Recognition, Analysis and Performance with Expressive Conducting Gestures

Paul Kolesnik and Marcelo Wanderley
Department of Music Technology, Faculty of Music, McGill University
pkoles@music.mcgill.ca
mwanderley@music.mcgill.ca

Abstract

Although a number of conducting gesture analysis and following systems have been developed over the years, most of the projects either primarily concentrated on tracking tempo and amplitude indicating gestures while not taking expressive gestures into account, or implemented individual mapping techniques for expressive gestures that varied from research to research. There is a clear need for a uniform process that could be applied toward analysis of both indicative and expressive gestures. The conducting gesture recognition system is implemented on the basis of Hidden Markov Model (HMM) process. An external HMM object is developed for Max/MSP software. Training and recognition procedures are applied toward both right hand beat- and amplitude- indicative gestures, and left hand expressive gestures. Continuous recognition of right-hand gestures is incorporated into a real-time gesture analysis and performance system in Max/MSP/Jitter environment.

1 Introduction

Conducting can be described as a way of controlling performance of multiple instruments with ones physical gestures but without direct contact with the instruments themselves. In a conductor-musician interactive environment, visual information perceived by musicians serves as the means of conveying directions for musical gestures that are created by conductors expressive physical gestures. First successful attempts to analyze conducting gestures with the help of a computer were made as early as 1980 with *A Microcomputer-based Conducting System* (Buxton et al. 1980) that was based on previous research in music synthesis carried out by Matthews in *Groove* project and *The Conductor Program* (Matthews 1976). Following Buxtons research, a number of conducting recognition systems have been developed (Bertini and Carosi 1992), (Borchers, Samminger, and Muhlhauser 2002), (Garnett et al. 2001), (Haflich and Burns 1983), (Ilmomene and Takala 1999), (Keane and Gross 1989), (Lee, Garnett, and Wessel 1992), (Marrin 2000), (Marrin and Paradiso 1997), (Matthews 1991), (Murphy, Andersen, and Jensen 2003), (Segen, Mujumder, and Gluckman 2000), (Usa and Mochida 1998). Those systems experimented with a number of different approaches towards beat tracking and expressive gesture analysis which advanced the field of conducting gesture recognition. All of the designed systems used MIDI prerecorded scores until 2002, when the first audio-based system called *Personal Orchestra* was created (Borchers, Samminger, and Muhlhauser 2002). Another significant advancement that took place over the years was a transfer from 2-dimensional to 3-dimensional positional analysis of gestures (Tobey and Fujinaga 1996).

Whereas the traditional school of orchestral conducting technique has developed a well-defined grammar of basic conducting gestures that can be used to set the required vocabulary for the recognition system, design of identification and recognition procedures for expressive gestures has been one of the main issues in the field of computer-based conducting gesture recognition. The reason a similar temporal segmentation technique that is used for beat and amplitude indicating gestures cannot be applied to expressive gestures is that most expressive gestures do not contain clearly defined temporal transition points indicated by their positional boundaries and largely vary in terms of their form.

2 System Design

The described system addresses the issue through implementation of a recognition and classification process based on Hidden Markov Model. This statistical observation sequence analysis process, widely known for its use in speech recognition (eg. (Deller, Hansen, and Proakis 2000)), has been also used in score following (Orio and Dechelle 2001) and sign language gesture recognition (Vogler and Metaxas 1999) systems, and has been successfully applied to right-hand beat

conducting recognition in *Multi-Modal Conducting Simulator* project (Usa and Mochida 1998).

2.1 HMM object

As the initial step of the project, an external HMM object was implemented for Max environment. The object was written as a representation of a discrete HMM model and served as an implementation of its three principal features—learning, finding an optimal sequence of states and recognition. Sizes of state and label vectors as well as the type of HMM (left-to-right or ergodic) were passed as arguments to the object. Observation sequence recognition was implemented with a forward-backward algorithm, calculation of the optimal sequence of model states used Viterbi algorithm, and model training was done through Baum-Welch reestimation procedure. Logarithm scaling techniques were used in order to avoid computational range errors which may occur for longer observation streams due to a recursive nature of the processes. A detailed overview of general HMM techniques can be found in (Rabiner and Huang 1986) and (Rabiner 1989), whereas scaling procedure and other practical issues are described in (Lien 1998) and (Huang, Ariki, and Jack 1990). The HMM object also provided features for storing, viewing, importing/exporting and editing of HMM models.

