An Experimental Set of Hand Gestures for Expressive Control of Musical Parameters in Realtime

Paul Modler
Hochschule fuer Gestaltung
Lorenz Str. 15
76135 Karlsruhe
pmodler@hfg-karlsruhe.de

Tony Myatt
Music Department
University of York
Heslington, York
tone@cage.york.ac.uk

Michael Saup
Hochschule fuer Gestaltung
Lorenz Str. 15
76135 Karlsruhe
michael@particles.de

ABSTRACT

This paper describes the implementation of Time Delay Neural Networks (TDNN) to recognize gestures from video images. Video sources are used because they are non-invasive and do not inhibit performer's physical movement or require specialist devices to be attached to the performer which experience has shown to be a significant problem that impacts musicians performance and can focus musical rehearsals and performances upon technical rather than musical concerns (Myatt 2003).
We describe a set of hand gestures learned by an artificial neural network to control musical parameters expressively in real time. The set is made up of different types of gestures in order to investigate:

- aspects of the recognition process
- expressive musical control
- schemes of parameter mapping
- generalization issues for an extended set for musical control

The learning procedure of the Neural Network is described which is based on variations by affine transformations of image sequences of the hand gestures.
The whole application including the gesture capturing is implemented in jMax to achieve real time conditions and easy integration into a musical environment to realize different mappings and routings of the control stream.
The system represents a practice-based research using actual music models like compositions and processes of composition which will follow the work described in the paper.

Keywords

Gesture Recognition, Artificial Neural Network, Expressive Control, Real-time Interaction

1. INTRODUCTION

Discussions relating to gestural data processing for musical applications have emerged in recent years. This discussion developed from the background of interactive computer music systems and their use in performance, but also from the experience of novel interfaces to control the generation of sound and musical processes. Several data processing paradigms have been established inspired by the availability of a larger range of sensor systems and the increasing processing power These all use gestural data to control musical parameters (or light etc.) within artistic environments.

Issues of mapping control data to musical parameters are related to these paradigms and the detection of higher level expressive information of music parameters are of great interest (RIMM, MEGA, Wanderley/Battier).

A range of sensors and processing algorithms are available for specific applications, each with advantages and drawbacks according to the context of their use. This is especially true in dance and installation environments where video systems are often used to track movements or objects of interest (*Fingerprint, Palindrom, SoftVNS, BigEye*).

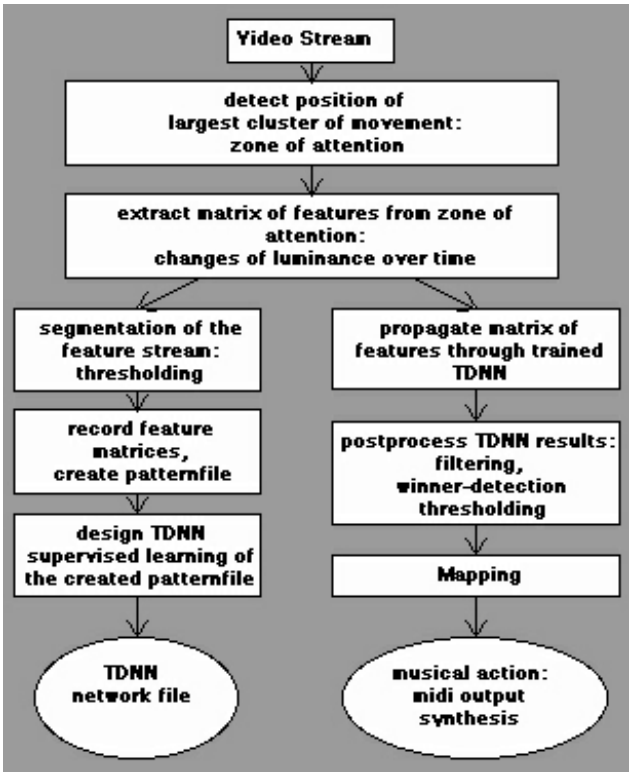

Figure 1: Dataflow

In this paper we describe a set of 17 gestures taught to a video based system developed for the recognition of gestural data in real-time

2. OVERVIEW OF THE GESTURE RECOGNITION PROCESS

The video stream from a standard dv-camera is analyzed and information relating to luminance magnitudes of consecutive video frames is extracted and presented to an artificial neural network.

The output of the neural network is evaluated using a post processing function that results in a binary output signifying the recognition of a trained gesture.

