# TOWARDS KANSEI INFORMATION PROCESSING IN MUSIC/DANCE INTERACTIVE MULTIMODAL ENVIRONMENTS

*Antonio Camurri, Roberto Chiarvetto, Alessandro Coglio, Massimiliano Di Stefano, Claudia Liconte, Alberto Massari, Claudio Massucco, Daniela Murta, Stefano Nervi, Giuliano Palmieri, Matteo Ricchetti, Riccardo Rossi, Alessandro Stroscio, Riccardo Trocca*

Laboratory of Musical Informatics - http://MusArt.dist.unige.it
DIST-University of Genova, Viale Causa 13, I-16145, Genova (Italy)

**Abstract.** This paper gives an overview of research projects and of recently developed systems at our Lab on interactive multimodal environments. In particular, we focus on (i) movement and gesture analysis, mainly for non-symbolic, expressive data from human movement and gesture and their relation to music performance; (ii) machine communication to humans, supported by semi-autonomous mobile on-wheels robots technology: how a machine move in the environment can be a further communication channel with humans.

# 1 Introduction

This paper describes our recent work on human-computer communication in interactive Multimodal Environments (MEs) (Camurri 1995; Camurri, Coglio, Coletta and Massucco 1997). In short, an ME is an active space populated by *agents* allowing one or more users to communicate by means of different modalities such as (full-body) movement, gesture, voice. Users get feedback from the ME in real time in terms of sound, music, visual media, lights control, and actuators (e.g. mobile scenography; on-stage robots on wheels behaving as actors or players, possibly equipped by music and multimedia output). High-level interaction requires ME agents to be capable of changing their "character" and social interaction over time. In this paper we describe such input and output communication channels in ME agents and we present systems we recently developed.

# 2 ME Input: Movement and Gesture Analysis

### 2.1 Human Gesture Taxonomies

Current sensor systems technology allows to capture in real time data on movement trajectories of (parts of) the body, velocity, acceleration, images, etc. Such low-level data requires analysis to extract high-level information, symbolic and/or interpretative (or related to Kansei). An open problem in the literature regards the recognition and classification of gestures (Nigay and Coutaz 1993; Schomaker et at 1995), and to identify their semantics according to the context. A classification proposed by Nigay and Coutaz (1993) uses the following gesture categories: *symbolic* for conventional, context-independent, and typically unambiguous expressions, like "OK" and peace signs; *deictic* for entities, like "put that there"; *iconic* for displaying objects, spatial relations, and actions, like illustrating the orientation of two robots at a collision scene; *pantomimic* for expressing a invisible objects or tools. References in the literature are usually specific to hand gesture taxonomies (posture, motion, hand orientation), and to handwriting and pen gestures. Recently, Coutaz proposed a gesture classification according to three major functions: *semiotic* (used to communicate meaningful information), *ergotic* (used to perform manipulation in the real world), *epistemic* (used in learning from the environment, by touching and manipulating objects).

### 2.2 KANSEI and movement analysis

The previous classifications do not include any reference to expressive intentions and interpretation. Our approach is towards this direction, and we shortly present here some of the issues we follow in our work.

The first concerns the extraction of high-level, whole-body features, gesture gestalts, from the observation of a dancer or a performer. This is crucial from the viewpoint of both sensor technology and body language and its integration with music language.
Here follow some examples related to overall, qualitative evaluations: "how *fast* the movement is", "how much *in tempo* the dancer (or part of her body) is moving", "*how* she occupies the stage space", "evaluation of how *smooth/nervous* is the movement", "measure of the coordination between arms", qualitative evaluation of her equilibrium, stability, her potentiality to move in the immediate future, etc. These qualitative information are typically obtained from integrating over time and space a number of different sensors. Typical observation time windows range approximately from 0,5-1s to 3-5s, and beyond. These time windows on which the agent input modules operate can vary dynamically, for example on the basis of the "quality" of the recognition: a decreased quality in the movement recognition (e.g., different agents or input modules return conflicting data) can cause the agent to feed back its input modules to vary their time granularity and/or their time window on which they operate.