2.2 Symbol Recognition

Symbol recognition was the initial system developed with the external HMM object. Absolute 2-D positional coordinates extracted from the movement of an input source (mouse or Wacom tablet) were used to calculate orientation values along the horizontal axis. Resulting data stream was then passed to a vector quantization external object written in Max that mapped observation stream to specifications of the codebook used by the HMM models. Each of the HMM objects represented an isolated symbol to be recognized. At the learning stage, HMM objects were individually trained with a number of symbol examples. At the recognition stage, an observation stream representing a symbol was passed to all of the HMM objects, and the one producing the highest probability was considered as the recognized symbol.

At the initial stage, English alphabet symbols were successfully used for training and recognition by the system. Furthermore, HMM recognition procedure performed equally well with examples of words—for example, it was able to differentiate between “who” and “why”, in spite of the fact that those two words share first two letters. In both cases of letter and word recognition, recognition rate was over 92.5%. Symbols that were incorrectly identified were the ones that shared many similar characteristics—for instance, in several cases a capital *C* was mistaken for a capital *G* and vice versa.

2.3 Gesture Recognition

At the following stage, the procedure was modified in order to be able to accept users movements as its input information. Inexpensive Logitech USB cameras were used for image input together with a colour glove worn by user for gesture tracking. However, the system was built to be compatible with higher precision 6-DOF positional trackers for use in further research. Image acquisition and processing was handled by Eyesweb freeware (Camurri et al. 2000) using blob colour tracking techniques. An Eyesweb patch extracted positional coordinates of colour glove and sent their values to Max/MSP software via OSC network.

As the initial step of the experiment, the procedure used by the symbol recognition system was replicated using a single webcam to capture a 2-D positional user input. Resulting recognition rates were similar to those obtained in the previous experiment with a mouse/tablet. The system was then extended to accommodate the 3-D nature of expressive conducting gestures. A second USB camera that was placed in profile view of the user (right or left, corresponding to the tracked hand) was used to capture additional positional information. The Max patch was modified so that there were two separate channels, front and profile, responsible for gesture recognition. Each of the channels contained an equal number of HMM objects, corresponding to the number of gestures the system was intended to recognize. At the recognition stage, probabilities of the corresponding pairs of objects were combined to obtain a final resulting probability that determined the choice of identified gesture.

Figure 1: Front and profile camera view of the user training the recognition system with left-hand expressive gestures

All of the conducting gestures used for positional recordings were performed by a doctoral conducting student at the Music Faculty of McGill University. Five left-hand expressive gestures were selected to be recognized. For right-hand beat indicating gestures, two beat patterns were chosen—a four-beat legato and a four-beat staccato patterns. A separate HMM object was used to represent each beat gesture of the two patterns. There were 20 training sets and 10 testing

sets for each gesture, and the system performed with a 98% recognition rate.

2.4 Gesture Analysis, Recognition and Performance

Continuous gesture recognition of right-hand beat identification gestures was incorporated into a gesture analysis and performance system, which was designed as a tool for conducting an audio (and optionally, a video) score of orchestral performance in real-time. The primary purpose of the gesture analysis section of the system was to extract beat amplitude and beat transition points from conducting gestures based on maxima and minima of their absolute positional values. Subsequently, gesture performance part of the system mapped the identified beat transition points and beat amplitude values to modifications in playback speed and volume of the audio score that was being conducted by the user. Real-time audio stretching/compression techniques that were used were similar to those introduced in *Personal Orchestra* project (Borchers, Samminger, and Muhlhauser 2002) and used in *Computer Vision* system (Murphy, Andersen, and Jensen 2003). As an optional feature, a prerecorded video of McGill Symphony orchestra performance was used as an output feature simultaneously with the corresponding audio score. Jitter objects received tempo modification information from Max/MSP objects and adjusted the playback of prerecorded video score.

