Towards soundpainting gesture recognition

Thomas Pellegrini
Université de Toulouse; IRIT
118 Route de Narbonne
31062 Toulouse, FRANCE
pellegri@irit.fr

Patrice Guyot
Université de Toulouse; IRIT
patrice.guyot@irit.fr

Baptiste Angles
Université de Toulouse; IRIT
baptiste.angles@gmail.com

Christophe Mollaret
Université de Toulouse; IRIT
christophe.mollaret@irit.fr

Christophe Mangou
christophe.mangou@
ensemble-amalgammes.fr

AM '14, October 01–03 2014, Aalborg, Denmark.

http://dx.doi.org/10.1145/2636879.2636899

Abstract

In this article, we describe our recent research activities on gesture recognition for soundpainting applications. Soundpainting is a multidisciplinary live composing sign language for musicians, actors, dancers, and visual artists. These gestures are produced by a soundpainter, which plays the role of a conductor, in order to lead a live performance. Soundpainting gestures are normalized and well defined, thus they are a very interesting case study in automatic gesture recognition. We describe a first gesture recognition system based on hidden Markov Models. We also report on the creation of a pilot corpus of soundpainting RGB/depth videos. The use of a computer could have many interesting applications listed in the paper. These applications are not limited to live performance, in which the computer would act as a performer. It could also help to investigate the balance between improvisation and planned creation in the particular context of soundpainting.

Keywords

Gesture recognition; soundpainting; Hidden Markov Models; music; interactive systems

ACM Classification Keywords

H.5.5 Sound and Music Computing

Introduction

In music perfomances, the balance between improvisation, as a spontaneous decision, and prepared material such as music scores has always been a source of reflexions, investigations, and creation. This exploration has produced, over the years, many different music styles and practices, from traditional and classical music to blues and free-jazz. Encyclopedia Britannica[1] defines music improvisation as "the extemporaneous composition or free performance of a musical passage, usually in a manner conforming to certain stylistic norms but unfettered by the prescriptive features of a specific musical text".

In our work, we are interested by an improvisation framework called *Soundpainting*, which was invented about fourty years ago and that is gaining momentum nowadays. It is a multidisciplinary live composing sign language for musicians, actors, dancers, and visual artists[2]. Presently, this *language* comprises more than 1000 gestures, although a hundred already allow to perform. These gestures are produced by a soundpainter, which plays the role of a conductor, in order to lead an improvisation. In a musical band or orchestra, the interaction between musicians is always a key point, which becomes crucial in improvised music [10]. Thus, the different roles and decisions to take in real time produce complex situations that are difficult to analyze and evaluate [8]. This remains true in the context of soundpainting, even if some supervision is brought by the soundpainter.

Computer-assisted improvisation systems have attracted a lot of attention over the last ten years [6, 13]. Examples are the framework provided by the Omax project, where an interactive system learns in real time from human performers [2], or the Continuator musical instrument that learns and interactively plays with a user in the user's style [11]. Recently, with the availability of RGB/Depth cameras at a low cost, gesture recognition/modeling real-time systems were developed with applications in music composition and improvisation. For instance, a gesture follower system was developed at IRCAM [4]. Nevertheless, to the best of our knowledge, no system was designed specifically for soundpainting applications.

In this paper, we report on a pilot corpus of soundpainting RGB/Depth videos, and on a first prototype of a gesture automatic recognition system in the context of soundpainting applied to music improvisation. We first briefly describe soundpainting and its main concepts. We give an overview on techniques used in gesture recognition and present our prototype. A proof of concept is reported along with the description of the pilot corpus created with a professional music conductor. The last section lists ideas of potential applications.

Bird's-eye view on soundpainting

Soundpainting is a sign language developed in the mid-1980s by the American Jazz composer Walter Thompson. Using a set of gestures, soundpainting allows the spontaneous conduction of an orchestra. The signs indicate the type of material desired by the conductor, known as the soundpainter, but it can also be used to conduct dancers, actors, poets, and visual artists. Soundpainting becomes more and more established as a code to be used in improvised performances, as shown by the emergence of dedicated orchestras and music festivals[3]. It gives an official and normed language to compose and improvise in real time, with the possibility to use previously prepared music as well as free improvisation. Beyond musical performance, soundpainting also provides tools for musical teaching.