Video digitization and visualization, recording and editing of data as well as real-time gesture recognition is realized on a Pentium-4 Linux system running at 2.8 GHz with 25 Frames per

seconds video resolution and running *jMax* 2.5.1. A standard consumer video camera as well as a low cost web-cam, had been used successfully.

A mapping patch combines the output of the neural network with continuous parameters derived from the gesture. Different setups of the mapping can be switched through the recognition process itself.

3. EXPERIMENTAL SET OF HAND GESTURES

Based on previous work demonstrating successfull recognition of a small set of four hand gestures we choose a larger set of hand gestures to be taught and recognised. The set is experimental, in such a way, that it is incomplete and does not represent an entire hirarchy or typography of gestures. But it was composed out of gestures representing different types like full handed or single finger oriented or functional like pointing or abstract like moving the hand in a certain way.

The gesture set is recorded two times on a consumer dv camera. One set is the trainings set and one set the test set. For each set every gesture is recorded 2 times at 3 different speeds: slow, medium, fast.

The training-set is then digitized using a *jMax* patch described below. For each class of gestures a *pattern-file* is created in the highest resolution. This gives the possibility to compose in a later stage various trainings pattern-sets in different resolutions or integrating later new recorded gestures.

The patterns are stored in the pattern-file format of the Stuttgarter Neuronale Netz Simulator.

Table 1: Experimental Gesture Set

	Gestures	**Abbr.**
1	Flat figure 8	fig8
2	index moving with hand right	wa-ri
3	index moving with hand left	wa-le
4	pointing down	in-do
5	pointing up	in-up
6	pointing left	in-le
7	pointing right	in-re
8	Index & middle moving “walking” (fingers down)	walk
9	index rotating in a circle	ro-ci
10	waving hand , finger down	wa-do
11	musical stop (grasping, circle)	mu-st
12	hand opening, hand horizontal, fingers to front	hh-op
13	hand closing, hand horizontal, fingers to front	hh-cl
14	hand opening, fingers up	open
15	hand closing, fingers up	close
16	waving, fingers up	waving
17	flatter, all fingers move randomly except thumb	flatter

4. PREPROCESSING AND FEATURE EXTRACTION

According to the layout of the algorithm in figure 1 the incoming video stream is searched to find the largest cluster of luminance variations in consecutive video frames.

For this a clustering process is used to detect the overall location of the gesture. The aim of this is to achieve a position independent sub-frame and also to increase the resolution of the relevant video data only for the relevant parts of the visual stream. In other words: the whole picture is inspected to indicate a small zone for high resolution data capture (zone of attention).

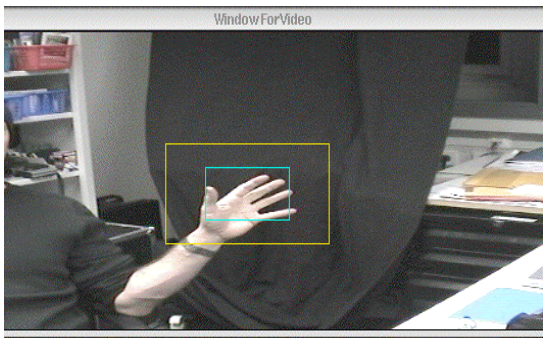

Figure 2: Video Input and Zone of Attention

Figure 3: Feature Extraction

In our case the whole video frame is searched to find the location where the gesture of the hand is produced. A certain area (zone of attention) is then cut out and used to extract the feature-map for the neural network. The coordinates of the zone of attention are also extracted and can be used as subgestural information to control additional parameters.

5. VARIATIONS OF PATTERNS BY AFFINE TRANSFORMATIONS

Recording and editing of gestures is time consuming. We applied linear affine transformations on each recorded gesture in the following order to multiply the recorded training sets for increased stability and quality of the recognition:

- *stretching* in the x, y direction
- *rotation* in the x, y plane around the center of gravity of the gesture

- *shifting* in the x, y plane

Each of this transformation operations are realized as an appropriate method in *jMax* to compute the transformation.

6. DESIGN OF THE ARTIFICIAL NEURAL NETWORK: TDNN

For the recognition of the gestures we use a Time Delay Neural Network architecture. The Time Delay structure was developed for phoneme recognition (Waibel 1989, Berthold 1994) but has been successfully applied in gestural processing (Modler, Zannos 1997) as well as in musical audio applications (Marolt 1999). A certain form of Time Delay Networks was succesfully applied for recognition of image sequences for gestural control (Vassilakis, Howell, 2001)

This neural network architecture provides recognition of timed patterns at low processing power requirements independent from the pattern speed reference.