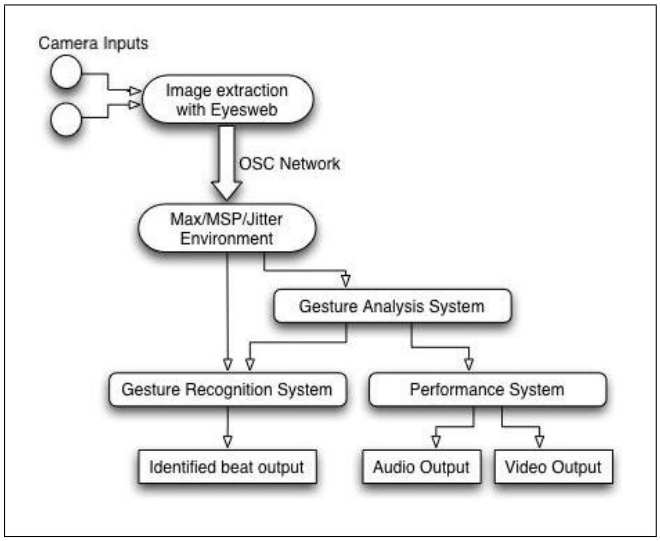

Figure 2: Schematic representation of the system

The same right-hand beat indicating gestures that were used for isolated gesture recognition were applied in the identification stage of a continuous recognition process. As a part of a continuous gesture recognition process, temporal segmentation of beats was done through information received from gesture analysis section of the system. The system was able to correctly identify the conducting gestures in real-time with a 94.6% recognition rate.

3 Future Works

One of the future goals of the project is to design a gesture recognition process that can be implemented in a continuous conducting movement environment in combination with the developed gesture analysis and performance system. Whereas the issue of temporal segmentation of continuous gesture observation stream can be easily solved for right-hand beat indicating gestures through the use of information extracted by another process (such as tracking positional maxima and minima of the gestures), there is no simple way of using a similar technique for expressive gestures, since there is no clear uniform indication of positional transitions between them. A solution to this problem lies in the capability of HMM process to automatically segment an entire observation stream into isolated gesture states. This technique (Vogler and Metaxas 1999) involves training HMM models separately with isolated gestures, and then chaining the trained models together into a single network of states. Viterbi algorithm can then be used on the entire observation stream, so that the temporal segmentation problem is simplified to computing the most probable path through the network of states produced by the data stream.

Upon completion of the continuous process of gesture recognition, the eventual goal of the work will be to develop a classification library of conductors gestures for computer conducting gesture recognition systems. This part of the project will address the need for development of a uniform set of conducting gesture definitions in terms of their positional information and mappings to the music score. The proposed library will be based on the existing well-developed grammar of traditional conducting technique, and will be introduced as a standardized set of gesture definitions to be used for future research in the field of conducting gesture recognition. Positional 3-D recording of the library gestures will be done with Vicon Motion Capture and Polhemus Liberty systems soon to be available at McGill Music Technology Labs.

4 Conclusion

The main achievement of the work is development of an HMM-based procedure that can be applied to analysis and classification of expressive conducting gestures. In particular, HMM training and recognition processes was applied to anal-

ysis of both right hand beat indicator gestures and left hand expressive articulation gestures. This brings an improvement over existing systems, since whereas right hand movements had been analyzed with HMM (Usa and Mochida 1998) and Artificial Neural Net (Ilmomene and Takala 1999) (Garnett et al. 2001) techniques in the past, there has been no previous research involving high-level recognition and classification techniques applied to left hand gestures. The designed HMM object, which is available for free distribution, is intended for use as a general tool in Max/MSP environment. It could be applied not only towards positional data classification but also towards any other process that involves pattern recognition—such as speech recognition, timbre recognition or score following. The resulting set of analysis, HMM-based recognition and performance tools will be directed towards future research in development of standardized classification of conducting gestures.

References

Bertini, G. and P. Carosi (1992). Light baton: A system for conducting computer music performance. In *Proceedings of the International Computer Music Conference*, pp. 73–76. International Computer Music Association.

