Using IR Optical Marker Based Motion Capture for Exploring Musical Interaction

Ståle A. Skogstad,[♮] Alexander R. Jensenius,[♯] Kristian Nymoen[♮]

[♮]University of Oslo
Department of Informatics
Pb 1080 Blindern, 0316 Oslo, Norway
{savskogs, krisny}@ifi.uio.no

[♯]University of Oslo
Department of Musicology
Pb 1017, Blindern, 0315 Oslo, Norway
a.r.jensenius@imv.uio.no

ABSTRACT

The paper presents a conceptual overview of how optical infrared marker based motion capture systems (IrMoCap) can be used in musical interaction. First we present a review of related work of using IrMoCap for musical control. This is followed by a discussion of possible features which can be exploited. Finally, the question of mapping movement features to sound features is presented and discussed.

1. INTRODUCTION

Motion capture (MoCap) is a term often used to describe the process of recording human body movement and storing it in the digital domain. Many different disciplines make use of MoCap systems, and they can briefly be divided into two groups: *analysis* and *synthesis*. The first approach (analysis) is typically found in fields working on bio-mechanical research questions, e.g. medicine, rehabilitation and sports science. The second approach (synthesis) can be found in the entertainment sector, where MoCap systems are used to create lifelike animations in movies and computer games.

Many different MoCap technologies exist [1], and we will here choose to split them into two different groups: *optical* and *non-optical* systems. Among the non-optical systems, one of the most affordable solutions is that of *inertial* sensor systems, based on sensors such as gyroscopes, accelerometers and magnetometers. While each such sensor outputs relevant movement data in themselves, MoCap systems based on such sensors typically perform *sensor fusion* on the raw data. Sensor fusion means that data from the individual sensors are combined such that it is possible to integrate the data to calculate position (and sometimes orientation) with fairly little drift. On the positive side, such systems are often portable and flexible, and provide good value for money. Unfortunately, they often provide poorer spatial accuracy and precision than optical systems, and have problems with the measured position drifting over time.

Mechanical MoCap systems are based on directly tracking the angles of body joints through the use of flex sensors. Such systems are often flexible and durable, and have been used for many creative applications.

Magnetic systems calculate both 3D position and 3D orientation based on moving a coil in an electromagnetic field. They often give precise and reliable data, but have a comparably small capture volume. Another big drawback is the susceptibility to magnetic and electrical interference.

While they have many positive sides, inertial, mechanical and magnetic systems share one problem: they usually rely on fairly large sensors that have to be attached with cables to the computer. Exactly this is what makes *optical* MoCap systems preferable in many contexts, since they provide for a non-obtrusive and flexible solution.

Optical systems can be divided into visual *markerless* systems and *marker based* systems. Both these techniques rely on computer vision techniques for extracting movement features and tracking body parts. Although markerless computer vision techniques are in rapid development, the marker based solutions still make for more accurate, precise and fast tracking. Optical MoCap has been particularly popular for creative applications, due to the low cost, flexibility and availability of relevant tools, e.g. Max/MSP/Jitter and EyesWeb [3].

The technique which is often referred to as state of the art in the world of MoCap, is what could be called *optical infrared marker based motion capture* (IrMoCap). This is based on a group of cameras, typically no less than 6, surrounding the person(s)/object(s) to be tracked. The cameras emit infrared light which is bounced off reflective markers attached on the body of the person being observed and captured by the cameras. Through triangulation techniques the system calculates the absolute position in space, with submillimeter resolution and at speeds above 500 Hz. By combining multiple markers it is possible to uniquely identify certain objects, something which may also be accomplished using active markers that emit their own light.

We have experience with all of the above mentioned MoCap solutions, and see that they all have positive and negative effects. In our current research, however, we have decided to focus our attention on IrMoCap, since this is the technique which currently provides for the most precise, accurate and fast MoCap solution. On the negative side they are expensive and requires a controlled lab setting to work properly. This is because the system needs to be calibrated thoroughly and is sensitive to light pollution. Despite these drawbacks, we believe that the knowledge and experience gained from using such systems may be transferred to other more accessible and affordable MoCap technologies in the future.