In practice, the soundpainter realizes sequences of signs using hand and body gestures that can be thought of as gesture sentences that give instructions to the group of perform-

[1] Encyclopædia Britannica Online, http://www.britannica.com/EBchecked/topic, last visit: May 2014

[2] http://www.soundpainting.com, last visit: May 2014

[3] http://www.soundpaintingfestival.fr

ers. Basic gestures depicte musical concepts such as volume, tempo, pitch, and duration. More complex signals encompass notions of genre, style, key, memory, and more. The gestures are signed using the syntax of *Who, What, How* and *When*. They are grouped in two basic categories: *sculpting gestures* and *function signals*. Sculpting gestures indicate *What* type of material and *How* it should to be produced and function signals indicate Who will execute a set of instructions instructions and *When*. In our project, these different types of gestures must be recognized and semantically understood by an automatic system.

Gesture recognition

Gesture recognition opens new perspectives for natural human-machine interactions. Indeed, we constantly use gestures to communicate without requiring special attention or effort memory. It is therefore quite natural that they can be used to interact efficiently with a computer [5]. Applications of gesture recognition are manifold. Examples of the most common applications are video games, smart video surveillance systems, sign language, and virtual reality.

A broad review of the approaches used in automatic gesture recognition can be found in [9]. Since then, the arrival on the market a few years ago of inexpensive cameras that combine one or several RGB cameras with a depth camera (such as the Microsoft's Kinect device) allowed the real essort of 3D gesture recognition methods. Depth images are useful to ease the retrieval of the third dimension (depth) and unlike RGB cameras, they are not impacted by the color fluctuations induced by clothing, skin, or the environment. Machine learning techniques range from particle filtering to sequential generative models such as hidden Markov Models [9]. Recently, systems involving neural networks such as multilayer perceptrons or recurrent networks are investigated for their discriminative power. Deep learning is also a resurging paradigm at the core of actual studies in computer vision and other domains such as speech recognition [1].

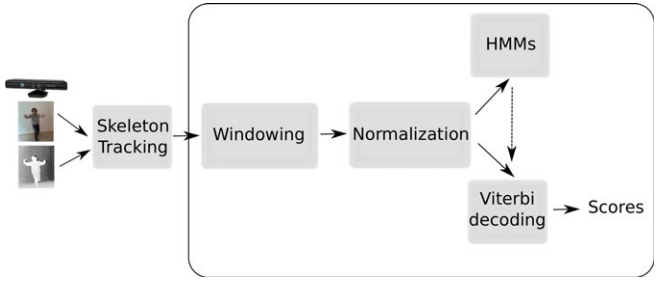

Figure 1: HMM-based prototype

HMM-based prototype

Figure 1 shows the architecture of our system. The soundpainter is filmed by a device that captures RGB and depth videos. The RGB and depth image frames are processed to retrieve the skeleton joints of the signer (we used the NITE and OpenNI libraries for this purpose). The frame-based skeleton absolute coordinates feed our gesture recognition software. Several sliding windows of different frame lengths are used in parallel. The analysis window sizes range from 5 to 10, along with windows of 20, 30 and 40 frames. The use of several frame lengths improved recognition performance as it allows to recognize a gesture performed at different paces. Then, the skeleton coordinates' frames from each sliding window are normalized. This normalization is a generalization to three dimensions of the normalization applied in the "$1" algorithm [12]. It consists of a resampling of the gesture to 32 coordinate vectors for each sliding window. Different kinds of spatio-temporal resampling methods were implemented and we use one that smoothly interpolates the joint trajectories. Follow a rotation according to the soundpainter shoulders' orientation and a translation to center the acquisition to a reference point. A symmetry to merge left- and right-handed gestures is then performed, and a final rescaling to a reference cube is done. We use the x-y-z spatial coordinates of 6

skeleton joints: the elbow, the wrist and the hand joints for both the left and right hands.

