Neural Networks for Mapping Hand Gestures to Sound Synthesis Parameters

Paul Modler

University of York

plpm1@york.ac.uk

Introduction

This paper expands on our work on mapping hand gestures to musical parameters in an interactive music performance and virtual reality environment[1]. A neural network architecture for subgestures will be introduced to provide a mechanism in order to achieve meaningful control parameters (e.g. aesthetic variations).

We use data obtained from a sensor glove, an input device for digitizing hand and finger motions into multi-parametric data. These data are processed in order to extract meaningful data to control musical structures. This is done by an extended neural network architecture for sub-gestures combined with the extraction of parametric values.

We focus on the mapping of gestural variations onto equivalent musical parameters which could be used in a performance. We set up a dictionary of symbolic and parametric subgestures. Different hand gestures of this dictionary and characteristic variations will be evaluated with respect to their applicability to intuitive control of musical structures.

The system is complemented with a 3D VRML environment, i.e. an animated hand model and behaving representations of musical structures. This 3D representation combines with the gesture processing module and the sound generation engine to produce "Behaving Virtual Musical Objects".

System Architecture

The interactive system we assume comprises the following components which are described below in greater detail.

- a dedicated sensor glove which tracks hand and finger motions;
- a design and control environment for the data glove including preprocessing features (written in JAVA);
- a data processing section based on neural networks for gesture recognition and postprocessing
- a real-time sound synthesis module for advanced synthesis algorithms;
- a virtual reality framework (VRML) for the interaction of the performer with virtual objects, it is included in the JAVA environment.

[1] Earlier versions of this paper were presented at IEEE Conference SMC 1998 and 5th Brazilian Symposium of Computer and Music 1998. We thank the audiences for their helpful criticism.

Reprint from : *Trends in Gestural Control of Music*, M.M. Wanderley and M. Battier, eds.

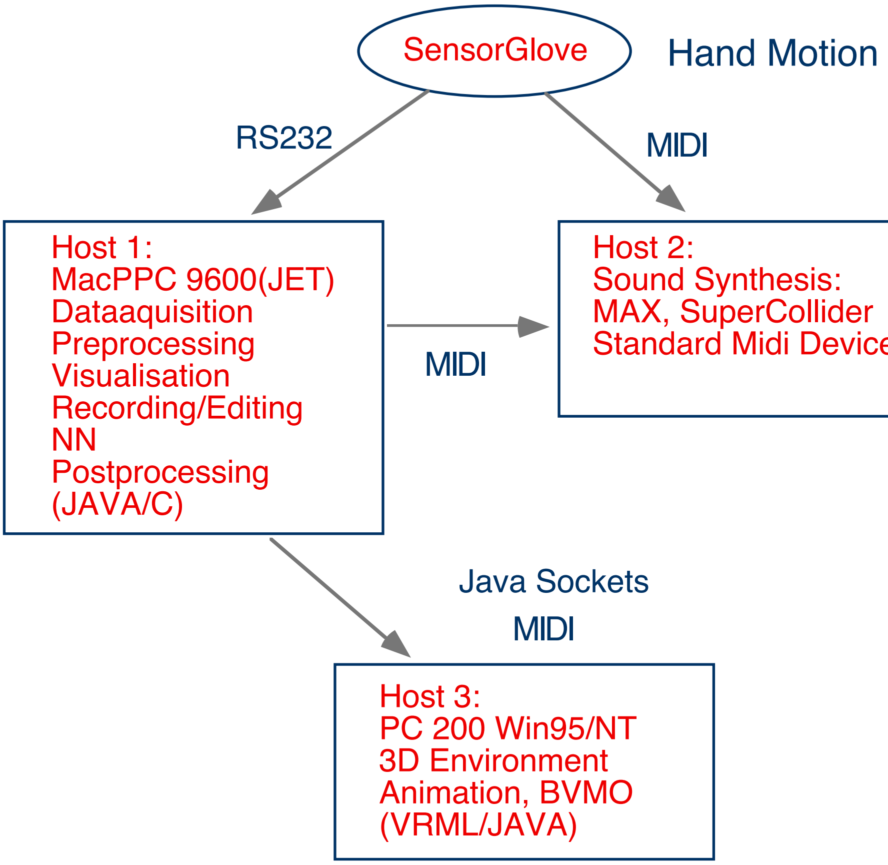
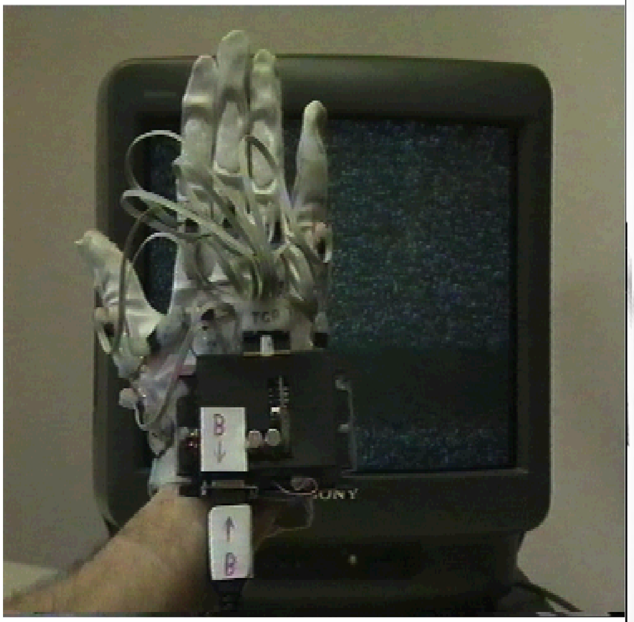


Fig. 1. System Architecture.

