchy of qualitative attributes, and mapping the results of the analysis into actions that produce consistent human-computer interaction and aesthetically pleasing visualizations.

We believe that the mWorlds framework has implications far beyond the performing arts. In developing generic components that enable dynamic, multimodal control, analysis, and action-response mappings, we extend the general capabilities of immersive and augmented environments. Future work may include collaborations with scientists, educators, or other practitioners to explore the many possibilities inherent in this framework.

Acknowledgements

Special thanks to the Frances P. Rohlen Visiting Artists Fund, CAS George A. Miller Visiting Professors Fund, the UIUC Campus Research Board, the Institute for Advanced Computing Applications and Technologies (IACAT), the National Center for Supercomputing Applications (NCSA), and eDREAM (Illinois Emerging Digital Research and Education in Arts Media Institute) for supporting our collaborative work.

References

[1] Garnett, G., McGrath, R.E., and Pietrowicz, M. mWorlds: Novel Human Interaction with Virtual Worlds MardiGras 2009.

[2] Garnett, G., Smith, B., Pietrowicz. M., and Toenjes, J. Bluelights in the Basement Late Night Krannert HASTAC III: Traversing Digital Boundaries, April 2009.

[3] http://news.illinois.edu/news/10/0122astral.html

[4] http://www.chass.uiuc.edu/hastaciii.

[5] http://www.cs.waikato.ac.nz/ml/weka

[6] http://www.hastac.org/events/hastac-2010-grand-challenges-and-global-innovations-conference

[7] http://www.ncsa.illinois.edu/News/09/0420Visiting artist.html

[8] http://www.trishabrowncompany.org

[9] Pietrowicz, M., McGrath, R., and Garnett, G. A Framework for Enabling Multimodal Gestural Interaction, submitted to Natural User Interfaces: the prospect and challenge of touch and gestural computing, Workshop at CHI 2010

[10] Pietrowicz, M., McGrath, R.E., Smith, B., and Garnett, G. Transforming Human Interaction with Virtual Worlds in Workshop for Computational Creativity Support at CHI 2009.

[11] Puckette, M. Max at Seventeen Computer Music J (26:4), Dec. 2002, pp 31-43.

[12] Schiphorst, T., Toenjes, J., Hook, S., and Pietrowicz, M. Wearable Computing for Art and Performance, a Workshop in Computer Movement Analysis and Intermedia Performance Sept. 2009, University of Illinois Krannert Center for the Performing Arts.

[13] Wikipedia, "Laban Movement Analysis", http://en.wikipedia.org/wiki/Laban_Movement_Analys

[14] Wright, Matthew, Open Sound Control: an enabling technology for musical networking. Organised Sound, 10 (3):193-200, 2005/12/01 2005.