Our main research goal is to explore the control potential of human body movement in musical applications. By combining high quality MoCap data with advanced machine learning techniques, we try to explore multidimensional mappings between motion features and sound features. Here we are interested in exploring everything from

NIME2010, Sydney, Australia

direct control, like playing an instrument, to more *indirect* control, i.e. controlling more global features in the sound and musical structures. We want, in other words, to explore the possibilities of using new technologies to increase the connection between human motion and musical expression.

2. RELATED WORK

We have only found a few studies that have been published on using IrMoCap systems in musical interaction, and we have chosen to separate this into two categories: non-real-time and real-time.

2.1 Non-real-time control

Dobrian et al. describes a system where data recorded with an IrMoCap system can be mapped to MIDI signals [5]. Their software makes it possible to choose which marker and its associated motion feature that should be mapped. The motion features include marker position, velocity, acceleration, and distance and between markers (in one, two or three dimensions). In addition to linear mappings, the software also allows for reversed, exponential, logarithmic mappings.

An important point that Dobrian et al. reflects upon, is that performing on a 'touchless' instrument both provides a challenge, but also opens for interesting musical explorations. We also share their interest in trying to develop strategies for keeping multidimensionality (e.g. data from 30 3D markers) throughout the mapping process.

One of the challenges when working with IrMoCap is the massive amounts of data that has be to handled, e.g. 30x3 marker values for each recorded frame. Bevilacqua et al. report on developing techniques for segmentation of the movement stream and what they call 'gestural segmentation' in [2]. Here they describe some of the numerical problems of computing velocity and acceleration from noisy data and point out that filtering is important, but that it also adds latency to the system. They experimented with using principal component analysis (PCA) for feature extraction, and using the output for controlling MIDI systems and signal processing.

2.2 Real-time interaction

The first example we have found of using IrMoCap in real-time musical applications is a project by Qian et al., in which they used "a number of static human body gestures (poses) to drive the interactive system" [21]. They divided the body into 10 rigid 'objects,' and used angular relations as features for the pattern recognition classification. This was used to control granular and additive sound synthesis, where pitch material were selected through a simple genetic algorithm. Unfortunately, we have not been able to find any video examples of their performance to evaluate the approach.

Other examples of real-time applications include Woolford's use of IrMoCap to visualize and sonify body motion in installations [25], and Downie's experimentation in a stage setting [7]. We see that many research groups get access to and set up projects around IrMoCap technologies, one example being the *Embodied Generative Music* project at IEM in Graz [9]. They have been experimenting with an installation where you prerecorded music is 'laid out' in physical space, and where it is possible to explore the "tactile" feeling of sound in space.

2.3 Sonification

A related but still different approach is that of Kapur et al., where the goal is to build the necessary infrastructure to study the use of sonification for understanding human motion [17]. They are interested in studying how the musician's posture and movement during performance affect the sound produced, as well as the emotional content of the performance. They also hope that studying sonification of IrMoCap data can aid individuals with motor disorders. The study did not involve real-time examples but used recorded data of people performing music (tabla and violin), dancers acting out different emotions, and individuals having impairments in sensory motor coordination. The sonifications consist of mapping marker positions to control sinusoidal oscillators, FM synthesis, phase vocoders and physical models of instruments.

In the same direction we find work related to sonification of IrMoCap data from musicians' 'ancillary gestures', with the aim of providing an alternative perspective when analyzing movements of musicians [23, 11]. This was also done by Larkin et al. in a project where IrMoCap data of string performers were sonified, intended as an interactive feedback to the performer [18]. Vogt et al. have a similar approach with applications in physiotherapy and other training contexts [24].

3. MOTION EXTRACTION

Our research goal is to study the capabilities of IrMoCap in the context of musical expression. The challenge then is to develop solutions for extracting meaningful information from the continuous stream of data, and map these to relevant features in the musical sound. This is both a question about making an interpretation of the data, but also a technical challenge when it comes to handling marker occlusion problems, data noise, latency and computational and numerical challenges.

In the context of optical MoCap, Camurri et al. [4] have suggested a four-layer framework that can be useful for our application:

- Layer 1: Physical signals
- Layer 2: Low-level features
- Layer 3: Mid-level features
- Layer 4: Concepts and structures

Separating between the different layers may help to structure some of the challenges, both conceptual and technological, and will form the basis for our thinking about IrMoCap data processing in the following sections.