User Input Device: The Sensor Glove

The sensor glove, developed by Frank Hofmann at the Technical University of Berlin, is used as an input device (figures 2 and 3).

By tracking 24 finger angles and 3 hand acceleration values, gestures of the hand can be processed by the connected system. As a first approach, direct mapping of single sensor values to sound parameters was used. Although good results concerning the possibilities of controlling parameters of the sound algorithm (frequency modulation, granular synthesis, analog synthesis) have been obtained, drawbacks of such a direct connection emerged as well. E.g., intuitive control of multiple parameters simultaneously turned out to be hard to realize. The data from the Sensor Glove are fed into a postprocessing unit which provides feature extraction and gesture recognition abilities.



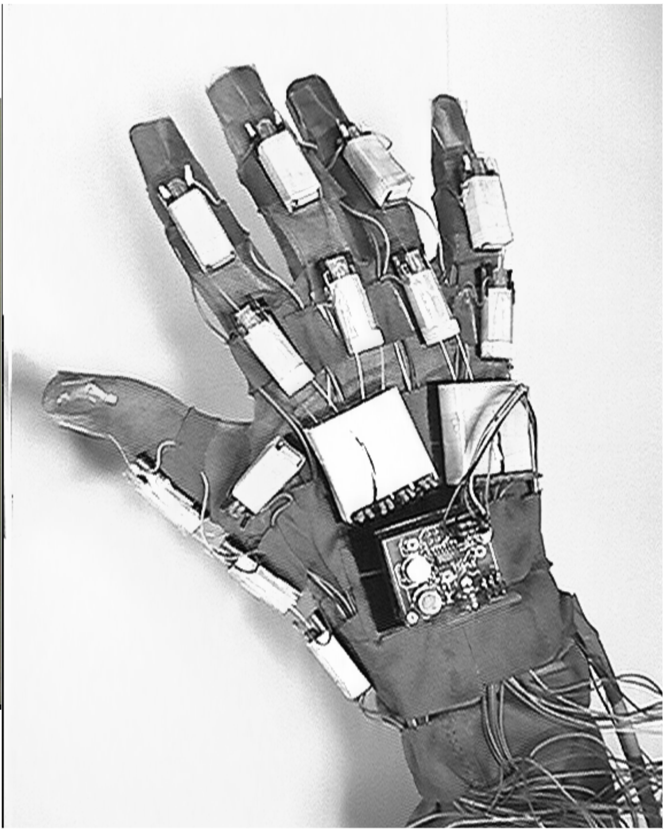


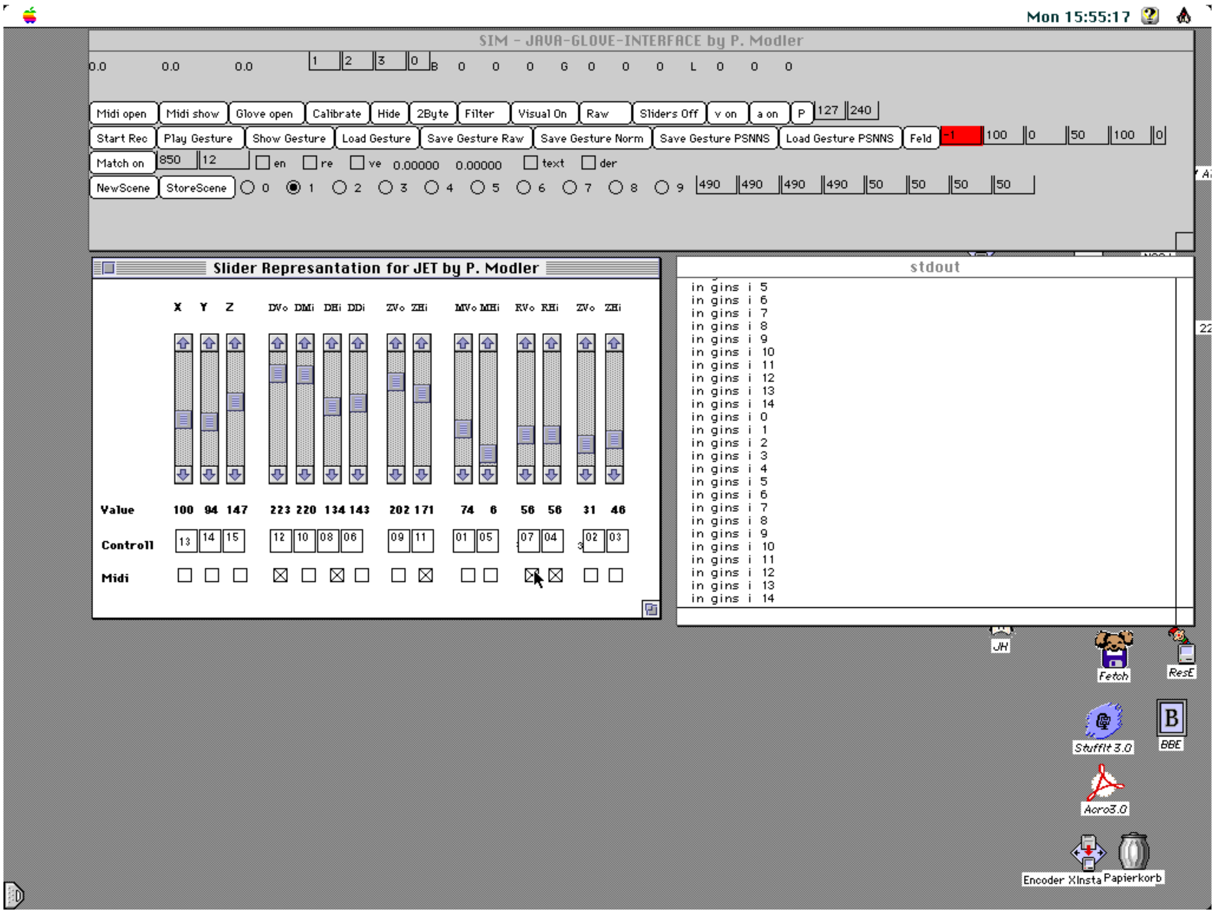
Fig. 2. Sensor Glove version 2. Fig. 3. Sensor Glove version 3 (by Frank Hofmann).