A further issue concerns studies on body movement in the direction like, for example, the work of Laban, with particular regard to his *Theroy of Effort* (Maletic 1987). It includes studies on the relation between body movement and intentions in communication.

Another important issue concerns the physics and body response (Winkler 1995): actions involving movement and gesture are executed by parts of the body, each

having its limitations in terms of range of motion, speed, force, as well as weight and privileged directions. Furthermore, an action is characterized by features like ease and accuracy of execution, repeatability, measures of fatigue, required energy, etc.

(Sawada et al. 1995) argue that information on dynamics rather kinematics is closer to the Kansei aspects in movement detection.

**2.3 Our Approach: the Technology**

We aim at using and developing sensor technologies for MEs according to the following guidelines:

1. to adopt or to develop robust, low-cost, as simple as possible technologies;
2. to let users (dancer, director, choreographer, composer) free to move and think of their specific tasks (e.g., dance and music performance, design of a choreography), instead of technology. This implies the use of mainly spatial sensors not requiring any light or dress limitation, non intrusive, or small, wireless on-body sensors;
3. an ME usually needs multi-sensor fusion, to assure that all the needed input can be gathered, to be integrated for as complex as necessary gesture and movement "style" analysis.

In our systems we adopted, developed, and integrated systems based on active infrared, ultrasound, pressure, capacitive, radio, and camera-based sensors.

# 3. Movement and Gesture Detection

This section introduces systems we recently developed for movement and gesture analysis, following the guidelines and issues skethed in the previous sections. In real applications, these different systems can be integrated for gathering more effectively movement information.

**3.1 Camera based systems: the EyesWeb**

We started experimenting with camera-based sensors by using special purpose devices (Costel, MacReflex) originally designed for bioengineering applications (Camurri et al 1986; Camurri et al. 1993). *EyesWeb* is a software module and library for real-time movement and gesture analysis, based on the Matrox Meteor frame grabber board for PC Pentium and a videocamera. We use both standard color and infrared camera. The system works as follows to extract high-level information about movement: a preprocessing phase grabs in real time the data from the camera and tries to recognise the posture of the human figure in the current frame as belonging to one of a few stereotypical posture or clusters (examples are given in figure 1). This classification is done by extracting a small number of parameters and then send them as inputs of a self-organizing neural network. Then, in the second phase, (i) a segmentation can be done on the stream of frames according to the frames of change of cluster, and (ii) different analysis algorithms can be applied to extracting features to the different video segments.

The low-level preprocessing software is roughly structured in the following steps:

0. In a preliminary, off-line phase, an image of the background is acquired and stored as a reference background array (Images are here stored as 320x200 buffers of color pixels);
1. As soon as a human body enters in the observed space (e.g. a portion of the stage), the system starts a continuous image grabbing;
2. Each body image is copied in a separate foreground image buffer;
3. A two-level quantization resulting from the difference between the current image (foreground, with the dancer) and the background image (acquired in step 0) is done;
4. Denoise the resulting 2 level image with a threshold based algorithm using a mask;
5. Trace the sub-image containing the whole human body figure and calculating its dimensions;
6. Compute the percentage of black pixels in the rectangle, the processing time, the equilibrium parameter (defined mainly in terms of the distance between the last pixel – from left to right - of the right foot of the body and the first pixel of the left foot, and of the amount of pixels on the ground of each foot), and a few other qualitative parameters.
7. Both live grabbing image and this resulting 2 level quantized image containing the relevant rectangle and the parameters can be displayed in real time (as in figure 1).