3.1 Marker and Object Data

The first and second layers in the model of Camurri are related to the physical signals and low-level features, and is related to the output we get from a IrMoCap system: 3D positions of the markers that the cameras can see. These markers, passive or active, can be placed directly on the human body or placed on objects that can be moved in the space.

In addition to tracking the position of an object, it is also possible to find the angular orientation of an object by placing 3 or more markers on the object's surface. Here we are experimenting with having many objects, all with unique marker constellations, so that it is possible to uniquely identify all the objects. This will make it possible to play with all these objects in the motion capture area simultaneously.

3.2 Mapping Markers to a Kinematic Model

Instead of dealing with a vast amount of isolated markers and/or 6D objects, we are also exploring techniques for grouping them together and study how they move in relation to each other. This can be accomplished by defining one or more *kinematic models*, e.g. of the human body. But it can also be possible to define kinematic models for

other types of composite systems, e.g. a movable sculpture. Defining a kinematic model can be done by representing the data as several connected solid objects with the respected joint angles between adjacent solid body parts [21]. A benefit of such an approach is that it helps in decreasing the dimensions of the data set, and can provide us with more meaningful data.

3.3 Manipulation of Parameters

There are endless possibilities for manipulation of the above mentioned parameters: change the scale of axes, invert signals etc. It is also possible to extract different relationships between markers, e.g. relative distance and angles between points. Further on, it is possible to perform numerical calculation on the output streams to obtain properties like velocity, acceleration, jerk etc. All of these, however, are only numerical approximations, and noise from the data will propagate through the computations and possibly be amplified by the numerical algorithms [5]. These numerical computations should therefore be done with care. Filtering is a possible solution to get less noisy results, but a filter and other computations will at the same time add latency to the system.

3.4 Spatial aspects

Moving towards mid-level features, there are many questions when it comes to how to extract meaningful information from the continuous data sets. One approach here is to look at spatial aspects of the data. A kinematic model of the human body can be a good starting point for extracting information about specific body postures and placement of the body in space. Information about different body postures can for example be mapped to different sound features, and it may be possible to morph between discrete postures.

3.5 Temporal Aspects

Instead of (or addition to) the spatial aspects, we can work with temporal aspects. Placement of sonic objects in time is an underlying feature in the development of musical structures, so we need to find solutions for identifying, representing and utilizing temporal features from MoCap data. Here it can help to think about a three-level model of temporality: *sub-chunk*, *chunk* and *supra-chunk* [10]. Here the chunk level represents a time span of approximately 1-5 seconds, a time span which fits well with our working memory. The chunk level also (not coincidentally) happen to cover the time span of human actions, speech and music phrasing. In this model of time, the sub-chunk level is related to short sensations, while the supra-chunk level can be thought of as made up of a series of chunks. If we think about the continuous stream of MoCap data as the sub-chunk level, then segmentation of this stream into action segments that fall within the range of 1-5 seconds would correspond to the chunk level.

3.6 Pattern Recognition

As mentioned above, pattern recognition techniques have been used for mapping motion to sound [2, 21]. The typical goal here would be to recognize various types of expressive features from body movement and map these to relevant sounds. Here the dimensionality of the feature space is important for the robustness of recognition rates [8]. For example using 30 3D marker streams directly as features to the classifier can be problematic. This can be solved by reducing the dimensionality in the spatial and/or temporal domains, as mentioned above. Also, standard dimensionality reduction techniques from the field of pattern recognition can be used to find the features that work best.

An important conceptual question is how pattern recognition algorithms can support our goals. Using pattern recognition can certainly give us more options for the mapping to musical features, but how can it be used in an interesting way? We believe it is important that the final artistic results should be something new that we cannot do with traditional techniques. Simple one-to-one mappings, and trigger based systems would not do justice to the richness and complexity afforded by the IrMoCap system. The artistic result can end up just being a demonstration of technology with (hopefully) more than 90% correct recognition rate. An added challenge is that we are not good at reproducing our action precisely [19].