The network for the gesture set we designed to have one input layer with 900x6 input units, one hidden layer with 50x4 units and one output layer with 17 units. For each frame of the video stream the network is presented a new set of pixels plus the previous 5 sets. This can be seen as a windowing function over the whole data stream. (Zell, 1994)

The TDNN is design and trained using the Stuttagrter Neuronale Netz Simulator (SNNS) with the pattern-sets created as described above.

The time needed to teach the TDNN varies depending on the size of the network, the size of each pattern and the number of instances for each gesture type, and on the number of cycles a pattern set has to be taught before the pattern set is learnt sufficiently.

One Cycle of a whole pattern- set for the set of 17 gestures needs about 20 hours to be taught on a Pentium 4, 2.2GHz. A pattern-set we are using for training the network contains about 17 times approximately 1300 patterns giving a rough total sum of 22100 patterns.

7. POST-PROCESSING

The output of the neural network is processed with threshold and filter functions.

Together with the overall level of the energy of the hand gesture the onset and offset of a gesture as well as the type of the gesture is estimated.

The output of the recognition process can be displayed on the screen as well as be sent to external devices via Midi or it can be internally used to control routing, mapping or sound parameters.

8. RECOGNITION RATES

Various parameters of the hand gestures, like distance, location, rotation and size of the hand in the video frame have influence on the recognition results.

Overall light conditions, gesture speed and the number of taught gestures etc, also have an impact on recognition rates.

Figure 4 shows the jMAx patch displaying the resulting values of the 17 output neurons and the energy of the gesture and the index of the winner neuron as time diagrams. On the right side a column with the abbreviations of the gestures names can be seen. On the left side the actual values are displayed. In the shown diagrams three instances of the gesture "waving" are presented and successfully recognized. In the middle part of the patch the energy of the gesture is displayed and below the index of the winner.

Smaller excitation peaks of non valid neurons, like "flatter", "close" or "mus-st" are not strong enough to disturb the correct recognition.

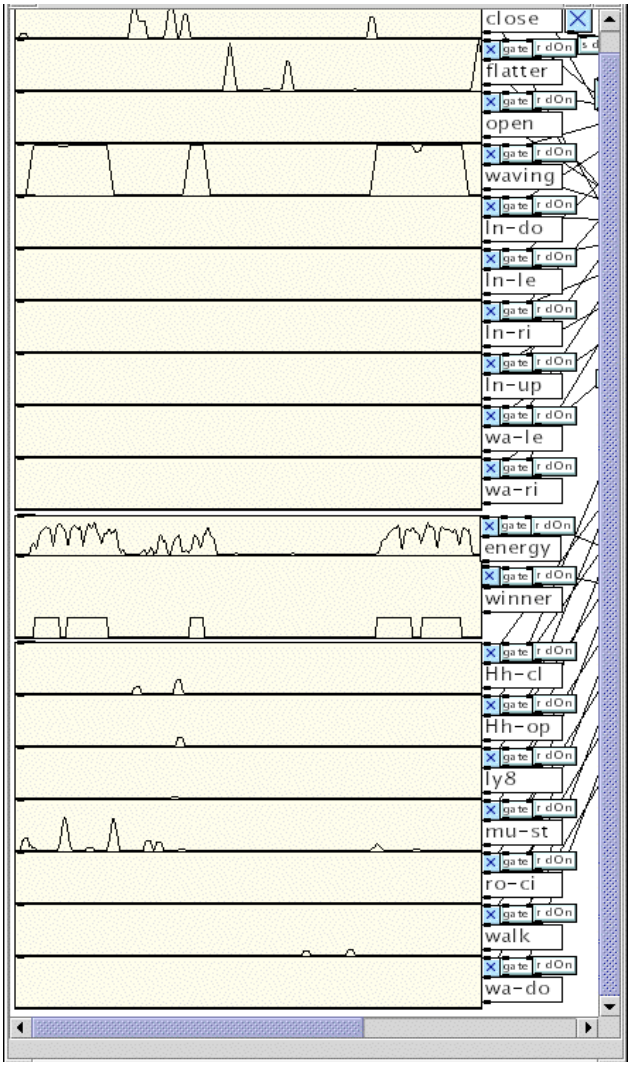

Figure 4: Results of the Output Neurons, Energy and Index of Winner over Time for 17 Gestures

An output value from the neuron close to 1.0 (upper line) indicates the network assuming the output unit as recognized.

For the estimation of recognition rates we use different means. First we process the original gestures played back from the dv camera and directly processed through the system. This gives a recognition rate of about 100%.