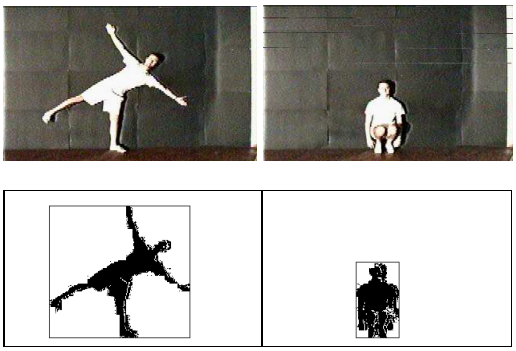

*Figure 1Two examples of different kinds of posture and the output of the preprocessing module.*

As an example, some of the computed values for the static single frame analysis of figure 1a are the following:

*Rectangle Height = 169 pixels*
*Rectangle Witdth = 175 pixels*
*% of black pixels = 17 pixels*
*Equilibrium = 12*

These are some of the inputs to the neural network module for coarse-grain classification of the posture. This allows to classify the input posture among a number of classes of stereotypical postures (e.g. crouched in front of the camera; standing in front with no overlap between arms and body; lateral view; etc.), some of them shown in figure 1. Of course, each

posture is characterized by different feature extraction algorithms.
A segmentation and subsequent dynamic analysis is then performed on the informations extracted from the different frames to detect dynamic features on the evolution of the movement.
The system is typically used integrated with analogic sensors on the floor which contribute with further data to the recognition processes (e.g. to disambiguate orientation of the torso, to give information on the third dimension – depth, etc.).
The processing time for the preprocesing phase is about 50ms per frame on a Pentium Pro 200 / Windows NT (up to 20 frames per second).

### 3.2 The DanceWeb hardware

DanceWeb is a data acquisition system based on the Intel 87C51Fx microcontroller, capable to manage (in its standard version) up to 40 ultrasonic sensors and 16 digital inputs. The ultrasonic sensors are configured in 8 groups of 5 sensors. Each sensor can be individually enabled. The digital inputs are always enabled.
The groups are sampled in a loop, with a sampling time which can be set via software between 15ms and 10s. The digital inputs are read at every sampling time.
The system has a serial RS 232 standard port for data and control communication, which can be configured via software up to 115 Kbaud (a FIFO UART is used).
Sampled data are collected and shared by a Win 95 and NT compatible software DLL, allowing any Win32 application to directly use the system.

### 3.3 The V-scope

V-scope is a wireless infrared/ultrasound sensor system developed by Lipman Ltd. for the real-time tracking of the position of up to eight markers placed on the human body (e.g. on the articulatory joints) or in general on moving objects (e.g., on a music instrument). The hardware is composed of the markers, three tx/rx towers for real-time detection of marker position, and a main processing unit connected via a serial link to a computer. The sampling rate can vary from 5 to several hundreds of msec per marker and the range of measuring depth can vary from 2 to 5m. More details can be found in (Camurri 1997).
A neural network-based module has been developed to process the inputs from Vscope. It adapts to gesture recognition an approach originally developed for handwriting: in analogy with *graphemes* and *strokes* (handwriting primitives), it is possible to train self-organizing nets with gesture primitives (*gestlets*). In the current implementation, neural modules have been developed for the classification of hands trajectories.
Another interesting category of information - in the direction of Kansei perception - is based on force field and abstract potential metaphors: for example, we investigated the mapping of (x,y,z) coordinates of parts of the body of the dancer on a field. For example, a field might map the stage in different areas, each characterized by a different reactivity. In a simple mapping, the (x,y) coordinates of the torso of the dancer can be directly mapped on the field-stage. In more sophisticated applications, the field map can be a metaphor of music spaces, which can be navigated. The types of movement and gesture of the dancer can correspond to different navigation processes in the map: moving from an area to another means a continuous change in the way the observer agents of the MEs interpret the dance and in changes in music rules and in the music objects database.

### 3.4 DressWare

DressWare (De Rossi et al. 1997) refers to a novel family of sensor technologies consisting of wearable piezoresistive fabrics.
An ongoing project at our Lab concerns the experimentation in MEs of the DressWare sensors systems developed by Danilo DeRossi and collegues.