Visual Representation of Gesture Data

The gesture data obtained by the Sensor Glove can be visualized in several ways. The following description is based on glove version 2 providing data form 12 inductive and 3 acceleration sensors.

Slider Representation

Each sensor is represented through a vertical slider. Additionally the numerical value of each sensor is displayed in the bottom line. For midi output an input box and a checkbox is attached to each slider. The input box determines the desired midi controller number, the checkbox controlls activation of MIDI output for that sensor.



	X	Y	Z	DVo	DMi	DHi	DDi	ZVo	ZHi	MVo	MHi	RVo	RHi	ZVo	ZHi
Value	100	94	147	223	220	134	143	202	171	74	6	56	56	31	46
Controll	13	14	15	12	10	08	06	09	11	01	05	07	04	02	03
Midi	□	□	□	☒	□	☒	□	□	☒	□	□	☒	☒	□	□

```
in gins i 5
in gins i 6
in gins i 7
in gins i 8
in gins i 9
in gins i 10
in gins i 11
in gins i 12
in gins i 13
in gins i 14
in gins i 0
in gins i 1
in gins i 2
in gins i 3
in gins i 4
in gins i 5
in gins i 6
in gins i 7
in gins i 8
in gins i 9
in gins i 10
in gins i 11
in gins i 12
in gins i 13
in gins i 14
```

Fig. 4. Slider Representation and MIDI Control of the Sensor Glove.

Sonogram Style Representation of Multi-dimensional Gesture Data: 'Gesture'-gram

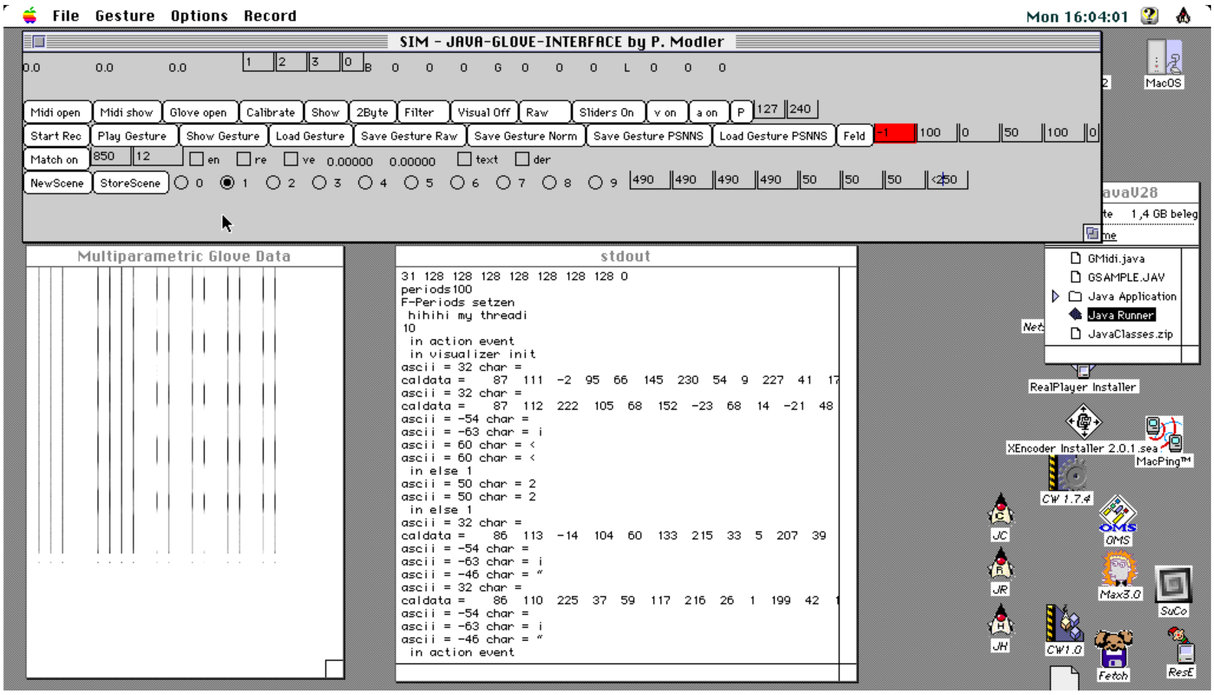
The value of each sensor over the time is represented by a line. The value of a sensor at a certain time is mapped to the greyscale of one line (figure 5).