Our system is based on Hidden Markov Models, one per gesture. We chose this approach for the temporal sequence modeling capabilities of HMMs. Initially, we did some Matlab prototyping based on the HMM implementation from Dan Ellis[4]. Then, we implemented a Java version from scratch. Our final HMM approach is similar to the one reported in [3]. The main difference lies in the fact that we model each gesture by HMMs with a few states only, typically two or three states, instead of using one state per frame. Gesture recognition can be performed with Viterbi decoding or with the forward algorithm that gives slightly better results. We use a multivariate Gaussian distribution as observation probability density. We plan to implement Gaussian Mixture models and also non-parametric methods in a near future.

Proof of concept

Our objective is to adapt our system to soundpainting gesture recognition for music creation. We initiated a fruitful collaboration with a certified soundpainter, Christophe Mangou[5]. He is one of the directors of the reknown French classical orchestra *"Orchestre national du Capitole de Toulouse"*[6]. Besides his activity in classical music, he started a soundpainting project in 2005 named *Amalgammes*, which is a variable-size orchestra that combines the use of written scores and improvisation led by the soundpainting vocabulary[7].

[4] http://www.ee.columbia.edu/ dpwe/e4896/practicals.html#prac10
[5] http://www.christophemangou.com/en.html
[6] http://onct.toulouse.fr
[7] http://www.ensemble-amalgammes.fr
[8] http://gesture.chalearn.org/

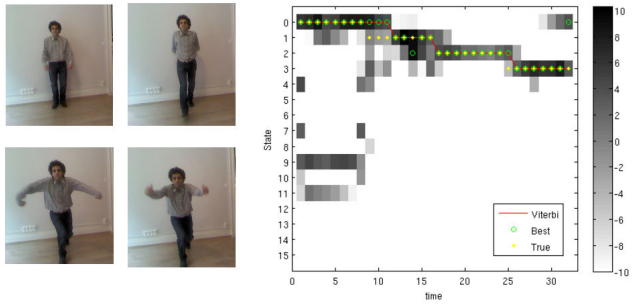

Figure 2: HMM states visited during a realization of a 'Play' gesture.

We recorded a pilot corpus with him, which consists of video recordings of Christophe Mangou performing about 20 gestures repeated five times each, such as 'Play', 'Whole Group', etc. We also recorded a set of soundpainting "sentences", which are sequences of gestures that could be used in a performance. Isolated gestures and gesture sequences present differences in their realization that reminds co-articulation in speech or in sign language.

Figure 2 shows an example of a realization of 'Play'. This gesture was modeled by three states, from state 0 to state 2. The right-hand side of the figure shows the visited states as found by a Viterbi decoding. As one can see, the three states are sequentially visited. We used this kind of analysis to determine the best number of states for each gesture. When a state is not visited, it is removed from the model.

We further enhanced our system by adding features derived from the tracking of the hands such as hand blob areas, hand barycenter coordinates and spatial moments. We are cur-

rently participating to a competition named ChaLearn that focuses on the recognition of a vocabulary of 20 Italian cultural/anthropological signs[8].

Future applications

Computer as a musician

A first direct application to soundpainter gesture recognition is the real time production of sound and music by a computer. In the case of gestures corresponding to *function signals* (who and when), the interpretation of these gestures is straightforward. For example, with a given sign or simply by pointing towards the camera, the soundpainter could designate the computer directly, indicating that a sound should be played by the machine.

The case of *sculpting gestures* (what and how) is a more complex question. Indeed, the use of a computer to produce or play sounds offers a wide variety of possibilities. These possibilities could be reduced to a single instrument for instance. In this case, a wide range of sculpting gestures such as long tones or notes in a pointillist style could be used. We can also imagine that the computer could play the role of several instruments. In this case, specific gestures should be used. The soundpainter could then ask for a drum or a double bass, but also for iconic sounds, recorded music or environmental sounds such as rain or broken glass.

The improvised aspects are an important point to take into consideration as it is a key point in soundpainting. It can be guided by random propositions, but could also depend of the context and previous material, as it is the case in the Omax project [2]. Moreover, the interaction between musicians is fundamental in a soundpainting performance. This raises very interesting and difficult questions when introducing a computer in the loop.