## 4. Low-level software: Mummia

*Mummia* is a software system and reusable library developed by Claudio Massucco in cooperation with the Laboratorio di Informatica Musicale and Soundcage. Mummia is a software environment for the development and supervision of real-time, dance- and gesture-driven performances (Rowe 1993). Material provided by a composer is dynamically rearranged according to composer's rules and external inputs, typically dancers' movement.
Mummia's model is based on *Virtual sensors* and *Midi agents* (note that an ME agent may embed a dynamic network of Midi agents: see (Camurri, Coglio et al 1997)). A *Virtual sensor* can be either a physical sensor, an elaboration of a physical sensors (e.g., its derivative), or the fusion of data from different sensors (e.g., the reconstruction of a complex body movement). A *Midi agent* is an agent able to produce a MIDI output given (i) its internal state, (ii) continuous parameters and (iii) triggers that modify its behavior. Typically, a single Midi agent only controls a small portion of the global score/performance, for a given duration of time. Of course, a music parameter can be controlled by different agents, e.g. in a sort of superposition of effects.
A Mummia *performance object* maps virtual sensors values in parameters and triggers for Midi agents, and schedule agents to produce the final music output.
Mummia is in some extent similar to Opcode's Max. Several reasons led us to the development of such a novel software program and library. A Max patch is equivalent to one or some Mummia Midi agents, and, in later version of Max, also to Midi agents and virtual sensors. Max is designed to easily develop, test, and use one or a few agents, while Mummia is designed to manage the combinations and the dynamic remapping of many agents.
Further, Mummia is an open and scalable architecture: the Mummia library and scheduler is for example used as a low-level platform in several applications (e.g. Camurri Coglio Coletta Massucco 1997). New hardware (analog or digital sensors, actuators, etc.) can be easily added either connecting them to the

acquisition cards and software drivers supported by the library, or by writing C/C++ code for interfacing new ones. The Mummia reusable library supports communication via MIDI, serial (up to 114KB/s), via sockets, and the Microsoft DCOM standard. It runs on PC under Windows 3.11, 95, and NT environment. It provides an editor and a compiler for real-time performances.

# 5. ME Output: Robots and Effectors

An important aspect of MEs concerns the possibility to control the movement of physical objects, including robots and in general effectors. Several applications we recently developed integrate semi-autonomous mobile systems: for example, special robots on wheels capable of interacting with actors, dancers on stage, with the public of a theatre, with the visitors of a museum area exhibition, or in other ME application scenarios. We recently used in a concert a small, wireless mobile robot navigating on stage, "wearing" audio amplifier and loudspeakers in a real, "physical" music spatialisation (vs. virtual audio spatialisation). Its movement (smoothness, paths, way of avoiding obstacles, etc.) depends on both the interaction with performers or dancers along the performance and on its "personality" and goals it has to fulfill. In a performer-robot communication, the movements of the trombone performer on stage, detected by V-scope with markers on the instrument, as well as the sound produced are inputs for the robot, which can therefore change its behaviour.

We developed such a machine prototype for multimedia concerts, theatrical as well as museal MEs. It consists of a small robot on wheels (Pioneer 1 from Real World Interface Inc. designed by Kurt Konolige), equipped with on-board sensors, a computer for the local low-level processing and communication of sensorial data, audio radio links, on-board amplifier and loudspeakers. Such system is capable to move, navigate, and react in real time to events happening on stage (e.g. actions performed by the actors or the spectators), to acquire sounds from the environment, and perform sound and musical tasks (e.g., music "physical" spatialisation). To avoid the system get lost (simple odometry is often not sufficient for real tasks), it is possible to displace in the environment small "lighthouses" placed on the ceiling in crucial points (e.g., near doors or area entrances, goal stations, edges of the environment). This helps the system to maintain updated its position during navigation in a number of environmental situations. Our localization system consists of simple IR receivers (the "lighthouses") displaced on the ceiling in the environment, each capable of receiving the IR message from the robot when it passes under the receiver. The agent has a map of the area, which includes the positions of all the lighthouses. Each IR receiver on the ceiling correspond to a separate digital input, so, when activated, the agent roughly knows where the robot is and therefore can update its position in the map. In museal applications, in cases when the robot is temporarily lost and is in a recovery phase, the agent must continue to maintain a believable character (Bates 1994), by continuing to interact with visitors while searching for a wall, a corridor, a lighthouse.

