Neuroklang : Real-time Timbre Control using Neural Networks

Visda Goudarzi

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University
{ visda}@ccrma.stanford.edu

Abstract. Neuroklang is a video-based interface for the sonification of hand gestures for real-time timbre control. Neuroklang is being used to make computer-based instruments with interactive control. It also allows the musicians to create sonorous and visual compositions in real time. The system consists of a laptop's camera, the filtering of camera input via the open source software known as Processing, a Neural Network for analyzing gestures, the sending of OSC control messages to the audio-processing program ChucK, and finally the parameter-mapping and sound synthesis enabled by ChucK.

Keywords: Neural Network, Image Processing, Gesture Recognition, Gesture Controller, Gesture-based Interaction, Real-time Performance, Video-based Interface, Musical Mapping, ChucK, Processing, Human Computer Interaction, Computer-Human Interaction, Sonification

1 Introduction

Interactive control over digital media has always been very appealing in performing art. Rapid evolution and accessibility of personal computers in the last decade has led to the use of the computer as the main interface for musical control in interactive media, with special emphasis on the real-time manipulation of digital video.

Controlling computer-based instruments through sensors can become very inexpressive by limiting the freedom of the human body to express various gestures. Therefore using digital camera to capture motion and applying image recognition techniques on the captured data is a good option.

The idea of Neuroklang originated from trying to make interactive performances with dancers and making sound according to their motions. an attempt to make computer-based interfaces for controlling music in a simple and sophisticated manner. Troika Dance Company has developed Isadora (the same metaphor of patch programming such as Max/MSP.) Although Neuroklang achieves the same goal, it uses available open source programming languages such as Processing and ChucK.

While Neuroklang tends to be simple and understandable for users, it also increases interactive feedback to introduce more sophisticated musical understanding for them. The first thoughts for developing this system came from musings about human interaction and communication, particularly a notion about a mute person talking to a visually-impaired person: how could that be possible without a translator? We would need a camera to capture the gestures of sign language that the mute person uses and a system to translate these gestures to words and/or sounds. This type of

work has been explored separately for persons who are visually-impaired and those who are mute, each in a separate and different context. A very good example is vOICe [1] - a location-finding, image sonification software for visually-impaired people. vOICe allows users to see with their ears and offers the experience of live-camera views through sophisticated image-to-sound rendering. The input comes from a digital camera mounted on a user's head.

The software scans the image with a vertical line and sonifies features in the image by representing vertical position as pitch, horizontal position as time (within the scan,) and brightness as volume. An example for speech recognition for vocally-impaired persons is Data-Glove [2]. It is a system for recognizing gestures of American Sign Language in real-time from video using Hidden Markov Models and Neural Networks. The network is mapped to a speech synthesizer's parameters in order to sonify the gestures. There are several versions of Data-Glove. Glove TalkII is one of them. It consists of three neural networks, one mapped to consonants, one to vowels and one to their relation to map hand gestures to speech. Additionally pitch and loudness are controlled to make the system more expressive.

All these gesture recognition systems are very interesting since they sonify discrete classified gestures and map them directly or indirectly to sound. A big issue which is not discussed in these systems is that the aesthetic aspects of a flowing analog motion is not taken into account. This is not an issue as long as the goal of a project is making scientific sounds or speech and not producing musically interesting data out of the gestures.

1.1 Overview of the System

Neuroklang consists of two main components:

Gesture and Image Processing: This part of the system consists of a laptop's video camera and an Open Source software Processing [3] to filter and calibrate data received from the camera. In the current prototype of Neuroklang, the input screen is divided into different sections, each represents a different instrument. In each section, the relative and absolute brightness and the amount of change compared to the previous frame in red, green and blue is measured. Furthermore, four different blobs - each detecting a different color (white, red, green and blue) - show up on the screen. By moving objects with the same color as a blob, color tracking those objects with blob tracking is possible, so there are four more possible parameters to map to sound.

The Neural Network part of the project is implemented on top of the project Wekinator [4]. In order to train the system, a couple of hand gestures are fed into the network and the incoming gestures get compared with those. Chuck and Processing communicate via OSC [5] messages sent from Processing to ChucK in order to manipulate sound and send the opposite direction to control the video output to make the instrument more expressive.
Data Processing and Sound Synthesis: ChucK[6] programs are used to manipulate data received from Processing to synthesize sound.

1.2 Interface Design and Architecture

Neuroklang is a work-in-progress and there is a lot more to be done to formulate expressive sounds from expressive gestures. Each section on the video frame is mapped to a different instrument. So far, the modules for four types of instruments are implemented. One is a drone-like sound that has already been tested. The second instrument is a randomly generated, particle-like sound. The timbre and reverb of this sound is manipulated with gestures. In the future, the density of these random sounds will be indirectly mapped to the density of motion in the image. The third instrument is a beat-detecting instrument tracking the beats in motion. The fourth instrument is a voice instrument. The voices are manipulated with a granular synthesizer, and grain parameters are mapped to blob motions received from the video. Therefore moving hands can manipulate the length and randomness of grains.

The second part of the interface manages the Neural Network. It recognizes four specific gestures and maps them to four subtle timbres which melt into the overall sound.