4. MAPPING MOTION TO SOUND

After evaluating some of the challenges when it comes to retrieving, processing and exploring data from an IrMoCap system in the previous section, we will here look at some of the challenges when it comes to mapping such data to sound features. This is a broad field and we will only touch on some of its complexity.

4.1 Sound-producing actions

Looking at the sound-producing actions used when performing a musical instrument, they can typically be divided into two groups: *excitation* and *modification* actions [15]. We can further distinguish between two types of excitations: *discrete* (e.g. triggers) or *continuous* excitation (e.g. bowing).

The raw data from an IrMoCap system is a continuous stream of numbers, so if we want to trigger signals we need to identify discrete actions through segmentation. The question, then, is whether using such a system for trigging predefined sounds is particularly interesting, or whether we might be better off by using an extra controller with simple buttons. This touches some of the challenges when it comes to designing connections between motion and musical features; to be effective the mapping should somehow match our mental model of what we want to control [22]. At the same time, several studies have shown that users find more complex and composite mappings more musically challenging and interesting [12, 16].

4.2 Touchless Actions

We can define *touchless action* as an action 'in the air' and where we cannot use the haptic and tactile response of a normal physical controller to guide us. In a musical context this implies a virtual relationship between sound and action since the relationship between the two is not bound by physical laws like we find in acoustic instruments [14].

When designing control interfaces for normal desktop computers, the design goals are rather straight forward. The interfaces should be ergonomic and effective, properties which are relatively easy to measure. Musical interfaces, on the other hand, have the extra requirement of being artistically interesting to use, a quality which is hard to evaluate and determine [16]. One design aspect which is especially important for virtual instruments is how the instrument's functionality can be understood mentally [22]. If the instrument is virtual, our whole comprehension of the instrument must either come from the sonic feedback or from our bodily experience of using the instrument. It seems plausible that the understanding of the connection between action and sound is a crucial point for the playability of a virtual instrument, but equally so for the audience watching the performance [6].

If we want to use touchless action as the basis for controlling musical features, it may be relevant to consider to

what degree we are conscious about our own body and its motion. If we use physical properties of tracked motion we need to take into account how these properties are understood by the users. For example so called *naive physics*, the untrained human perception of basic physical phenomena, can differ from what the data tells us [13]. Therefore, when using features like acceleration it is not certain that the user's understanding of these features reflects the numerical values.

A question connected to the potential of using touchless actions as control data is how many dimensions our actions consist of. Or maybe more important, how many dimensions are we able to exploit as control data? It may be appropriate to study the *informational theoretical* content. What is the needed sampling rate and how many bits per second are our touchless actions able to communicate?

Several groups of people are trained in touchless action. Dancers are experts in doing technically difficult actions, hearing impaired are experts in sign language and all of us use body language in our everyday life. To be able to exploit touchless action in a musical setting is certainly an interesting idea. But probably new paradigms are needed to map these actions to meaningful musical features. Until then it may be a good idea to design virtual instruments by mimicking aspects of our physical world so that we can take advantage of our established ecological experience of living in the world [19, 13].

4.3 Mapping to Sound Features

Let us briefly look at some possibilities when it comes to translating various types of motion and action features to sonic and musical features. A simple example is to map absolute marker position to the pitch of a sound. This may seem like a trivial task, but involves many different possibilities: should it be continuous control of pitch or in steps? How does pitch space relate to physical space? What types of pitch resolution and scales should be used? Instead of using absolute marker position to control sound features, it is also possible to look at the relative distance or angular position between two or more markers. These and many other similar questions will be the subject of some of our systematic studies of relationships between motion and sound in the coming years.

4.4 Spatialization

Another approach we are going to investigate in future studies include that of *spatialization*, i.e. placement of sound in space. The addition of a 32 channel speaker system in our motion capture lab provides the opportunity to explore control of sound through position and motion of the body in space. This may include moving sound sources around in the space, but also studying more complex relationships between physical and sonic space.

One approach to start such exploration may be to start by randomly setting up mappings between motion and sound features, much in the same way as the video to sound sonification suggested by Pelletier [20]. Instead of using optical flow we can let the marker displacement be sonified with additive or granular synthesis, something which may hopefully result in a rich combined motion and sound experience. Here marker occlusion and noise will also not be so problematic as long as a high percentage of the markers is properly tracked.