New tool for music composition

Beyond live performance, it would be interesting to use the computer in music composition led by soundpainting. Indeed, such a framework could be used to build new interfaces for sound design based on soundpainting gestures.

Computer-assisted soundpainting learning

Gesture recognition could be used by the soundpainter as a tool to learn or to improve his gestures. The sounpainters can be assisted in real time by a feedback given by the computer. Specific strategies should nevertheless be implemented in order to provide relevant and informative feedback.

Analysis of practice

As a recent musical practice, soundpainting aroused the interest of musicologists and social scientists. It provides a context where improvisation and live interaction can be analyzed. A first PhD Thesis on the subject was written by Marc Duby, who investigated how soundpainting operates as a system for collaborative creation of music in performative situations [7]. Moreover, soundpainting opens the way to a sociosemiotic analysis, in which it is susceptible to examination in the light of some theories of language. Lastly, practice of soundpainting raises questions, such as the evaluation of a live performance by an audience [8]. A soundpainter gesture recognition system could help in these studies by indexing performance video recordings automatically, and also by reproducing and creating new performances when the computer is used as a performer.

Conclusions

In this paper, we described a new application of gesture recognition to the context of soundpainting. Soundpainting gestures are a very interesting case study in automatic gesture recognition as showed by the encouraging experiments that we conducted with an HMM-based system. The intro-

duction of a computer could have many exciting applications that we listed in the paper. The computer could act as a performer, thus, very interesting and difficult questions would be raised by the fact that performers interact with the soundpainter but also between each other. These applications are not limited to live performance. A machine would also help to investigate the balance between improvisation and prepared music creation in the particular context of soundpainting.

Acknowledgment

This work was partly supported by a grant from the ANR (Agence Nationale de la Recherche) with reference ANR-12-CORD-003.

References

[1] Arel, I., Rose, D., and Karnowski, T. Deep Machine Learning – A New Frontier in Artificial Intelligence Research. *Computational Intelligence Magazine 5(4)* (2010), 13–18.

[2] Assayag, G., Bloch, G., Chemillier, M., Cont, A., and Dubnov, S. Omax brothers: a dynamic topology of agents for improvization learning. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia* (2006), ACM, pp. 125–132.

[3] Bevilacqua, F., Baschet, F., and Lemouton, S. The augmented string quartet: experiments and gesture following. *Journal of New Music Research 41*, 1 (2012), 103–119.

[4] Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., and Rasamimanana, N. Continuous realtime gesture following and recognition. In *Gesture in Embodied Communication and Human-Computer Interaction*, vol. 5934 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 73–84.

[5] Biswas, K., and Kumar Basu, S. Gesture Recognition using Microsoft Kinect. In *Proc. of ICARA* (Wellington, 2011), pp. 100–103.

[6] Blackwell, T. Swarm music: improvised music with multi-swarms. *Artificial Intelligence and the Simulation of Behaviour, University of Wales* (2003).

[7] Duby, M. *Soundpainting as a system for the collaborative creation of music in performance*. PhD thesis, University of Pretoria, 2006.

[8] Eisenberg, J., and Thompson, W. F. A matter of taste: Evaluating improvised music. *Creativity Research Journal 15*, 2-3 (2003), 287–296.

[9] Mitra, S., and Acharya, T. Gesture Recognition: A survey. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS 37:3* (2007), 311–324.

[10] Moran, N. *Measuring musical interaction: Analysing communication in embodied musical behaviour*. PhD thesis, The Open University, 2007.

[11] Pachet, F. The Continuator: Musical Interaction with Style. In *Proc. of ICMC* (Göteborg, 2002), pp. 211–218.

[12] Wobbrock, J., Wilson, A., and Li, Y. Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In *Proc. of UIST* (Newport, 2007), pp. 159–168.

[13] Yee-King, M. J. An automated music improviser using a genetic algorithm driven synthesis engine. In *Applications of Evolutionary Computing*. Springer, 2007, pp. 567–576.