1.3 Recognition of Gestures using Neural Network

Sign languages are the main form of communication among the Deaf community. However these languages are not widely known outside of these communities, hence a communications barrier can exist between Deaf and hearing people. Even a bigger communication barrier exits between Deaf and visually impaired people. The hand tracking technologies enable the possibility of creating portable devices which can convert sign language to speech, as an approach to overcoming these difficulties. There are several studies on analyzing and sonifying sign languages. The goal in Neuroklang was not to sonify a specific sign language, but understanding the analysis of sign languages helped to define postures that could be useful for Neuroklang. According to linguistic analysis of sign languages, signs can be described in terms of four basic features; The hand shape defines the configuration of the joints of the hand. Orientation specifies the direction the hand and fingers are pointing, whilst the place of articulation is the location of the hand relative to the body. Motion, which consists of a change over time of any combination of the other three features. [7]
Based on basics of sign language recognition systems Neuroklang has been created with the capablity of classifying four different hand postures.

The predetermined input samples are open hand from front and back, and hand fist from front and back. 10 different pictures of each of these postures are pixelized and analyzed in Processing. The result array of numbers which represent the brightness of the pixels in each picture are being used as the data input into the Neural Network.

1.4 Training The Neural Network

Network design and training is the first step towards constructing a network for real-time mapping. For this, the user will need some predetermined input samples and their targets, and a defined architecture to start the training. An approach with trials and errors makes the network converge to the desired behavior. According to Freed and Wessel, such a choice of network architecture and parameters would save significant time in realizing the network. In Neuroklang and most musical applications, a maximum of one hidden layer suffice. [8] While use of excessive number of neurons makes the network converge more rapidly, it degrades the generalization in most applications. Therefore, choosing the number of layers and neurons is very critical.

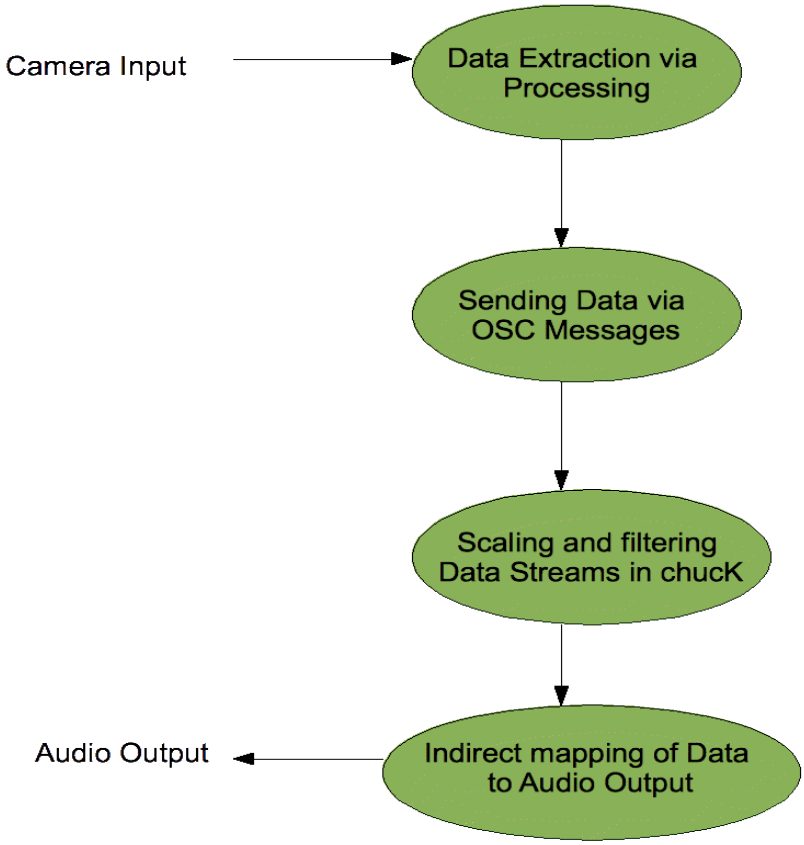

Fig. 1. Schematic picture of Data Flow in Neuroklang

2 Conclusion and Future Work

New music controllers bring new possibilities to making music; this project only explored a few of the new and different ways to do so. The next steps are to explore the expressivity and the degree of complexity possible when performing on each of the instruments. Iterations in development of the prototype with different types of instruments and accompanying tests are planned.

Neuroklang is capable of classifying four postures and sonifying them. The design of the system allows for future enhancements, both in terms of expanding its vocabulary, and improving the recognition accuracy. The vocabulary could be developed with better sonification strategies. The major area in which accuracy could be improved is in the classification of motion and dynamic gestures rather than only postures.

References

1. Kurze, M.: TDraw: a Computer-based Tactile Drawing Tool for Blind People. Proceedings of 2nd Annual ACM Conference on Assistive technologies. ACM Press. Canada (1996) 131- 138

2. Fels, S.S., Hinton, G.E.: Glove-Talk: A Neural Network Interface between a Data-glove and a Speech Synthesizer. IEEE Trans. On Neural Networks, Vol. 4, No. 1 (1993)

3."Processing" website: http://www.processing.org

4. Fiebrink, R., G. Wang, and P. R. Cook. "Support for MIR prototyping and real-time applications in the ChucK programming language." Proceedings of the International Conference on Music Information Retrieval (ISMIR), Philadelphia, September 14–18, 2008.

5. Wright, M., Freed, A.: Open SoundControl: A New Protocol for Communicating with Sound Synthesizers. ICMC. Thessaloniki (1997)

6. Wang, G., Cook, P.R.: ChucK: A DAFx, Concurrent, On-the-fly Audio Programming Language. Proceedings of the ICMC (2003)

7. T Johnston , Auslan: The Sign Language of the Australian Deaf Community, PhD thesis, Department of Linguistics, University of Sydney (1989)

8. Lee, M., Freed, A., Wessel, D. "Real-Time Neural Network Processing of Gestural and Acoustic Signals", Proceedings of the 17th International Computer Music Conference, Montreal(1991)