Context-Aware Gesture Recognition in Classical Music Conducting

Álvaro Sarasúa
Sonology Department, Escola Superior de Música de Catalunya
Music Technology Group, Universitat Pompeu Fabra
Padilla 155, Barcelona, Spain
alvaro.sarasua@upf.edu

ABSTRACT

Body movement has received increasing attention in music technology research during the last years. Some new musical interfaces make use of gestures to control music in a meaningful and intuitive way. A typical approach is to use the orchestra conducting paradigm, in which the computer that generates the music would be a *virtual orchestra* conducted by the user. However, although conductors' gestures are complex and their meaning can vary depending on the musical context, this context-dependency is still to explore. We propose a method to study context-dependency of body and facial gestures of conductors in orchestral classical music based on temporal clustering of gestures into actions, followed by an analysis of the evolution of audio features after action occurrences. For this, multi-modal data (audio, video, motion capture) will be recorded in real live concerts and rehearsals situations using unobtrusive techniques.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Interaction styles, Theory and methods; H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing

General Terms

Algorithms, Theory

Keywords

gesture, classical music, conducting, music information retrieval

1. INTRODUCTION

Body movement plays a key role on the performance and perception of music. We can think of the movement musicians have to make in order to create sounds from their instruments, the way in which dancers move following what they hear or just the different movements (tapping, headbanging...) that we make when listening to music. The interest for these movements has increased in the last years in music technology research and we can find several studies dealing with music-related *gestures*. One of the reasons for this growing interest on musical gestures is that a multimodal analysis of music that includes not only sound or score but also gesture-related information can reveal information about co-expressive elements that are present in the communication process of music [5].

Our work studies body movement in the orchestral classical music scenario relating it to expressivity on the resulting music, with a special focus on conductors' gestures. To do so, we will carry experiments in real live concerts situations where we will record multimodal data (multitrack audio, video, motion). From these recordings, we will study correlation between conductors' gestures and expressivity-related audio features. Moreover, we will study how similar gestures can get different meanings depending on the musical context.

This work is done in the scope of the PHENICX project, which aims at creating a methodological and technical framework for live classical music concerts in such a way that they become enriched multimodal, multi-perspective, multilayer digital artifacts to be explored, (re)enjoyed and shared in different ways. The output of this particular work will be used for an application that allows impersonation of the conductor in a complex and meaningful way. Also, the study of relations between conductors' gestures and musical expressive parameters will be useful for live visualizations of conductor-ensemble interaction within the performance.

This paper is organized as follows. In Section 2, we review and discuss works related to this topic. In Section 3, the proposed methodology is explained. Finally, in Section 4 future work is briefly discussed.

2. RELATED WORKS

Different approaches have been considered in the study of musical gestures. For instance, there are cases in which the sound-producing gestures of a particular instrument are obtained and analyzed in detail in order to synthesize their sound in a realistic way. Maestre et al. [17] use bowing parameters of violin performances to achieve realistic sample-based synthesis. In [10], Guaus et al. present a system to track left-hand fingering in guitar and study expressivity. A similar approach is taken by García et al. in [9], with the recorder as a case study and using pressure sensors.

There are some other approaches, however, that study different music-related gestures which are not just those in-

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502216 .

volved in sound generation as the ones above. Jensenius [11] carried a comprehensive study of *music-related movements* including observations of *air instrument* performances, *free dance* to music and *sound-tracing*, showing that people are able to associate different movements with sound features regardless of their musical expertise. Dahl et al. [6] made experiments with participants rating perceived emotions from mute video performances and showed that some basic emotions (happiness, sadness and anger) can be communicated via movements only whereas some others are not (fear). A similar experiment was done by Luck et al. [16]: the kinematics of conductors' expressive gestures (represented by features extracted from motion capture data) were related to the perceived expression (achieved by continuous ratings made by participants). Their results suggest some correlation between gesture features and perceived expressivity (e.g. more expressive = vertically extended and accelerated hand movement). Similarly, Camurri et al. [4] showed that emotions can be perceived in dancing movement following features as *quantity of motion*.