5. CONCLUSION

Infrared optical marker based motion capture technology is currently the state of art of motion capture systems, and despite some limitations, we believe such systems may provide for interesting and inspirational exploration of what other motion capture technologies can be used for. This paper has provided a review of some related work, and has covered some of the challenges related to using such systems in musical interaction. Much research still remains to make good musical use of such technologies. Here we believe it is reasonable to start by mimicking the already known physical world.

6. REFERENCES

[1] http://en.wikipedia.org/wiki/motion_capture.

[2] F. Bevilacqua, J. Ridenou, and D. Cuccia. 3D motion capture data: motion analysis and mapping to music. In *SIMS*, 2002.

[3] A. Camurri. Toward real-time multimodal processing: Eyesweb 4.0. In *Proc. AISB*, 2004.

[4] A. Camurri et al. Multimodal analysis of expressive gesture in music and dance performances. *LNCS*, 2004.

[5] C. Dobrian and F. Bevilacqua. Gestural control of music: using the vicon 8 motion capture system. In *NIME*, 2003.

[6] C. Dobrian and D. Koppelman. The 'e' in nime: musical expression with new computer interfaces. In *NIME*, 2006.

[7] M. Downie. *Choreographing the Extended Agent: performance graphics for dance theater*. PhD thesis, Massachusetts Institute of Technology, 2005.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[9] G. Eckel and D. Pirr. On artistic research in the context of the project embodied generative music. In *ICMC*, 2009.

[10] R. I. Godøy. *Systematic and Comparative Musicology: Concepts, Methods, Findings.*, chapter Reflections on chunking in music., pages 117–132. Peter Lang, 2008.

[11] F. Grond, T. Hermann, V. Verfaille, and M. Wanderley. Methods for effective ancillary gesture sonification of clarinetists. In *LNCS*, 2009.

[12] A. Hunt, M. M. Wanderley, and M. Paradis. The importance of parameter mapping in electronic instrument design. 2002.

[13] R. J. Jacob, A. Girouard, L. M. Hirshfield, M. S. Horn, O. Shaer, E. T. Solovey, and J. Zigelbaum. Reality-based interaction: a framework for post-wimp interfaces. In *CHI*, pages 201–210, 2008.

[14] A. R. Jensenius. *ACTION - SOUND, Developing Methods and Tools to Study Music-Related Body Movement*. PhD thesis, University of Oslo, 2007.

[15] A. R. Jensenius, M. M. Wanderley, R. I. Godøy, and M. Leman. Musical gestures: concepts and methods in research. In R. I. Godøy and M. Leman, editors, *Musical Gestures: Sound, Movement, and Meaning*, pages 12–35. Routledge, New York, 2010.

[16] S. Jordà. Digital instruments and players: Part i - efficiency and apprenticeship. *NIME*, 2004.

[17] A. Kapur and G. Tzanetakis. A framework for sonification of vicon motion capture data. In *Proc. DAFX05*, 2005.

[18] O. Larkin, T. Koerselman, B. Ong, and K. Ng. Sonification of bowing features for string instrument training. In *ICAD*, 2008.

[19] A. G. Mulder, S. S. Fels, and K. Mase. Design of virtual 3D instruments for musical interaction. In *GI*, 1999.

[20] J.-M. Pelletier. Sonified motion flow fields as a means of musical expression. In *Proc. NIME*, 2008.

[21] G. Qian, F. Guo, T. Ingalls, L. Olson, J. James, and T. Rikakis. A gesture-driven multimodal interactive dance system. In *ICME*, 2004.

[22] S. Skogstad. Models for the design of interfaces for musical expression. *To be submitted*.

[23] V. Verfaille, O. Quek, and M. M. Wanderley. Sonification of musicians' ancillary gestures. In *Proc. ICAD*, 2006.

[24] K. Vogt, D. Pirr, I. Kobenz, R. Hlldrich, and G. Eckel. Physiosonic - movement sonification as auditory feedback. Copenhagen, Denmark, 2009.

[25] K. Woolford. Will.0.w1sp - installation overview. In *ACM Multimedia (MM'07)*, 2007.