Our ME requirements do not include area surveillance, disabled transportation, and other applications available with devices that are, from the robotics and mechanics viewpoint, much more sophisticated and expensive (roughly more than one order of magnitude). The robotics platform can therefore be small and "light", with a required payload of only a few kilograms, with satisfactory battery autonomy..

Our system software uses the Saphira navigation software library (developed by Kurt Konolige at SRI). The system has been equipped with a set of sound and multimedia devices: the current setup features an audio radio link, a stereo audio amplifier, a small, high-quality loudspeaker, the lighthouse IR transmitters/receiver system, and possibly infrared radio-controls for multimedia equipment located in the stage or exhibition area.. A separate radio link is used to send audio signals to the machine (voice, sound and music signals produced in real time on the supervision workstation running the agent software).

The machine should exhibit different "behaviors" related to its movement, since movement is one the main communication channels of its emotional state, integrated with voice, sound, music, activation and control of multimedia equipment in the environment. We have developed movement attributes and behaviors like "go there" with smooth, nervous, lazy, straight or tail-wagging movements; "follow that moving object", etc. A set of such behaviors has been developed starting from the Saphira software, which provides tools for designing navigation behaviors, managed within our agent architecture (Camurri Coglio et al 97).

Initial public experimentation of a previous prototype based on a LabMate robotic platform was carried out in a museum exhibition (Palazzo Ducale, Genova), 19-22 December 1993. The current system based on Pioneer 1 has been experimented in public exhibitions (Salone del Lavoro Ercole, Magazzini del Cotone, Porto Antico, Genova, January 1996; Imparagiocando, Palazzo Ducale, Genova, March-April 1996), in public concerts, and will be installed in the museum Città dei Bambini, Porto Antico, Genova (Camurri, Dondi and Gambardella 1997).

### 5.1 Robot's Emotional Component

A companion paper (Camurri, Ferrentino, Dapelo 1997) presents a computational model of artificial emotions, used to implement the emotional component of agents. Here we sketch a simplification of this model designed for the *cicerone-robot* of the Music Atelier of the "La Città dei bambini" (Camurri Dondi and Gambardella 1997), a permanent exhibition in Genova. Please refer to the original for a deeper discussion.

It is developed on a plane whose coordinates are E (*vital energy*) on the abscissas axis, and S (*emotional stability*) on the ordinates axis, following studies of psychologists.
The emotion space is here a circle divided into eight sectors (mood zones) representing the possible emotional states (two opposite to the center zones are antithetical moods) and in three rings representing the mood intensity (weak-average-strong): combining zones and rings we can obtain 24 cells characteristic of the main emotional states.

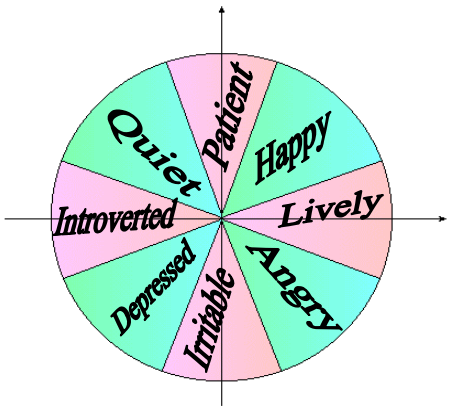


***Figure 2***

Starting from this (simplified) model, we can build different characters simply changing the size of the mood zones: for example, an irascible character has the angry zone very wide and the quiet zone very narrow. Further, an attractor (a sort of magnetic pole) denoting the most common and the most representative emotional state of the personality of the agent is introduced: a lack of inputs (carrots/sticks) for a certain time will cause the current mood to drift toward the attractor. At the extremes of the circle, in any zone, the emotion state is very instable and can cause in certain cases jumps of the current mood toward very distant zones. The mood dynamics in the Emotion space is also characterized by a sort of anisotropy.