The lines are grouped according to the physiological arrangement of the hand fingers, preceeded by the 3 accelleration sensors.

- 3 acceleration (x, y, z);
- 4 sensores for the thumb (front, middle, back, diagonal);
- 2 index (front, back);
- 2 middle (front, back);
- 2 ring (front, back);
- 2 small (front, back).

This representation offers good visual resolution for time and values. (1 pixel == 1 timeslice == 20ms). The displayed "gesture"-gram gives an immediate impression of the gesture and permits easy detection of significant parts.

Computed higher derivatives of gestures can be visualized in the same way using the "gesture"-gram object.



```
31 128 128 128 128 128 128 128 0
periods100
F-Periods setzen
 hihihi my threadi
10
 in action event
 in visualizer init
ascii = 32 char =
caldata =    87  111  -2  95  66  145  230  54  9  227  41  17
ascii = 32 char =
caldata =    87  112  222  105  68  152  -23  68  14  -21  48
ascii = -54 char =
ascii = -63 char = i
ascii = 60 char = <
ascii = 60 char = <
 in else 1
ascii = 50 char = 2
ascii = 50 char = 2
 in else 1
ascii = 32 char =
caldata =    86  113  -14  104  60  133  215  33  5  207  39
ascii = -54 char =
ascii = -63 char = i
ascii = -46 char = "
ascii = 32 char =
caldata =    86  110  225  37  59  117  216  26  1  199  42  1
ascii = -54 char =
ascii = -63 char = i
ascii = -46 char = "
 in action event
```

Fig. 5. "Gesture"-gram.

The Concept of Symbolic and Parametric Subgestures

We assume that a gesture consists of subgestures of symbolic and parametric nature (cf. Modler & Zannos 1997). The symbolic category may include static poses as well as non-static subgestures, including time-varying gestures which carry symbolic meaning for the performer

The parametric type should always be time-variant for the control of sound parameters.

With this subgesture architecture sequences of gestures can be recognized using only a part of the hand as a significant symbolic sign, while other parts of the hand movement are used as a parametric sign.

An example

A straightened index finger indicates mouse down (symbolic subgesture) and moving the whole hand determines the alteration of the position of the mouse (parametric subgesture).

Or a straightened index finger selects a sound object (symbolic subgesture) and determines the volume and pitch of the object through the hand position (parametric subgesture).

Subgestures allow for both the application of smaller neural nets as well as the possibility of using trained nets (subgestures) in various compound gestures.

We aim at establishing a set of gestures, suited for the gestural control of musical parameters. The gestures are subdivided into symbolic and parametric subgestures as described above. We show how a dedicated neural network architecture extracts time varying gestures (symbolic gestures). In addition, secondary features such as trajectories of certain sensors or overall hand velocity will be used to extract parametric values (parametric gestures).

Special attention is given to the way a certain gesture can be emphasized or altered with significance for both emotions and music. We are investigating whether the concept of symbolic and parametric gestures can adequately describe the situation of an emotionally expressive performance.

The gestures will be evaluated regarding their potential of providing meaningful symbolic and parametric subgestures, as well as how these subgestures can deal with gestural variations.

Dictionary of Symbolic Subgestures

A set of 15 to 25 gestures was selected and used as a prototype dictionary. For a classification the following categories were used to organize the gesture dictionary.

A) gestures with short (not relevant) start and end transition phases and a static state (pose) (e.g. finger signs for numbers);
B) gestures with repetitive motions (e.g. flat hand moving up and down [slower]);
C) simple (most fingers behave similar) gestures with relevant start and end state and one transition phase (e.g. opening hand from fist);
D) complex (most fingers behave different) gestures with relevant or not relevant start and end states and transition phase (e.g. opening a fist one finger after the other starting with the small finger);
E) compound gestures with various states and continuous transitions (e.g. opening a fist, turning the hand and waving it, then closing the hand to a fist again).