Then the recorded test set is processed. It contains also 17 gestures recorded 2 times at 3 different speeds. The test-set is recognized at a rates of about 93 %.

As a third test we feed directly the gestures from a live performer. The performer chooses a set of 30 gestures at random and presents them to the system through the dv camera. The outcome varies depending on how the gestures are presented to the network, and what gestures are presented. At this stage 80% recognition rates have been demonstrated.

We envisage improvements through future developments of the algorithm and data gathering and feature extraction process. This order of recognition rate may be sufficient for free structured musical works like improvised and semi-improvised music which will be determined through future practice-based research. The communication of meaningful performance data to an interactive computer system is of such significance that musical benefits from this techniques are likely without increased recognition rates.

Testing a selected subset of the gestures for example a set of open handed gestures like "waving,, flatter, stop, flat-figure-8, waving-down" rises the recognition rate up to about 90%.

Figure 5: Motion-Energy, Winner-Index, Output of Index-Up and Index-Down Neurons over Time

9. ROBUSTNESS

Camera distance

Adjusting the video frame through the cameras zoom function the distance of the camera to the performer can be increased for the recognition phase.

For a network trained with gestural data recorded at a distance of 1.6m we achieved a loss in recognition of about 0%-10% increasing the cameras distance to up to 8m.

Light Intensity and Light Direction

The system is robust against light intensities different from light intensities of the training data. We had no significant loss in recognition rates when reducing the light-intensity from 840 Lux (training data) to about 180 Lux (recognition data) which is equivalent to a focal reduction of 2.5 steps.

The system is sensible against different light directions. since they produce different overall shapes of the extracted features of the hand.

10. RECOGNITION DELAY

A crucial question is the processing time of a recognition process especially in a musical real-time environment. Due to the structure of the Time Delay Neural Network gestural data is presented continuously to the network. This can be seen as the recognition process can start at any point in the data-stream. It is possible to recognize the gesture in before the gesture is finished or before it is "felt" to be finished.

In figure 5 the output of the motion energy of the attention rectangle, the index of the estimated gesture (winner), motion parameters of the attention rectangle (arVX, arVY, energyAR) and the output of the Neurons for the gestures "Index-Up" and "Index Down" are shown over time.

The diagram shows that the gestures are recognized before the gestures are fully completed. The maximal motion energy of the gestures are corresponding with the recognition results.

11. INTEGRATION INTO A PERFORMANCE ENVIRONMENT

The whole system is integrated into the jMax environment For that we realized following objects as external objects in jMax: (Modler, 2002):

Table 4: External jMax Objects

Object	Purpose
grabber	video input
window	video ouput and data visualisation:
recorder	recording editing and saving of multidimensional data
feature	feature extraction
nn	TDNN and patternfiles, gesture recognition

The object realizing the Time Delay Neural Network is based on the kernel of the Stuttagarter Neuronale Netz Simulator (SNNS, 1994). It is integrated into the jMax environment.

The results of the recognition process as well as values of the extracted features can be sent to external devices through the standard *jMax* midi-port.

In a test setup we trigger different audio samples in jMax from the output of the gesture recognition process . Additional parameters, like the position of the Zone of Attention or the volume of the luminance inside the Zone can be sent via midi or ethernet as sub-gestural control information to remote units. Depending on the load of the processor audio synthesis can be triggered and controlled on the same machine.

12. MAPPING STRATEGIES FOR EXPRESSIVE CONTROL OF PARAMETERS

To control musical performances we took 3 schemes for connecting the output of the neuronal network to musical parameters into account.

Triggering - Symbolic

For each trained gesture an audio file is associated. If the recognition algorithm results a valid winner, the appropriate soundfile is played. For this a larger set is desirable, since a restricted set of gestures reduces the possibility of variations.

Triggering & Continuous Control - Symbolic & Parametric

A combination of symbolic commands derived from the recognition process and additional continuous parameters are used. A selection situation is a paradigm for such a mapping: For example the x coordinate of the hand location in the video image is mapped to the index of the sound selection.

The gesture "pointing down" selects the sound and "open" plays it back with a volume proportional to the energy of the "open" gesture. The latter mapping is used for expressive control of the onset of the sound. Additional sensor data gathered from sensor devices like accelerometers, can be integrated into the recognition and mapping process.

Direct mapping - Parametric

Since the results of the output neurons is a continuous stream of floating point values we directly can map this stream to a set of continuous parameters. For this we connect them to the volume controls of a set of sine wave generators. Each output neuron then controls the loudness of a sine wave generator.

