Head-Tracking for Gestural and Continuous Control of Parameterized Audio Effects

David Merrill
MIT Media Lab
20 Ames St. E15-313
Cambridge, MA 02139-4307
dmerrill@media.mit.edu

ABSTRACT

This paper describes a system which uses the output from head-tracking and gesture recognition software to drive a parameterized guitar effects synthesizer in real-time.

Keywords

Head-tracking, gestural control, continuous control, parameterized effects processor.

1. INTRODUCTION

Researchers and performers working with gestural control of music have a history of training digital video cameras on themselves and inventing interesting mappings for the output. A notable early example is David Rokeby's Very Nervous System [5]. This paper presents a working system that uses a modified real-time head-tracker to drive a parameterized guitar effects processor. A conceptual model for a similar system was proposed by Lyons [7], and he built a system that used a visual mouth-tracker to drive parameterized audio effects [8].

2. SYSTEM OVERVIEW

The system consists of a modified real-time head-tracker which communicates via a TCP/IP socket connection to a custom server program. The server is responsible for managing the mapping of sensed gesture onto appropriate control messages, which it sends to a guitar effects processor via MIDI messages.

The FaceSense program [1] runs under Linux, and uses an IR-sensitive camera based on the BlueEyes camera from IBM [4]. Pupil positions and sizes are tracked using a difference images, and eyes/eyebrows are tracked using templates. At a higher-level, detection of head nods, head shakes, and eye blinks is also implemented. The system runs at 29-30 frames-per-second on our IBM Netvista 1.5GHZ machine.

The mapping server is written in Java, and is responsible for handling incoming data from the FaceSense program and managing the state of the guitar effects processor. Modules implemented include running-average filters to smooth the incoming data (implemented efficiently with a ring buffer), an amplifier-selector-manager to handle the state machine used in the gestural interaction, a midi device manager to provide easy midi message transmission, and a general-purpose sensor-value-to-midi-message mapping class.

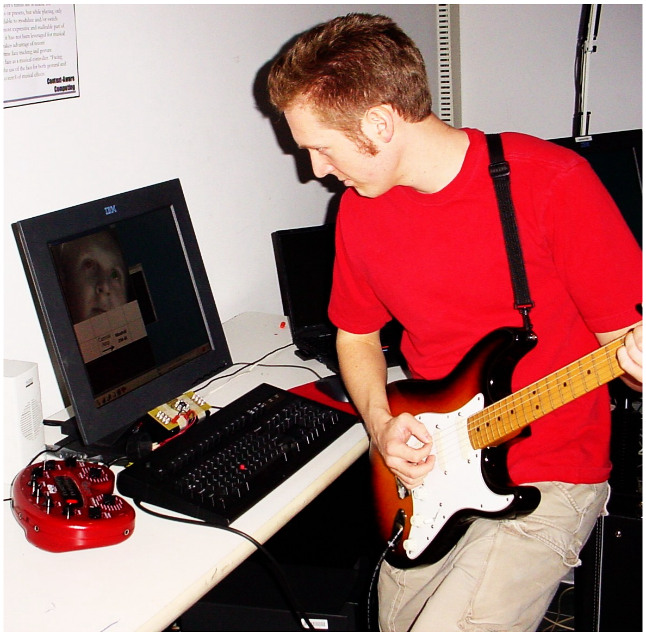

Figure 1: The author using the system. The IR camera is just beneath the monitor, and the effects processor is on the left.

The Line6 Pod 2.0 [6] proved to be a useful guitar effects processor for this project, since all settings are externally-controllable via MIDI messages (see figure 1).

3. MOTIVATIONS AND MAPPING

The dialogue between humans and machines is fundamentally social in nature [2], and it has been argued that the man-machine interface can be improved by leveraging human expectations of natural human social cues when designing technologies that interact with people. Such a cognitive "scaffolding" can engage the user's existing behaviors and expectations about interaction to enhance interaction with a computer system [3]. This work explores ways in which a camera-to-audio-mapping interface can respond to the performer socially by reacting in the following ways:

Reaction to personal space: People are acutely aware of how close the face of their conversational partner is to their own, and an excessively close partner can cause anxiety or excitement. This system monitors maps distance between the performer's head and the camera onto a continuously-varying parameter (wah and volume were tried), tracking the continuously-varying level of comfort associated with personal space intrusions.

Reaction to face orientation: Cognitive psychology has shown that faces and other shapes become less recognizable as they are rotated off-axes. This system maps the tilt of the performer's face onto a continuously-varying parameter (volume and distortion level were tried).

Gestural communication: Humans frequently communicate semantic content by gesture, especially musicians in an environment too loud for spoken language. Musicians often cue each other with a head nod, signaling a musical change. This system engages the performer in a gestural dialogue to switch between amplifier presets. The head nod/shake gestures which the performer uses to drive the interaction borrow the scaffolding of everyday human-human yes/no interactions (see figure 2).