The dictionary is based mainly on gestures of categories A, B and C.

Several gestures of category A (poses) have been selected so that a performer can use them as clear signs. Only few examples of the category D have been chosen, because of their more complex character.

Examples:

Following examples use a shorthand notation for the postion of the fingers:
"o" means curved, "I" straight finger.

The fingers are listed for one hand in the order:
thumb, index, middle, ring, small, beginning with the thumb at the left position. For example ooooo is a fist, Ioooo used as a pose for "ok" or "1".

A) Pose	
Ioooo	thumb straight ("ok" or "1")
IIooo	thumb and index straight
IIIoo	thumb, index, midlle straight
IIIIo	all straight, except ring
IIIII	all fingers straight
ooooo	fist
oIooo	indicate
oIIoo	
oIIIo	
IoooI	
oIIII	
oIooI	
B) Repetitive:	
flat hand down/up	slower
flat hand up/down	faster
oIIoo II-moving	walking
oIooo I-rotating	circle
8	eight (tracing shape in air)
IIIII IIIII-moving	stretched fingers
IIIII IIIII-moving	curved fingers
C) Simple:	
ooooo <-> IIIII	opening and closing fist
IIooo	forming a circle with index and thumb, opening and closing this circle
D) Complex	
oooo->IIIII->IIooo->IIIII	opening a fist, closing it except thumb and index, then opening it again

Dictionary of Parametric Subgestures

Besides the symbolic gestures a set of parametric gestures has been selected for building a dictionary for classification in which the following categories for subgestures are available:

a) alteration of the absolute position of the hand: translation;
b) alteration of the absolute position of the hand: rotation;
c) alteration of velocity (energy);
d) alteration of the cycle duration.

In the dictionary of parametric subgestures we included instances of categories a, b and c. Additional work has to be done regarding the extraction of repetitive cycle time and detection of resulting timing variations.

For the prototype implementation, a module which supervises the combination of symbolic and parametric subgestures has been included. This coordinating device recognizes the influence the extracted subgestures have on the output of the symbolic gestures.

Pattern Recognition of Gestures by Neural Networks

Neural Network Architecture

Based on a prototype implementation for the recognition of continuous performed time-variant gestures (cf. Modler & Zannos 1997) we extended the proposed architecture to deal with the demands of the selected dictionary. Additional input layers have been added for the recognition of subgroups of the gesture dictionary. The layers of the subgroups are connected by separate hidden layers. Figure 6 shows the network and its excitation for a gesture (opening and closing hand) with 15 input neurons from sensors.The three layers are visible, starting from the left with the input layer then hidden layer and output layer with 3 detection neurons.

The input layer shows no activation for the hand movement (column 1-3) followed by 4 columns showing the activation of thumb sensors. Columnns 8 to 15 show the activation of the remainder of the fingers. Dark colour indicate a curved finger and light colour straightened finger. The output layer shows the neuron of the detected gesture in light colour.

