Music Moves

Wim Looman, Richard Green
University of Canterbury

***Abstract*—This report**

I. Introduction

A. Goal

B. Background

A lot of different methods for producing movement from sound have been researched. Some of these involve using the hands to "shape" the music, for example having a 2d pointer (which could easily be generated from a hand tracking algorithm) that affects the music based on a set of simple gestures [1]. Others attempt to provide a more traditional musical interface, albeit without the actual instrument. This can be a simple interface such as a DJ's deck **[yetton]_**, or a more complex interface like a full piano **[godoy]_** or guitar **[pan]_**.

Whilst the first approach will probably lead to the most valuable research, it would also require a deep understanding of music theory. Therefore the second approach was the one chosen for investigation. One instrument that hasn't been widely studied for video based playing is the air drum. It would have to be wondered why; as the drum simultaneously provides both a complex interface, the 3d position of the imaginary sticks, and a simple interface, only around 3-10 different drums/cymbals to hit.

Similar to previous research in this area **[pan]_**, hand tracking was decided on as the simplest method of decoding the input video. This would then produce two sets of (x, y) coordinates per frame, one for each hand. These would be fed into a neural network, along with the positions for the previous set of frames. This would allow the neural network to detect the changes in hand position corresponding to the hitting of the imaginary drum set.

1) Hand Tracking:

2) Neural Network: To simplify the creation of the neural network it was decided that this should be a separate program. Writing the network in Erlang would be much easier than using C, and by having a separate program the training would be simpler to setup.

The neural network would have a large input layer, four nodes per frame of the video. To provide context it is assumed that at least a second of video will be required, at say 30 frames a second this means the input layer would be at least 120 nodes. The actual length of the context will be part of the optimisation of the neural network.

II. Results

III. Conclusion

IV. Docutils System Messages

system-message
ERROR/3 in `introduction.rst`, line 12
Unknown target name: "yetton".

system-message
ERROR/3 in `introduction.rst`, line 12
Unknown target name: "godoy".

system-message
ERROR/3 in `introduction.rst`, line 12
Unknown target name: "pan".

system-message
ERROR/3 in `introduction.rst`, line 29
Unknown target name: "pan".

References

[1] Jensen, "Jensen's article," *The journal*, 2678.