Finally, the interface helps solve the "laptop musician problem" in which the performer-computer interaction inhibits performer-audience interaction. Even if the performer chooses to look at the computer screen during the interaction, the computer is also "looking back", and the camera's eye view of the performer can be projected for the audience to enjoy as well. Furthermore, a shoulder-mounted display would allow the performer to walk the stage and continue to interact directly with the audience.

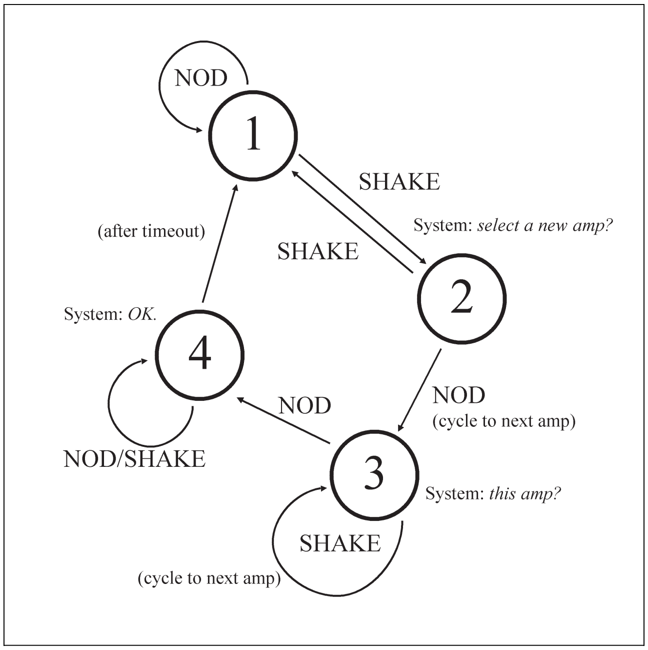

Figure 2: State-machine diagram for gestural interaction

4. EVALUATION

The system is usable and entertaining. It is also relatively user-independent, although the FaceSense head-tracker works best in a dark or dim environment free of specular light sources. The mapping of head-distance from the camera to a wah effect has been received particularly well. A problem with the head-tilt interaction is that it can be difficult for the performer to keep his face within the camera's field-of-view. The visual feedback of the performer's face on the computer screen makes this coordination possible with some practice, but a head or shoulder-mounted device [8] could improve on the usability of the system.

In addition, the gestural "yes/no" interaction, which currently requires the performer to be looking at the screen, could be improved by moving this feedback to another channel, such as audio or tactile.

5. CONCLUSIONS AND FUTURE WORK

This project represents an early step in using head-tracking software for both continuous and gestural mappings to musical output. It would be interesting to continue the current work by using a true face-tracker which could extract socially-meaningful representations of facial expressions (anger, fear, surprise, etc..). However, even with more meaningful feature-extraction, finding compelling mappings for the output of such a system will continue to be a challenge. As parameterized audio effects processors swell in sophistication, and the human input and acoustic output of these music-creation tools become decoupled to a greater extent, the problem of mapping becomes increasingly complicated. In time, machine-learning-based tools could be developed to supplement or perhaps even replace the human in the difficult mapping task. A system which collects either implicit or explicit feedback from the user could learn optimally pleasing mappings of the detected features.

6. ACKNOWLEDGMENTS

Thanks to Ashish Kapoor for his generous help with the FaceSense tracking software.

7. REFERENCES

[1] A. Kapoor. Real-Time, Fully Automatic Upper Facial Feature Tracking. Proceedings of The 5th International Conference on Automatic Face and Gesture Recognition 2002, Washington D.C., May 20-21, 2002.

[2] B. Reeves, C. Nass. The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places. Cambridge University Press/CSLI, New York, 1996.

[3] C. Breazeal. Designing Sociable Robots. MIT Press, 2000.

[4] C. Morimoto, D. Koons, A. Amir, and M. Flickner. Pupil detection and tracking using multiple light sources. Technical report, IBM Almaden Research Center, 1998.

[5] D. Rokeby. Personal website. http://www.interlog.com/~drokeby/

[6] Line 6 Pod v2.0, http://www.line6.com/

[7] M. Lyons and N Tetsutani. Facing the Music: A Facial Action Controlled Musical Interface. Proceedings, CHI 2001, Conference on Human Factors in Computing Systems, March 31 - April 5, Seattle, pp. 309-310.

[8] M. Lyons, M. Haehnel & N. Tetsutani. The Mouthesizer: A Facial Gesture Musical Interface. Conference Abstracts, Siggraph 2001, Los Angeles, p. 230.