Works closer to the Human-Computer Interaction (HCI) field have looked for meaningful ways of controlling music performed by a computer using gestures. A typical intuitive approach is to use the orchestra conducting paradigm, in which the computer would be a *virtual orchestra* that the user conducts using gestures captured by different kinds of devices. There have been several works in this direction [2, 3, 13, 19] that either use video or special batons to allow interaction with virtual orchestras by usually modifying tempo and dynamics. Fabiani [7] used the same conductor paradigm going a step beyond and creating mappings between gestural data and semantic descriptors to let the listener control how music is performed. However, this mapping, which uses the valence-arousal model of emotions introduced by Russell [23] by for example assigning quantity of motion to valence and position of the hands to arousal, remains simple compared to the tasks performed by a real conductor and allows limited music interaction when it comes to more creative applications. He uses the KTH rule system [8] (a computational model for expressive performance) in two different approaches: (1) synthesizing sound from MIDI information and (2) modifying tempo and dynamics of real audio.

The gestures of the conductor and their relation to the music have also been studied from a computational perspective. Luck [14] has done a series of works on computational analysis of conductors' temporal gestures by extracting features from audio and motion-capture recordings. The main focus of these works is the relation of the gestures to the beat and the musicians' synchronization. From the results, communication of the beat seems to be related to acceleration along the trajectory and synchronization is clearly influenced by the synchronizer's previous experience. However, conducting manuals [22, 21] tend to focus more on the expressive aspects of conducting rather than on just time-beating.

Regardless of how interactive systems commented so far map gestures to sound features, this mapping remains the same throughout all the interaction. However, as stated in cited manuals and in works such as Poggi's [20], the same gesture can have different meanings and effects depending on the moment. For example, the director can frown to express aggressiveness (thus asking to play loud) or disgust (giving feedback to the musicians about something (s)he does not like). For a system to be able to gather different meanings from a same gesture depending on the musical context, it should be able to analyze music at different time scales in a way that allows it to know what part of the piece or the phrase is playing at the moment. Also, it should be able to get real-time information about at least tempo, loudness and articulation, which account for about 90% of the communication of emotions in expressive music performance [12]. The analysis of expressiveness-related audio features has been broadly studied for computational models of expressive performance. Widmer offers a good overview on the topic in [24]. Furthermore, and again according to the manuals and Poggi's work, facial gestures carry an important amount of expressive information while computational approaches to classical music conducting have not addressed them in a comprehensive way.

3. METHODOLOGY

In this Section, we explain how we propose to study conducting gestures with context dependency. Concretely, we explain how the data will be acquired and stored, how we will extract meaningful features from it and how we will relate gestural and audio information in order to extract conclusions. An initial dataset from a concert has been created in June 2013 including almost 2 hours of data.

3.1 Recordings

Most works including motion capture data have been done in controlled scenarios instead of live concert situations. The main reason for this is that motion capture systems used so far are somehow intrusive as they use sensors on the body, special batons, etc. The appearance of depth-sense cameras (popularized by Microsoft's Kinect[1]) allow unobtrusive 3-D gesture capturing with an easy setup. We will perform recordings at rehearsals and concerts in the High School of Music of Catalonia (ESMUC). Thanks to having access to the concerts and rehearsals of ESMUC, we will be able to record performances from different conductors (including students). The recordings of every performance will include:

- Multi-track audio. The number of tracks will depend on the performed piece, but there will be at least one track per section and, in the case of the concerts, one track from every microphone used for PA (stereo pair and overheads).
- Multi-perspective video. Concerts and rehearsals will be recorded with several cameras. A dedicated camera on the conductor will be used to identify basic facial gestures.
- Conductor's motion capture. Microsoft's Kinect will be used to record conductor's body movement. This device tracks 20 joints (points in the body) at 30 frames per second.

All these data will be collected in RepoVizz[2][18], an online database and visualizer of multimodal recordings developed by the Music Technology Group. Figure 1 shows a screenshot of RepoVizz's visualizer with data of a string quartet performance. RepoVizz automatically extracts features of audio files using Essentia[3].

[1]http://www.xbox.com/kinect
[2]http://repovizz.upf.edu/
[3]http://www.mtg.upf.edu/technologies/essentia

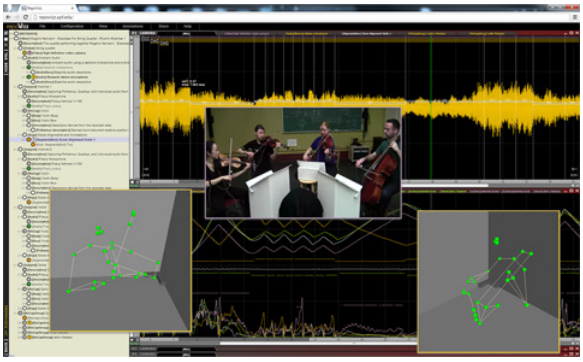

Figure 1: Screenshot of RepoVizz's visualizer, featuring some signals from computed audio descriptors, an audio file with annotations, video and motion capture display.