13. CONCLUSIONS

Our aim was to investigate the use of a small experimental set of hand gestures in combination with a video and neural network based gesture recognition system.

Gestures for the set are chooen according to the scheme to combine gestures of different types.

To increase the robustness of the recognition results and to reduce the necessary number of gesture recordings we successfully used affine transformations on the training-set for creating variations to extend the set of recorded gesture patterns. Through a set of external objects for jMax we provide all facilities to record, edit and generate the desired patternfiles.

For the recognition we designed a Time Delay Neural Network architecture and trained it with the generated patterns successfully. The trained Network is loaded into the jMax environment and used successfully for the gesture recognition process in real-time. The recognition rates differ from about 90% for the laboratory training-set to about 80% for the real performing situation with live camera input.

We showed that the dictionary of hand gestures can be used in different types of mappings to control musical parameters. The combination of symbolic triggering of musical events with the parametric use of continuous sub-gestural data like the location of the hand in the x,y image or the energy of the hand gesture offer expressive musical control.

Further work in this area will provide details about the maximum number of different gestures which can be learned sufficiently from a chosen network design and about the use of different gesture types, such as hand gestures or full body gestures. Also more detailed setups for the musical mapping have to be investigated.

The system provides a promising environment for experimental setups using gesture recognition for expressive control of musical parameters.

14. REFERENCES

[1] Berthold, R. Michael.: A Time Delay radial basis function network for phoneme recognition. In Proceedings of IEEE International Conference on Neural Networks, volume 7, pages 4470-- 4473, Orlando, FL, 1994. IEEE Computer Society Press.

[2] de Cecco, M., Dechelle, F., jMax/FTS Documentation, http://www.ircam.fr, 1999

[3] Harling, P.A., Edwards, A.D.N., (Eds), Proceedings of Gesture Workshop'96, Springer-Verlag (Pub.), 1997.

[4] Hofmann, F.G, Hommel, G.: Analyzing Human Gestural Motions Using Acceleration Sensors., Proc. of the Gesture Workshop '96 (GW'96), University of York, UK

[5] Marolt, Matia, A Comparison of feed forward neural network architectures for piano music transcription, Proceedings of the ICMC 1999, ICMA 1999

[6] MEGA, Multisensory Expressive Gesture Applications, V Framework Programme IST Project No.1999-20410, 2002, http://www.megaproject.org/

[7] Modler, Paul, Zannos, Ioannis, Emotional Aspects of Gesture Recognition by Neural Networks, using dedicated Input Devices, in Antonio Camurri (ed.) Proc. of KANSEI The Technology of Emotion, AIMI International Workshop, Universita Genova, Genova 1997

[8] Modler, Paul, A General Purpose Open Source Artificial Neural Network Simulator for jMax, IRCAM-Forum, Nov. 2002, Paris

[9] Myatt, A: Strategies for interaction in *construction 3*, Organised Sound, Volume 7 Number 3, CUP, Cambridge UK 2002, pp157-169

[10] Palindrome, http://www.palindrome.de/

[11] The RIMM Project, Real-time Interactive Multiple Media Content Generation Using High Performance Computing and Multi-Parametric Human-Computer Interfaces, European Commission 5th Framework programme Information, Societies, Technology 2002, http://www.york.ac.uk/res/rimm/

[12] Rokeby, David, SoftVNS Motion Tracking system, http://www.interlog.com/~drokeby/softVNS.html

[13] SNNS, Stuttgarter Neural Network Simulator, User Manual 4.1, Stuttgart, University of Stuttgart, 1995.

[14] Vassilakis H.,Howell, J. A., Buxton, H. I, Comparison of Feedforward (TDRBF) and Generative (TDRGBN) Network for Gesture Based Control, Proceedings of the Int. Gesture Workshop 2001,

[15] Wachsmuth, I., Froehlich, M. (Eds), Proceedings of the Int. Gesture Workshop'97, Lecture Notes in Artificial Intelligence, Springer-Verlag (Pub.), 1998.

[16] Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, Phoneme recognition using time-delay neural networks, IEEE Transactions On Acoustics, Speech, and Signal Processing, Vol 37, No. 3, pp. 328-339, March 1989.

[17] Wanderley, Marcelo, Battier Marc, Trends in Gestural Control of Music, CD-Rom, 2000, IRCAM, Paris,

[18] Zell, Andres, Simulation Neuronaler Netze, Bonn, Paris: Addison Wesley, 1994.