# References


Bates, J. (1994). The role of emotions in believable agents, *Comm. of the ACM*, 37(3): 122-125.

Camurri, A., P.Morasso, V.Tagliasco, and R.Zaccaria, (1986) Dance and movement notation, in P.Morasso and V.Tagliasco (Eds.), *Human Movement Understanding*, North Holland.

A.Camurri, F.Giuffrida, G.Vercelli, R.Zaccaria (1993). A System for the Real Time Control of Human Models on Stage, Proc. *X Colloquio di Informatica Musicale*, AIMI.

Camurri, A. (1995) Interactive Dance/Music Systems. Proc. *Intl. Computer Music Conference ICMC'95*, Banff, ICMA Press.

Camurri, A., Coglio, A., Coletta, P., Massucco, C. (1997) An Architecture for Multimodal Environment Agents. Proc. *AIMI Intl. Workshop on Kansei – The Technology of Emotion*, DIST-Univ. of Genova.

Camurri, A., Dondi, G., Gambardella, G. (1997). Interactive science exhibition: A playground for true and simulated emotions. Proc. *AIMI Intl. Workshop on Kansei – The Technology of Emotion*, DIST-University of Genova.

Camurri, A., Ferrentino, P., Dapelo, R. (1997). A computational model of artificial emotions. Proc. *AIMI Intl. Workshop on Kansei – The Technology of Emotion*, DIST-University of Genova.

De Rossi, D., A.Della Santa, A.Mazzoldi (1997) DressWare: Wearable Piezo- and Thermoresistive Fabrics for Ergonomics and Rehabilitation. Proc. *Nineteenth Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago.

Konolige, K. Saphira software documentation and demos. http://www.ai.sri.com/~konolige/saphira/#publications

Maletic, V. (1987). *Body Space Expression.* Mouton de Gruyter.

Nigay, L., & J.Coutaz (1993) A design space for multimodal systems - concurrent processing and data fusion. Proc. *INTERCHI-93 - Conf. on Human Factors in Computing Systems*, Amsterdam, 172-178, Addison Wesley.

Riecken, D. (1992) Wolfgang - A system using emoting potentials to manage musical design, in *Understanding Music with AI: Perspectives on Music Cognition*, M.Balaban, K.Ebcioglu, O.Laske (Eds.), AAAI/MIT Press, 1992.

Rowe. R. (1993). *Interactive music systems*. The MIT Press, Cambridge, MA.

Sawada, H., S.Ohkura, S.Hashimoto (1995) Gesture Analysis Using 3D Acceleration Sensor for Music Control", Proc. *Intl. Computer Music Conference ICMC-95*, Banff, ICMA.

Schomaker, L., J.Nijtmans, A.Camurri, F.Lavagetto, P.Morasso, C.Benoit, T.Guiard-Marigny, B. Le Goff, J.Robert-Ribes, A.Adjioudani, I.Defee, S.Munch, K,Hartung, J.Blauert (1995) *A Taxonomy of Multimodal Interaction in the Human Information Processing System*, A Report of the Esprit BRA Project 8579 MIAMI, WP1, February 1995 (available on internet at the site http://www.nici.kun.nl/~miami/reports/reports.html).

Sloboda, J.A. (1985). *The Musical Mind. The Cognitive Psychology of Music*. Oxford University Press.

Winkler, T. (1995) Making Motion Musical: Gesture Mapping Strategies for Interactive Computer Music. Proc. *Intl. Computer Music Conference ICMC-95*, Banff.