Video Based Recognition of Hand Gestures by Neural Networks for the Control of Sound and Music

Paul Modler
University of Media, Arts and Design
Department of Media Art/Sound
76135 Karlsruhe, Germany
pmodler@hfg-karlsruhe.de

Tony Myatt
University of York
Department of Music
YO 105 DD York, United Kongdom
am12@york.ac.uk

ABSTRACT
In recent years video based analysis of human motion gained increased interest, which for a large part is due to the ongoing rapid developments of computer and camera hardware, such as increased CPU power, fast and modular interfaces and high quality image digitisation. A similar important role plays the development of powerful approaches for the analysis of visual data from video sources. In computer music this development is reflected by a row of applications approaching the analysis of video and image data for gestural control of music and sound such as Eyesweb, Jitter, CV ([1,[2], [3]). Recognition and interpretation of hand movements is of great interest both in the areas of music and software engineering ([4], [5], [6]). In this demo an approach is presented for the control of music and sound parameters through hand gestures, which are recognised by an artificial neural network (ANN). The recognition network was trained with appearancebased features extracted from image sequences of a video camera.

1. A SET OF CYCLIC HANDGESTURES
Previous experiments showed that hand gestures may be combined as cyclic gestures such as waving the hand or pointing to the left and to the right with the index finger [7].

Gesture	Description	Short names of main states
Index up/down	Index finger moves up and down	indUp, indDo
Index left/right	Index finger moves left and right	indLe, indRi
Cut up/down	Flat hand moves up and down	cutUp, cutDo
Cut left/right	Flat hand moves left and right	cutLe, cutRi
Horizontal open/close	Hand with horizontal back opens and closes	horOp, horCl
Vertical open/close	Hand with vertical back opens and closes	verOp, verCl
Croco open/close	Hand with thumb opens and closes	corOp, corCl
Swing open/close	Hand turns and opens and turns and closes	swiOp, swiCl

Table 1: A set of cyclic gestures of the left hand

NIME08, June 5-7, 2008, Genova, Italy

For this, the motion of a gesture is grouped in at least two main states performed in a repetitive way. The whole gesture may then be seen as a progression through a cyclic state model with the aim to view the gesture not as an isolated event but in the gesture context and related motions.

2. VARIATION OF GESTURE INSTANCES
Each gesture was recorded at 3 lower positions and 2 upper positions of the gestural space of the hand and arm to obtain data reflecting the variance of the hand articulation at differing locations. All 5 recording instances were aimed to be in a plane parallel to the front of the camera. Blended hand positions for the static states of four cyclic gesture types are shown in Figure 1 to Figure 4.

Figure 1: Horizontal open/close

Figure 2: Vertical open/close

Figure 3: Croco open/close

Figure 4: Swing open/close

3. TIME DELAY NEURAL NETWORKS
Time Delay Neural Networks (TDDN) are feed-forward networks and incorporate the learning of time series through a series of data windows (delays) shifting in time over the data series. An exemplary TDNN would consist of 4x4 input units and 4 input delay frames. To apply such a TDNN to image features larger input frames were used i.e. 1024 or 256 input units and a hidden layer size of 50 units ([8].) The number of output units was in the range of 24 to 37 similar as in the shown

example. In our approaches each output unit was associated with a certain state of the training patterns for the network.

4. GESTURAL CONTROL OF SOUND

The system tries to realise the control of a sound generation process by using discrete bindings of gestures to sound parameters. The system uses gestures of the left hand which are recognised by the video analysis. Two identical sound generation processes for live sampling and sound modification are realised in a Max/Msp patch. The position space of the hand is divided through a dedicated object (Gitter) into 9 concentric fields (Figure 5).

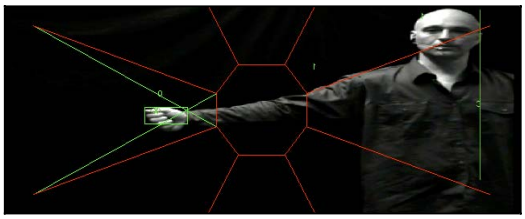

Figure 5: division of gesture plane into concentric fields

The binding of a parameter group to a *Gitter* field is aimed to locate the more important and more often-used parameters in the centre of the position space, and less-often used parameter groups around the central area.