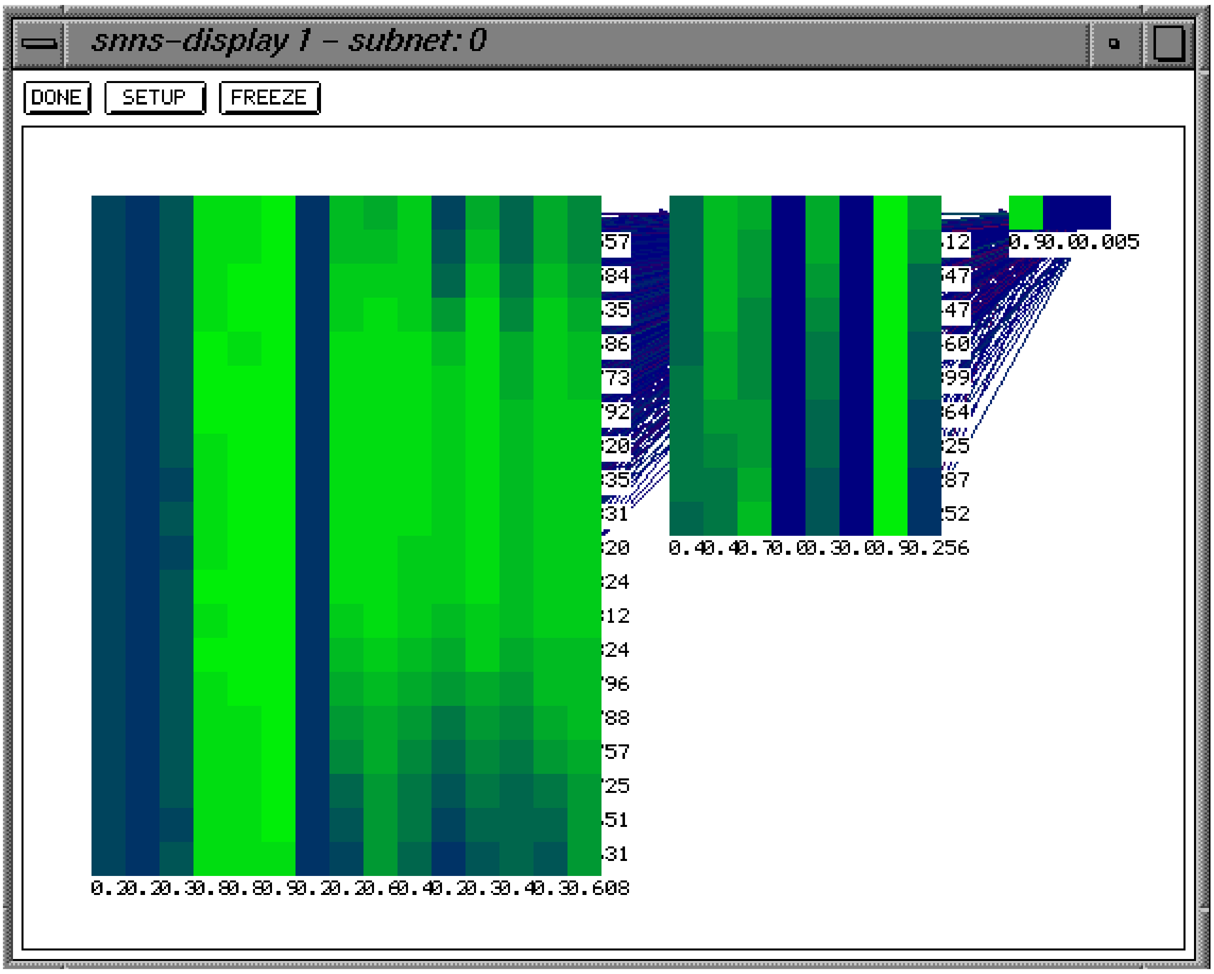


Fig. 6. Neural net for Timed Gesture.

Training Procedure

The Network is trained with a set of recorded instances of the gestures of the symbolic subgesture dictionary.

Both the 2D representation of the sensor data as well as the 3D-representation (see below) offered a good feedback about recorded instances.

Each gesture class was recorded two times in 4 different velocities.

The training of the Neural Net was conducted offline. The resulting net parameters were transferred to the Sensor Glove processing section and integrated in the C/JAVA environment.

Recognition of Subgestures by Neural Networks

For evaluation, time-varying continuous connected phrases of instances of the symbolic subgesture dictionary were presented to the trained net. This was realized online, i.e. the data were passed directly from the glove to the network.

For the selected part of the gesture dictionary the proposed net architecture offered good results, i.e. a recognition rate of about 90 %.

Combination of Extracted Parametric Subgestures with Symbolic Subgestures

The parametric subgestures as proposed above were derived from the sensor values. Further investigations will show whether neural networks can also provide the desired parametric parameters.

The combination of both parameters produced good results in both recognition of a subgesture as well as altering the overall gesture by changing the parametric subgesture (e.g. flat hand, fingers moving up and down [slower] combined with translation movements of the whole hand).

Results

The proposed combination of gesture recognition of symbolic subgestures with parametric ones achieved good results. In other words, they promise to promote and extend the possibilities and variability of a performance conducted with the Sensor Glove.

The concept of symbolic and parametric subgestures as well as the proposed categories offer the performer a guideline to fix a parameter mapping with connected sound synthesis and visualization modules.

The extension of the neural network for the processing of a larger number of features seems to be manageable, but an extension to a multiprocessing parallel architecture has to be considered too.

3D Representation: Integration of Virtual Reality Worlds

Animation of a Hand Model by the Sensor Glove

As a feasibility study, we have created a visual representation of a hand and Virtual Musical Objects in VRML language (figure 7).

The VRML language is a standardized toolkit which provides possibilities for creating three-dimensional environments called virtual worlds. VRML offers the advantage of a platform and browser independent application. Since VRML is so widely accepted, its disadvantage of reduced speed is acceptable.

The hand model is animated with the input from the Sensor Glove. This is realized by a JAVA-VRML interface.

This prototype world can be viewed with a VRML browser that has been integrated into the design and control environment and runs on a combination of JAVA and C.

The VRML - JAVA interface also offers the possibility of dynamically creating or altering existing VRML worlds, in other words, user-provided interaction models such as the animated hand model can then be introduced into unknown worlds (e.g. downloaded from somewhere else).

Complex worlds can be generated with special tools like COSMO Player, MAX3D or VREALM. which then can be animated, investigated, altered, and viewed with the VRML browser. Further down we describe how we use VRML worlds to represent musical objects.