Borchers, J., W. Samminger, and M. Muhlhauser (2002). Engineering a realistic real-time conducting system for the audio/video rendering of a real orchestra. In *Proceedings of the 4th International Symposium on Multimedia Software Engineering*, pp. 352–362. International Computer Music Association.

Buxton, W., W. Reeves, G. Fedorkov, K. C. Smith, and R. Baecker (1980). A microprocessor-based conducting system. *Computer Music Journal 4*(1), 8–21.

Camurri, A., P. Coletta, M. Peri, M. Ricchetti, A. Ricci, R. Trocca, and G. Volpe (2000). A real-time platform for interactive performance. In *Proceedings of the International Computer Music Conference*, pp. ???–??? International Computer Music Association.

Deller, J. R., J. H. Hansen, and J. G. Proakis (2000). *Discrete-time Processing of Speech Signals*. New York: IEEE Press.

Garnett, G. E., M. Jonnalagadda, I. Elezovic, T. Johnson, and K. Small (2001). Technological advances for conducting a virtual ensemble. In *Proceedings of the International Computer Music Conference*, pp. 167–169. International Computer Music Association.

Haflich, F. and M. Burns (1983). Following a conductor: The engineering of an input device. In *Proceedings of the International Computer Music Conference*. International Computer Music Association.

Huang, X. D., Y. Ariki, and M. Jack (1990). *Hidden Markov Models for Speech Recognition*. New York: Coumbia University Press.

Ilmomene, T. and T. Takala (1999). Conductor following with artificial neural networks. In *Proceedings of the International Computer Music Conference*, pp. 367–370. International Computer Music Association.

Keane, D. and P. Gross (1989). The midi baton. In *Proceedings of the International Computer Music Conference*, pp. 151–154. International Computer Music Association.

Lee, M., G. Garnett, and D. Wessel (1992). An adaptive conductor follower. In *Proceedings of the International Computer Music Conference*, pp. 454–455. International Computer Music Association.

Lien, J. J. (1998). *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity*. Ph. D. thesis, The Robotics Institute, Carnegie Mellon University.

Marrin, T. (2000). *Inside the Conductors Jacket: Analysis, Interpretation and Musical Synthesis of Expressive Gesture*. Ph. D. thesis, Massachusetts Institute of Technology.

Marrin, T. and J. Paradiso (1997). The digital baton: A versatile performance instrument. In *Proceedings of the International Computer Music Conference*, pp. 313–316. International Computer Music Association.

Matthews, M. V. (1976). The conductor program. In *Proceedings of the International Computer Music Conference*, Cambridge, Massachusetts.

Matthews, M. V. (1991). The radio baton and the conductor program, or: Pitch—the most important and least expressive part of music. *Computer Music Journal 15*(4), 37–46.

Murphy, D., T. H. Andersen, and K. Jensen (2003). Conducting audio files via computer vision. In *Proceedings of the 2003 International Gesture Workshop*, Genoa, Italy, pp. (in print).

Orio, N. and F. Dechelle (2001). Score following using spectral analysis and hidden markov models. In *Proceedings of the International Computer Music Conference*, pp. 125–129. International Computer Music Association.

Rabiner, L. R. (1989). The digital baton: A versatile performance instrument. *Proceedings of IEEE 77*(2), 257–285.

Rabiner, L. R. and B. H. Huang (1986). An introduction to hidden markov models. *IEEE Acoustics, Speech and Signal Processing Magazine 3*(1), 4–16.

Segen, J., A. Mujumder, and J. Gluckman (2000). Virtual dance and music conducted by a human conductor. *Eurographics 19*(3).

Tobey, F. and I. Fujinaga (1996). Extraction of conducting gestures in 3d space. In *Proceedings of the International Computer Music Conference*, pp. 305–307. International Computer Music Association.

Usa, S. and Y. Mochida (1998). A conducting recognition system on the model of musicians process. *Journal of Acoustical Society of Japan 19*(4), 275–287.

Vogler, C. and D. Metaxas (1999). Parallel hidden markov models for american sign language recognition. In *In Proceedings of the International Conference on Computer Vision*, pp. 116–122.