Remote to body	**Centre**	**Close to body**
Effect distortion (low)	Diffusion/panning (mid)	Reverb (low)
Volume (low)	Selection of sound (mid)	Recording (mid)
Effect ring-modulation (low)	Filter (high)	Granulation (high)

Table 2: Binding of sound parameter groups

In Table 2 the estimated complexity and number of parameters is given in brackets. Each field in the gesture coordinate extends into a list of selectable choices. In the stage setup (Figure 7) these choices are mainly implemented as settings for the parameter category associated with the field, representing a morphing state of the soundproceesing patch (Figure 6).

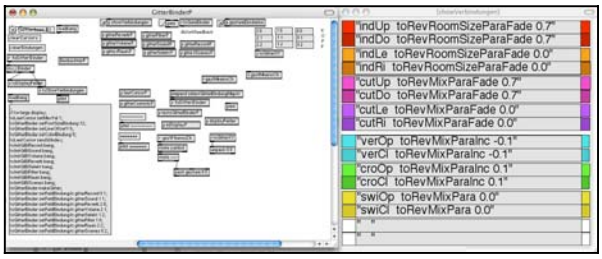

Figure 6: Display of sound actions (reverb) bound to the current field of the hand position

5. RESULTS

The demo system is a protoype for the usage of gesture recognition with artificial neural networks integrated in a sound generation context. The degree of required recognition precision varies between different performance paradigms. It may range from an aleatoric approach where it low recognition rate of 75% to 85% is sufficient to a strict binding of the gestures to a complex control of an elaborate instrument, where a single miss-recognition will disturb the whole musical concept or at least will be perceived as a hindering error. For both extremes, the aleatoric and the strict approach the binding of the body gestures to the musical actions have to be considered thoroughly.

A similar situation may be found for the required number of recognisable gestures, which differs between the musical intention and the role it assigns the gesture recognition. Two or three gestures may be enough to play a central role in a piece. For a complex control a larger number of gestures is required e.g. more than the 16 gesture states of the hand used for the training of the neural network of the demo system.

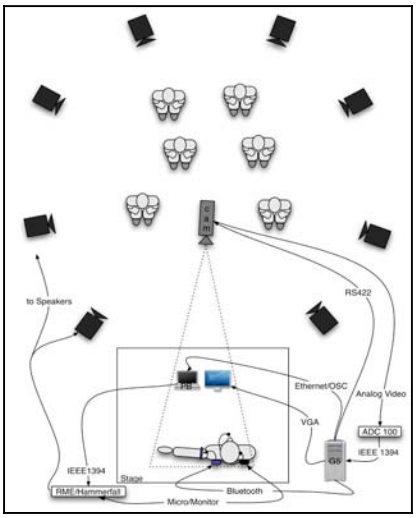

Figure 7: Stage setup

The goal in the applied setup was not to use as many as possible mappings provided from the gesture recognition or the visual analysis but to focus on the case, that visual musical control is achieved by an automatic recognition process. Gesture recognition may be only one part in the dramaturgy of a piece but it reflects the development of information technology, which approaches human mind and human body.

6. REFERENCES

[1] Camurri, A., Hashimoto, S., Ricchetti, M., Trocca, R., Suzuki, K., Volpe, G., EyesWeb toward gesture and affect recognition in interactive dance and music systems, Computer Music Journal, pp. 57-69, MIT Press, Spring 2000

[2] cv.jit, Computer vision for jitter by Pelletier, J. M., http://www.iamas.ac.jp/~jovan02/cv/ (2007)

[3] Cycling74,www.cycling74.com, (2008)

[4] Axel G. E. Mulder, S. Sidney Fels and Kenji Mase, Empty-handed Gesture Analysis in Max/FTS, AIMI international workshop on kansei the Genova, Italy, 1997.

[5] Cadoz C., Wanderley, M., Gesture-Music, in M. Wanderley and M. Battier (eds): Trends in Gestural Control of Music-Ircam - Centre Pompidou, 2000

[6] Mathias Kolsch, Matthew Turk, "Robust Hand Detection," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004

[7] Modler, P., Myatt, A., Saup, M., An experimental set of hand gestures for expressive control of musical parameters in realtime, Nime-2003, McGill University, Montreal, 2003

[8] Modler P, Myatt A., Image features based on 2-dimensional FFT for gesture analysis and recognition, Proceedings of the 4th Sound Music Computing Conf. Lefkada, 2007