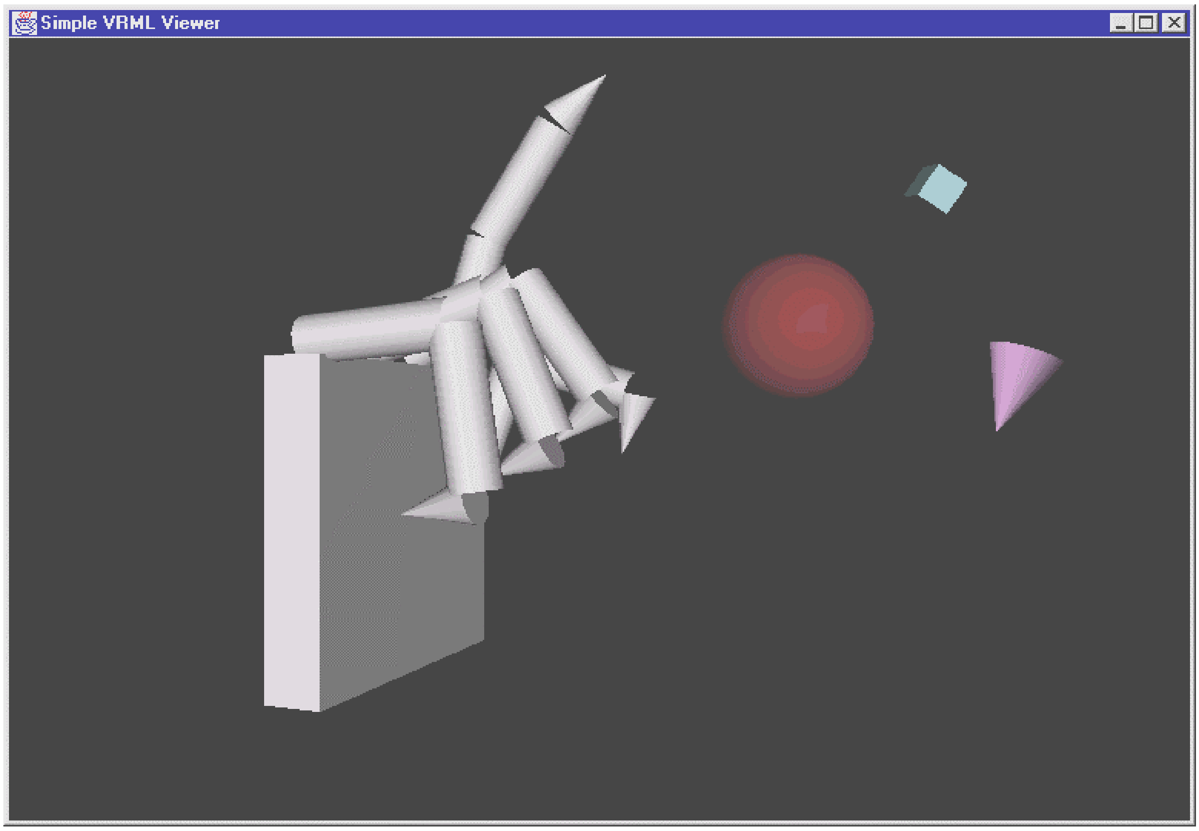


Fig. 7. VRML World with Animated Hand Model and Virtual Musical Objects.

3D Representation of Behaving Virtual Musical Objects (BVMO)

In addition to the hand animation, we developed a framework for the creation of Virtual Musical Objects (VMO). These objects together with the Sensor Glove, constitute the gesture processing section and the sound synthesis module. An extended form (EVMI) of a Virtual Musical Instrument (VMI) has been proposed by Alex Mulder (Mulder 1994).

The VMOs are variable in color, size, form, and position in the surrounding world. Additional features can be defined and controlled, e.g. time-varying alterations of a certain aspect such as color or motion trajectory. This can be regarded as a behaving VMO (BVMO). Through their defined behaviour BVMOs are alike autonomous agents.

For data passing from outside the VRML world to the BVMOs the JAVA-VRML interface is used. Good results have been achieved for animating the hand model and altering BVMOs by user input from the Sensor Glove.

Interactive Sound Synthesis System with Real-Time Control

The synthesis module is designed in order to provide the ability of controlling different sound algorithms in real-time by continuous control parameters. The algorithms are realized in SuperCollider, a special sound synthesis language offering real-time synthesis possibilities as well as LISP-style list processing and a limited SmallTalk like class/object structure (McCartney [4]).

We connected different synthesis techniques (frequency modulation, physical modeling and granular synthesis) with respect to their controllability through the Sensor Glove and the processing system.

Example 1: Frequency modulation

Two symbolic gestures triggered start off and ending of the sound generation. As parametric gestures the energy of the finger movements are mapped to the atack (start) and release (end) time, and the hand rotation to the volume of the generated sound.

Example 2: Physical model of a metal plate

The hit is modelled through the recognition of the symbolic gesture, in this case a straight index finger (oIooo). As a parametric gesture the energy of the finger movement is mapped to the damping factor of the plate model.