3.2 Gestures feature extraction

Previous work by Luck [15] on conductor's movement extracts features from gestures by deriving three-dimensional position, velocity and acceleration from markers at the body and the baton. Our motion capture data will allow to extract similar features from up to 9 points in the upper body (see Kinect's specifications).

From the conductor's video we will extract facial information using Kyle McDonald's MIT licensed libraries[4] for face tracking, including position of eyes, eye brows and mouth.

3.3 Gesture clustering

As explained in Section 1, one of our goals is to study how similar gestures can have different meanings depending on the musical context. The main problem here is to detect these "similar gestures" throughout an entire performance. Computational analysis of conductors' gestures so far do not consider *action units* as defined by Jensenius [11]: *goal-directed* movement excursions. In our approach, we will address this issue by looking for clusters of motion primitives from the conductor's gestures. Most common techniques for gesture clustering are based on Hidden Markov Models (HMMs) [1] and are suitable for real-time scenarios. Another approach we will consider for offline analysis is Hierarchical Aligned Cluster Analysis (HACA) [25], with which we can cluster time series of human motion in an unsupervised way. From time series of features explained in Section 3.2 we will look for points in time where similar gestures are being used.

3.4 Feature analysis

3.4.1 Relating gestures and audio

At this point, we will have two sets of data to relate audio with. On one hand, we will have direct gesture descriptors computed from the position of the body parts (upper body from Kinect and facial points from facial tracker). Cross-correlation between audio high level descriptors related to expressiveness and gestures descriptors will be computed and analysed in order to see how these gestures are actually influencing resulting audio. An analysis and correction of the lag (time difference between conductor's gestures and ensemble's response) similar to the one in [15] may be necessary. Figure 2 contains a block diagram that summarizes the overall process in this case. Features are extracted from audio and gestures and then correlated.

[4]https://github.com/kylemcdonald/ofxFaceTracker

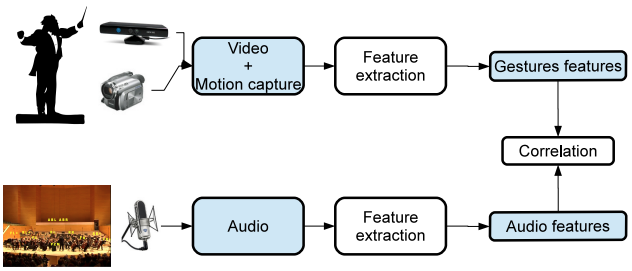

Figure 2: Block diagram of gestures-audio relation analysis. Features from audio and gestures are extracted and correlated.

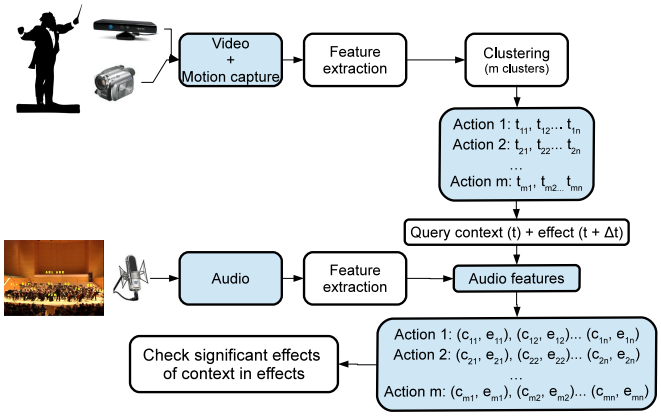

Figure 3: Block diagram of context-dependent analysis of gestures. Gestures are separated in m clusters corresponding to *actions*. Musical context and expressive change (effect) are represented by expressiveness-related audio features and their evolution in time after gesture occurrence.

3.4.2 Context-dependent analysis of gestures

On the other hand, we will have clusters of complex gestures and the different positions in time where they appear. Here, the approach will be slightly different and will include (a) identifying the context where the gesture occurs -part of the piece and features related to expressiveness-, (b) analysing the expressive changes in audio after the gesture and (c) studying similarity between context-effect pairs. We expect to find a significant effect of the context on the effect for a particular gesture, thus being able to define a set of rules of how an interactive system should react to that gesture depending on the context. The process now is more complex and implies more steps as showed in Figure 3. We start by extracting features from audio and gesture data. However, from the gesture data we build m clusters of *actions* that indicate the moments in time where different gestures occur. For a particular action and time t, the context will be defined by expressiveness-related audio features at the moment in which the gesture occurs. The effect will be defined by how these features evolve after some time Δt.

4. FUTURE WORK

The work so far has been focused on the definition of the presented framework and implementation of the gesture acquisition and storage software. In the near future, the plan is to work on the theoretical and practical definition of context for the orchestral music scenario. Also, using the initial dataset created this summer, experiments will be carried to evaluate different gesture feature extraction and clustering techniques that should, later, work in real-time scenarios. From the results of these experiments, changes in the acquisition procedure may be addressed in order to go towards a framework where gestures can be easily acquired, clustered and analyzed in relation to the resulting music.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7 / 2007- 2013 through the PHENICX project under grant agreement n^o 601166.

6. REFERENCES

[1] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. In *Gesture in embodied communication and human-computer interaction*, pages 73–84. Springer, 2010.

[2] J. Borchers, E. Lee, W. Samminger, and M. Mühlhäuser. Personal orchestra: A real-time audio/video system for interactive conducting. *Multimedia Systems*, 9(5):458–465, 2004.

[3] B. Bruegge, C. Teschner, P. Lachenmaier, E. Fenzl, D. Schmidt, and S. Bierbaum. Pinocchio: conducting a virtual symphony orchestra. In *Proceedings of the international conference on Advances in computer entertainment technology*, pages 294–295. ACM, 2007.

[4] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1):213–225, 2003.

[5] B. Caramiaux, F. Bevilacqua, and N. Schnell. Towards a gesture-sound cross-modal analysis. In *Gesture in Embodied Communication and Human-Computer Interaction*, pages 158–170. Springer, 2010.

[6] S. Dahl and A. Friberg. Expressiveness of musician's body movements in performances on marimba. In *Gesture-based communication in human-computer interaction*, pages 479–486. Springer, 2004.

[7] M. Fabiani. *Interactive computer-aided expressive music performance: Analysis, control, modification and synthesis*. PhD thesis, KTH, 2011.

[8] A. Friberg, R. Bresin, and J. Sundberg. Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology*, 2(2-3):145–161, 2006.

[9] F. García, L. Vinceslas, J. Tubau, and E. Maestre. Acquisition and study of blowing pressure profiles in recorder playing. In *Proceedings of the 2011 conference on New interfaces for musical expression*, 2011.

[10] E. Guaus, T. Ozaslan, E. Palacios, and J. L. Arcos. A left hand gesture caption system for guitar based on capacitive sensors. In *Proceedings of New interfaces for musical expression*, pages 238–243, 2010.

[11] A. R. Jensenius. *Action–Sound: Developing Methods and Tools to Study Music-Related Body Movement*. PhD thesis, University of Oslo, 2007.

[12] P. N. Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770, 2003.

[13] E. Lee, T. M. Nakra, and J. Borchers. You're the conductor: a realistic interactive conducting system for children. In *Proceedings of the 2004 conference on New interfaces for musical expression*, pages 68–73. National University of Singapore, 2004.

[14] G. Luck. Computational Analysis of Conductors' Temporal Gestures. *New Perspectives on Music and Gesture*, page 159, 2011.

[15] G. Luck and P. Toiviainen. Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis. *Music Perception*, 24(2):189–200, 2006.

[16] G. Luck, P. Toiviainen, and M. R. Thompson. Perception of Expression in Conductors' Gestures: A Continuous Response Study. *Music Perception*, 28(1):47–57, 2010.

[17] E. Maestre, A. Pérez, and R. Ramirez. Gesture sampling for instrumental sound synthesis: violin bowing as a case study. In *International Computer Music Conference*, 2010.

[18] O. Mayor, J. Llop, and E. Maestre. RepoVizz: A multimodal on-line database and browsing tool for music performance research. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011.

[19] L. Peng and D. Gerhard. A gestural interface for orchestral conducting education. In *Proceedings of 1st international conference on computer supported education (CSEDU), Lisboa, Portugal*, pages 406–409, 2009.

[20] I. Poggi. The lexicon of the conductor's face. *Advances in consciusness research*, 35:271–284, 2002.

[21] F. Prausnitz. *Score and podium: A complete guide to conducting*. WW Norton, 1983.

[22] M. Rudolf. *The grammar of conducting: a practical guide to baton technique and orchestral interpretation*. Schirmer Books, 1980.

[23] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[24] G. Widmer and W. Goebl. Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, 33(3):203–216, 2004.

[25] F. Zhou, F. De la Torre, and J. F. Cohn. Unsupervised discovery of facial events. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2574–2581, June 2010.