Example 3: Granular synthesis:

Synthesis Parameter	Symbolic Subgesture	Parametric Subgesture
pitch rate	"ok" (Ioooo)	energy of hand rotation
time rate	index finger and middle finger moving "walking" (oIIoo)	energy of finger movements
grain duration	straight thumb and straight small finger (IoooI)	energy of finger movements
volume	all fingers moving (IIIII)	energy of finger movements
source sound	fist (oooo)	angle of hand rotation

Realization of the System on a Distributed System, Parallel Processing Aspects

Currently the system is based on a 200 MHz Macintosh compatible 604PPC covering the data acquisition and processing from the Sensor Glove. It also realizes the gesture preprocessing recognition and postprocessing. The VRML 3D-environment is implemented on a 200 MHz PC which communicates with the SensorGlove machine by MIDI. The sound generation is realized on a 200 MHz Macintosh-compatible 604PPC which is controlled by the Sensor Glove machine as well as by MIDI (see figure 1). The training of a the neural network was executed on a Silicon Graphics Indigo2.

The MIDI connections were chosen because of their potential for combining the possible external tools as well as because of their standardized and low cost interfaces. Due to the high data rates produced by the Sensor Glove, a more powerful way of transmission has to be considered. Ethernet TCP, UDP or Firewire connections are possible solutions, since the C/JAVA environment can be connected with standard software tools over these connections.

For growing neural networks and additional processing of the control data, e.g. detection of beats or cyclic structures etc., a wider distribution of the computing tasks onto a larger number of parallel processing units has to be considered.

Besides that, the sound synthesis engine profits from a parallel architecture. Promising tests with a dedicated communication protocol for parallel distributed audio processes MIDAS (Kirk, Hunt [3]) have been accomplished.

Conclusions

Based on our experiments, we come to the following results and conclusions.

The subgestural concept for deriving symbolic and parametric gestures is a good approach for integrating gesture recognition into a performance situation. The neural network pattern recognition is combined with flexible and intuitive possibilities of altering material.

Specific control changes as well as intuitive overall changes can be achieved.

The proposed categories of subgestures offer the performer a comprehensive way to design the behavior of the sound generation. This provides a powerful alternative to the one-to-one mapping of single parameters.

The proposed dictionary of gestures provides the performer with an intuitive means for musical expressiveness and meaningful variations.

Behaving Virtual Musical Objects integrated in a virtual 3D world offer a promising way for novel visual representation of abstract sound generation algorithms. This includes specific control of a sound scene, but also facilitates memorizing and recalling of a sound scene and inspires the user to new combinations.

The combination of the proposed gesture mapping with the BVMOs constitutes a powerful environment not only for interactive preformances, but also for the design of sounds and sound scenes.

Acknowledgments

This work was part of the TEMA project supported by the European Union.
I would like to thank Ioannis Zannos, Frank Hofmann and the Federal Institute of Music Research Berlin who made this work possible and Ursula Kleinhenz for proofreading.

References

[1]

Hommel, G., Hofmann, F. and J. Henz. 1994. "The TU Berlin High-Precision Sensor Glove." In *Proceedings of the Fourth International, Scientific Conference*, Milan: University of Milan.

[2]

Hofmann, F., and G. Hommel. 1997. "Analyzing Human Gesture Motions using Acceleration Sensors." In P. A. Harling and A. D. N. Edwards, eds. *Progress in Gestural Interaction. Proceedings of Gesture Workshop'96*. London: Springer-Verlag Limited, pp. 39-60.

[3]

Kirk, R, and A. Hunt. 1996. "MIDAS-MILAN : An Open Distributed Processing System for Audio Signal Processing", *Journal of the Audio Engineering Society* 44(3): 119-129.

[4]

Mc Cartney, J. *SuperCollider Documentation*. Available at: http://www.audiosynth.com.

[5]

Modler, P. 1996. "Interactive Computer-Music Systems and Concepts of Gestalt." In *Proceedings of the Joint International Conference of Musicology*, Brügge.

[6]

———, and I. Zannos. 1997. "Emotional Aspects of Gesture Recognition by Neural Networks, using dedicated Input Devices". In Kansei, *The Technology of Emotion. Proceedings of the AIMI International Workshop*, A. Camurri, ed. Genoa: Associazione di Informatica Musicale Italiana, October 3-4, pp. 79-86.

[7]

Mulder, A. 1994. "Virtual Musical Instruments: Accessing the Sound Synthesis Universe as a Performer." In Proceedings of the First Brazilian Symposium on Computer Music, pp. 243-250.Available at: http://fassfu.ca/cs/people/ResearchStaff/amulder/personal/vmi/BSCM1.rev.html.

[8]

Schmidt-Atzert, L. 1996. *Lehrbuch der Emotionspsychologie*. Stuttgart, Berlin, Köln:Kohlhammer.

[9]

SNNS, Stuttgarter. 1995. *Neural Network Simulator, User Manual 4.1*. Stuttgart: University of Stuttgart.

[10]

Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. 1989. "Phoneme recognition using time-delay neural networks." *IEEE Transactions On Acoustics, Speech, and Signal Processing,* 37(3): 328-339.

[11]

Wassermann P.D. 1993. *Neural Computing, Theory and Practice*. Van Nostrand Reinhold.

[12]

Wundt, W. 1903. *Grundzüge der Physiologischen Psychologie*. Leipzig: Verlag von Wilhelm Engelmann.

[13]

Zell, A. 1994. *Simulation Neuronaler Netze*. Bonn, Paris: Addison Wesley.